КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Спеціальність: 122 «Комп'ютерні науки»

Освітня програма: «Інформаційна аналітика та впливи»

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему:

"Розробка технології оптимізації портфоліо методами машинного навчання"

Студента 2-го курсу групи ІАВ-21

Ящука Миколи Олександровича

(прізвище, ім'я, по батькові)

Науковий керівник:

Доктор Технічних наук

(науковий ступінь, вчене звання)

Осауленко Ігор Анатолійович

(прізвище, ім'я, по батькові)

_____ _____ _____

(підпис студента) (дата) (підпис)

Попередній захист:

_____

(Висновок: «До захисту в Екзаменаційній комісії»)

Завідувач кафедри

технологій управління _____ _____ _____

(підпис) (прізвище, ініціали) (дата)

Київ – 2023

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет інформаційних технологій

Кафедра технологій управління

Освітньо-кваліфікаційний рівень Магістр

Спеціальність 122 - Комп'ютерні науки

Освітня програма Інформаційна аналітика та впливи

ЗАТВЕРДЖУЮ

Завідувач кафедри

професор Морозов В.В.

_____

«____» _____20___року

З А В Д А Н Н Я

НА ВИКОНАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Студент _ Ящук Микола Олександрович

Група_ІАВ-21_

1. Тема кваліфікаційної роботи: "Розробка технології оптимізації портфоліо методами машинного навчання"

Затверджена наказом по від «___»_____20__ р. № ___.

2. Строк подання студентом готової роботи – "___"_____ 20___ р.

3. Цільова установка та вихідні дані до роботи: дослідження оптимізації портфоліо за допомогою прогнозування із застосуванням часових рядів; прогнозування здійсьнюється на основі продажів продуктів і може бути розширено на інші категорії для прогнозування

4. Зміст роботи: вивчення теоретичних засад аналізу часових рядів, таких як методи та підходи до аналізу, попередня обробка даних, побудова прогнозів та інше. Поглиблене дослідження методу xgboost. Також проводилося дослідження практичного застосування прогнозування часових рядів для прогнозування продажів, а зрештою оптимізацію портфоліо. В рамках дослідження вивчалися інструменти і технології, необхідні для розроблення інформаційного забезпечення, яке б забезпечило автоматизоване прогнозування часових рядів.

5. Перелік графічного матеріалу (слайдів)

_____

6. Календарний план виконання роботи:

№ п / п	Назва частин роботи	%	Виконання роботи	
			За планом	Фактично
1.	Вибір теми дипломної роботи	3	01.10.22	01.10.22
2.	Протокол кафедри ТУ про затвердження тем дипломних робіт та призначення наукових керівників	2	27.12.22	27.12.22
3.	Формування переліку Нормативних матеріалів, літератури з проблематики	10	08.01.23	08.01.23

	дипломної роботи			
4.	Складання розгорнутого плану кваліфікаційної роботи	5	18.01.23	18.01.23
5.	Ознайомлення наукового керівника з розгорнутим планом кваліфікаційної роботи. Внесення змін.	5	20.11.23	20.11.23
6.	Підготовка розділу 1 «ТЕОРЕТИЧНІ ВІДОМОСТІ ПРО ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ ДЛЯ ОПТИМІЗАЦІЇ ПОРТФОЛІО »	10	12.02.23	12.02.23
7.	Підготовка розділу 2 «ПРАКТИЧНЕ ЗАСТОСУВАННЯ ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ ДЛЯ ОПТИМІЗАЦІЇ ПОРТФОЛІО»	14	08.03.23	08.03.23
8.	Підготовка розділу 3 «РЕАЛІЗАЦІЯ МЕТОДІВ АНАЛІЗУ ТА МОДЕЛЮВАННЯ ВЕБРЕСУРСІВ ЕЛЕКТРОННОЇ КОМЕРЦІЇ»	14	01.04.23	01.04.23
10.	Оформлення кваліфікаційної роботи. Підготовка висновків і пропозицій	15	03.05.23	03.05.23
11.	Передача кваліфікаційної роботи науковому керівникові	2	04.05.23	04.05.23
12.	Передача кваліфікаційної роботи рецензенту для рецензування	2	11.05.23	11.05.23

13.	Попередній захист кваліфікаційної роботи	5	17.05.23	17.05.23

Дата видачі завдання «____»_____20__р.

Керівник роботи _____

(посада, прізвище, ім’я, по батькові)

_____

(підпис)

Завдання прийняв до виконання студент групи _____

_____

(прізвище, ім’я, по батькові)

ЗМІСТ

АНОТАЦІЯ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій
Кафедра технологій управління
Спеціальність 122 – Комп’ютерні науки,
освітня програма «Інформаційна аналітика та впливи»

Дипломна робота магістра Ящука Миколи Олександровича.

Тема роботи – «Розробка технології оптимізації портфоліо методами машинного навчання».

Мета дипломної роботи магістра – створити оптимізовану модель прогнозування продажів, як частину системи оптимізації портфоліо на основі аналізу даних.

Об’єкт дослідження – моделювання прогнозу продажів підприємства. Предмет дослідження – збір та аналіз даних для моделювання прогнозу продажів підприємства .

Наукова новизна полягає у дослідженні, визначенні конкретних методик та опису конкретного плану дій для аналізу та моделювання прогнозу продажів підприємства, адже наявні рішення не покривають потреби бізнесу.

Практична значимість роботи зосереджена в створенні чіткого алгоритму дій збору, аналізу і використання даних з подальшим моделюванням прогнозу продажів підприємства, що дозволить менеджерам досить якісно оптимізувати портфоліо власного виробництва для покращення планування і фінансових метрик.

Дипломна робота складається зі вступу, основної частини, що включає у себе три розділи, висновків, списку використаних джерел та додатків. Всього робота налічує 82 сторінки та перелік з 9 посилань.

Ключові слова: модель, портфоліо, оптимізація, гіперпараметри, продажі, методи.

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

XGBOOST – eXtreme Gradient Boosting

БД – база даних.

ОС – операційна система.

ПК – персональний комп’ютер.

КРМ – кваліфікаційна робота магістра.

КВ - ковзне вікно (sliding window)

ВСТУП

Оптимізація портфелю для підприємств є важливим завданням, яке вимагає здійснення ефективних інвестиційних рішень для досягнення максимального прибутку або мінімізації ризику. Застосування методів машинного навчання, зокрема аналізу часових рядів, у процесі оптимізації портфелю має кілька переваг, що робить цю тему перспективною.

- Обробка великого обсягу даних: При аналізі портфелю підприємств, велика кількість даних зазвичай доступна в формі часових рядів, що включають історичні дані про ціни продажі, фінансові показники, ринкові індекси тощо. Методи машинного навчання дозволяють ефективно обробляти і аналізувати ці дані, виявляти складні залежності і використовувати їх для прийняття рішень щодо оптимізації портфелю.

- Прогнозування та моделювання: Методи аналізу часових рядів дозволяють прогнозувати майбутні значення показників, таких як майбутні продажі. Застосування цих методів дозволяє підприємствам розробляти моделі для прогнозування ризику та доходності інвестицій, а також оцінювати ефективність різних стратегій портфелю.

- Управління ризиком: Однією з ключових задач управління портфелем є зменшення ризику і збереження капіталу. Методи машинного навчання можуть допомогти виявити складні залежності між різними активами і розробити стратегії управління ризиком, які дозволяють підприємствам знизити вплив волатильності ринку.

Підприємства оптимізують свої портфелі підприємств з метою досягнення максимального прибутку або мінімізації ризику. Процес оптимізації портфелю включає такі кроки:

Збір даних: Першим кроком є збір відповідних даних про активи, включаючи історичні продажі, фінансові показники, ризики та інші фактори, які впливають на ціну активів. Ці дані можуть бути зібрані з внутрішніх джерел підприємства або зовнішніх джерел, таких як фінансові бази даних, ринкові індекси та економічні показники.

Визначення цілей та обмежень: Підприємство повинно визначити свої цілі та обмеження щодо оптимізації портфелю. Наприклад, ціллю може бути максимізація прибутку, мінімізація ризику або досягнення певного рівня доходності при певному ризику. Обмеженнями можуть бути обмеження на види активів, ризикові настанови або розмір портфелю.

Вибір моделі оптимізації: Після визначення цілей і обмежень підприємство вибирає відповідну модель оптимізації портфелю. Це може включати аналітичні моделі, статистичні методи або методи машинного навчання, такі як аналіз часових рядів. Методи машинного навчання дозволяють побудувати моделі, які здатні виявляти складні залежності та прогнозувати майбутні значення показників.

Виконання оптимізації: Після вибору моделі оптимізації підприємство виконує сам процес оптимізації. В цьому кроці використовуються алгоритми та методи моделі для розрахунку оптимального розподілу коштів між різними активами в портфелі. Оптимізація може враховувати ризики, очікувану доходність, кореляції між активами та інші фактори.

Моніторинг та оновлення: Портфель підприємства потребує постійного моніторингу та оновлення. Ринкові умови, фінансові показники та інші фактори можуть змінюватися з часом, тому важливо періодично переглядати та оновлювати портфель, враховуючи нові дані та умови.

Застосування методів машинного навчання, зокрема аналізу часових рядів, в оптимізації портфелю дозволяє ефективно використовувати історичні дані, прогнозувати майбутні значення, виявляти складні залежності та приймати обґрунтовані рішення щодо розподілу активів в портфелі. Це допомагає підприємствам знизити ризики, збільшити доходність та досягти більш ефективного управління портфелем.

Прогнозування продажів є ключовою ідеєю щодо оптимізації портфоліо, оскільки має прямий вплив на дохід підприємства. Нижче наведено декілька причин, чому прогнозування продажів є важливим елементом оптимізації портфелю:

Планування виробництва та запасів: Прогнозування продажів дозволяє підприємствам зрозуміти очікувані обсяги продажів у майбутньому. Ця інформація є критичною для планування виробництва і визначення оптимального рівня запасів. Недооцінка або переоцінка попиту може призвести до непотрібного накопичення запасів або недостатнього обсягу товарів, що може вплинути на доходність підприємства.

Управління постачанням та ланцюгом постачання: Прогнозування продажів дозволяє підприємствам управляти постачальницьким ланцюгом, планувати поставки та підтримувати оптимальний рівень запасів у всій ланцюгу постачання. Це допомагає уникнути зайвих витрат на зберігання та недостач товарів, забезпечуючи ефективну операційну діяльність.

Фінансове планування: Прогнозування продажів є важливим елементом фінансового планування підприємства. Очікувані обсяги продажів дозволяють підприємству визначити очікуваний дохід, витрати та прибуток. Це дозволяє управляти фінансами підприємства, розробляти бюджет та забезпечувати фінансову стійкість.

Планування маркетингових стратегій: Прогнозування продажів надає підприємству інформацію про очікуваний попит на їх продукцію або послуги. Це дозволяє розробити ефективні маркетингові стратегії, спрямовані на досягнення цільової аудиторії і максимізацію продажів. Розуміння тенденцій та змін у попиті дозволяє підприємству адаптувати свої маркетингові зусилля та стратегії, що сприяє ефективному управлінню портфелем.

Враховуючи важливість прогнозування продажів, методи машинного навчання, зокрема аналіз часових рядів, можуть бути використані для розробки точних та надійних моделей прогнозування, що допоможуть підприємствам ефективно оптимізувати свій портфель та приймати обґрунтовані рішення щодо виробництва, постачання та маркетингу.

Відповідно, само з цієї аргументції тема «Розробка технології оптимізації портфоліо методами машинного навчання» є актуальною і це є темою моєї КРМ.

Мета роботи – створити оптимізовану модель прогнозування продажів, як частину системи оптимізації портфоліо на основі аналізу даних. Тому об’єктом дослідження є моделювання прогнозу продажів підприємства, а предметом дослідження – збір та аналіз даних для моделювання прогнозу продажів підприємства.

Наукова новизна полягає у дослідженні, визначенні конкретних методик та опису конкретного плану дій для аналізу та моделювання прогнозу продажів підприємства, адже наявні рішення не покривають потреби бізнесу.

Практична значимість роботи зосереджена в створенні чіткого алгоритму дій збору, аналізу і використання даних з подальшим моделюванням прогнозу продажів підприємства, що дозволить менеджерам досить якісно оптимізувати портфоліо власного виробництва для покращення планування і фінансових метрик.

Основні методи досліджень в КРМ є поглиблена аналітика для розуміння даних, а також застосування методів машинного навчання для часових рядів з метою найкращого прогнозування продажів.

Для даної КРМ були визначені та виконані наступні завдання:

- проаналізувати інформаційні джерела про принципи та особливості методів оптимізації портфоліо;
- оцінити існуючі описані рішення оптимізації портфоліо;
- створити план проведення аналізу даних для моделювання продажів;
- зібрати, обробити та підготувати дані для моделювання прогнозування продажів за допомогою часових рядів;
- створити структурну модель прогнозування продажів на основі проаналізованих даних;
- на основі отриманих результатів сформувати чіткий план дій (чек-лист) підготовки і аналізу даних з подальшим моделювання продажів для підприємства.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ВІДОМОСТІ ПРО ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ ДЛЯ ОПТИМІЗАЦІЇ ПОРТФОЛІО

1.1 Вступ до розділу

У цьому розділі проведено аналіз літератури, спрямований на прогнозування часових рядів для оптимізації портфоліо. Було ретельно вивчено різні джерела, включаючи наукові статті та публікації, з метою визначити різні методи та інструменти, що використовуються в цій галузі, а також оцінити досягнуті результати.

У процесі аналізу було приділено особливу увагу методам прогнозування часових рядів, які можуть бути застосовані для прогнозування продажів і подальшою метою досягнення оптимального портфоліо. Розглянуті методи включають алгоритми ансамблю, конкретно XGBoost, який дозволє прогнозувати майбутні значення часових рядів на основі історичних даних.

Також було проаналізовано застосування методів аналізу та прогнозування часових рядів в промисловості. Це дозволило виявити, які методи та підходи є найбільш ефективними в контексті оптимізації портфоліо в промисловому секторі та відповідають цілям нашого магістерського дослідження.

Крім того, були досліджені методи попередньої обробки часових рядів, зокрема обробка пропущених значень, нормалізація даних та виявлення та видалення випадкових шумів. Ці методи є важливими передумовами для отримання точних та надійних прогнозів. Обробка пропущених значень дозволяє вирішити проблему відсутніх або неповних даних, що можуть впливати на якість прогнозування. Нормалізація даних сприяє уніфікації значень часових рядів і допомагає уникнути викривлення результатів. Виявлення та видалення випадкових шумів забезпечує покращення чистоти та точності даних перед їх подальшим аналізом та прогнозуванням. Ці методи попередньої обробки даних сприяють покращенню якості прогнозів часових рядів і є важливими етапами в оптимізації портфоліо в промисловій галузі.

Для досягнення більш точних прогнозів часових рядів було вивчено ансамбль моделей, який є одним з ефективних підходів у прогнозуванні. Ансамбль моделей поєднує прогнози, отримані з різних моделей, для отримання більш об'єктивного та стійкого результату. Використання ансамблю моделей дозволяє скоригувати можливі недоліки окремих моделей та використовувати їх сильні сторони для досягнення кращої точності прогнозування.

Крім того, було вивчено інформаційні технології для автоматизації процесу аналізу та прогнозування часових рядів, зокрема автоматизований підбір гіперпараметрів. Автоматизований підбір гіперпараметрів використовує алгоритми оптимізації для автоматичного визначення оптимальних значень гіперпараметрів моделі. Гіперпараметри визначають конфігурацію моделі, таку як кількість шарів у нейромережі чи параметри регуляризації. Автоматизований підбір гіперпараметрів дозволяє знайти оптимальні значення гіперпараметрів шляхом систематичного пошуку в просторі можливих значень та оцінки їх впливу на точність та ефективність моделі.

Застосування цієї інформаційної технологій дозволяє зменшити людський вплив на процес аналізу та прогнозування часових рядів і підвищити ефективність оптимізації портфоліо на основі аналізу часових рядів.

Отже, цей розділ має на меті провести комплексний аналіз існуючих методів прогнозування часових рядів для оптимізації портфоліо, визначити їхню ефективність в різних випадках та визначити методи, які можна використовувати в нашому магістерському дослідженні.

1.2 Використання аналізу часових рядів у сфері бізнесу

Аналіз часових рядів має велике значення в сфері бізнесу, оскільки дозволяє прогнозувати майбутні події та тенденції на основі історичних даних. Однією з важливих галузей, де використовується аналіз часових рядів, є виробництво.

Виробничі підприємства мають велику кількість даних, які можуть бути представлені у формі часових рядів. Наприклад, дані про виробничі обсяги, рівень виробничої активності, продажі, запаси сировини та готової продукції можуть бути представлені у вигляді часових рядів. Аналіз цих даних дозволяє зрозуміти динаміку виробничого процесу, виявити сезонні залежності, тренди, аномалії та інші важливі закономірності.

Прогнозування часових рядів у виробничій сфері має на меті допомогти в управлінні виробничими процесами, прийнятті стратегічних рішень та плануванні ресурсів. Наприклад, на основі прогнозів можна планувати виробничі обсяги, оптимізувати запаси, розподіляти ресурси, планувати виробничий графік та планувати замовлення сировини.

Одним із важливих аспектів аналізу часових рядів у виробництві є виявлення та управління випадковими шумами. Виробничі процеси часто піддаються впливу випадкових факторів, таких як технічні несправності, збої у постачанні сировини, зміни вимог споживачів тощо. Аналіз часових рядів дозволяє виявити такі шуми та врахувати їх в прогнозуванні, що сприяє зниженню ризиків та покращенню якості прийнятих рішень.

Окрім того, використання аналізу часових рядів у виробництві допомагає здійснювати ефективний моніторинг процесів та оцінювати їх продуктивність. Виробничі дані можуть бути аналізовані в режимі реального часу, що дозволяє оперативно реагувати на зміни та вживати відповідних заходів для досягнення поставлених цілей.

У підсумку, аналіз часових рядів виробничих даних є потужним інструментом для виробничих підприємств. Він дозволяє розуміти динаміку виробничого процесу, а також забезпечує оперативний моніторинг та контроль процесів. Застосування аналізу часових рядів у виробництві сприяє підвищенню ефективності, зниженню ризиків та досягненню стратегічних цілей підприємства.

1.3 Оцінка практичного застосування аналізу часових рядів при оптимізації портфоліо

Оцінка практичного застосування аналізу часових рядів в оптимізації портфоліо є ключовою темою в фінансовому менеджменті. Аналіз часових рядів дозволяє прогнозувати та оцінювати ризики та доходність різних активів, що допомагає інвесторам та управляючим фондами приймати обґрунтовані рішення щодо складання та оптимізації інвестиційного портфелю.

Прогнозування продажів є одним з основних використань аналізу часових рядів в оптимізації портфеліо. Багато компаній, особливо ті, що працюють у сфері роздрібної торгівлі, залежать від точних прогнозів продажів для планування запасів, виробництва та управління постачанням. Використання аналізу часових рядів дозволяє аналізувати історичні дані про продажі, виявляти сезонні залежності, тренди та інші фактори, що впливають на продажі. На основі цих аналізів можна побудувати моделі прогнозування, які допоможуть прогнозувати майбутні продажі з високою точністю. Це дозволить підприємствам планувати свої дії, оптимізувати виробництво та постачання, уникнути надмірних запасів або дефіциту товарів, а також приймати обґрунтовані рішення щодо розподілу ресурсів та маркетингових зусиль.

У контексті оптимізації портфеліо, прогнозування продажів є критичним фактором для інвесторів. Вони мають зацікавленість у складанні оптимального портфелю, що максимізує дохідність та мінімізує ризики. Аналіз часових рядів дозволяє інвесторам прогнозувати майбутні доходи та волатильність різних активів. За допомогою таких прогнозів можна побудувати ефективну стратегію розподілу активів, враховуючи ризики та доходність кожного активу. Це дозволить інвесторам диверсифікувати свій портфель, розподіляючи ресурси між різними активами з урахуванням їхньої прогнозованої доходності та ризиків. Такий підхід допомагає зменшити загальний ризик портфелю і забезпечує більш стабільний дохід для інвестора.

Одним із важливих аспектів оцінки практичного застосування аналізу часових рядів в оптимізації портфеліо є врахування фактору невизначеності та змінності. Ринки та економічні умови постійно змінюються, тому необхідно мати можливість адаптувати стратегії портфелю до нових умов. Аналіз часових рядів дозволяє виявляти зміни у трендах та сезонності, а також аналізувати реакцію активів на різні події та фактори. За допомогою цього аналізу можна здійснювати регулярну перебалансування портфелю, враховуючи нову інформацію та ризики. Такий гнучкий підхід дозволяє забезпечувати оптимальну адаптацію портфелю до змінних умов та досягати кращих результатів.

У підсумку, використання аналізу часових рядів при оптимізації портфеліо є важливим інструментом для інвесторів та управляючих фондів. Прогнозування продажів та оцінка ризиків та доходності активів допомагають приймати обґрунтовані рішення щодо складання та оптимізації портфелю. Застосування аналізу часових рядів дозволяє досягати більш ефективного розподілу ресурсів, знижувати ризики та максимізувати доходність портфелю. При цьому важливо враховувати невизначеність та змінність ринкових умов і здійснювати регулярну адаптацію стратегій портфелю.

1.4 Використання методів XGBoost для прогнозування часових рядів

Модель XGBoost є потужним інструментом, що знайшов широке застосування у прогнозуванні часових рядів. XGBoost є алгоритмом машинного навчання, який поєднує в собі силу градієнтного бустингу і оптимізовані реалізації для досягнення високої точності та швидкості обчислень.

Прогнозування часових рядів є важливою задачею в багатьох галузях, таких як фінанси, економіка, маркетинг, логістика тощо. Це допомагає компаніям та організаціям зрозуміти майбутні тенденції, попит на продукцію або послуги, а також приймати обґрунтовані рішення щодо планування виробництва, управління запасами, реклами та багато іншого.

Модель XGBoost відрізняється своєю здатністю працювати з великими обсягами даних та здатністю до автоматичної роботи зі змінними часу. Це дає змогу враховувати сезонність, тренди, циклічність та інші характеристики часових рядів, що є важливими у прогнозуванні.

Однією з основних переваг XGBoost є його здатність автоматично розпізнавати та моделювати складні взаємозв'язки між змінними. Він може ефективно використовувати багатовимірні дані та використовувати їх для покращення якості прогнозування. Модель XGBoost також може працювати з різними типами змінних, включаючи категоріальні, числові та текстові.

Іншою важливою характеристикою XGBoost є його здатність до оптимізації та управління гіперпараметрами. Гіперпараметри визначають конфігурацію моделі, таку як глибина дерев, швидкість навчання, кількість ітерацій тощо. Існує багато методів для автоматичного підбору оптимальних значень гіперпараметрів, таких як випадковий пошук, градієнтний спуск та баєсовська оптимізація. Це дозволяє покращити якість прогнозування та зменшити ризик перенавчання моделі.

Застосування моделі XGBoost в прогнозуванні часових рядів для оптимізації портфоліо може мати значний вплив на досягнення кращих результатів. Вона дозволяє враховувати динаміку ринку, змінність та непередбачуваність факторів, що впливають на продажі та доходність активів. Поєднання XGBoost з іншими методами та моделями також може покращити точність прогнозування та зменшити вплив шумів та неточностей.

Отже, модель XGBoost є потужним інструментом для прогнозування часових рядів, що знайшов широке застосування у сфері оптимізації портфоліо. Її здатність до роботи зі змінними часу, автоматичне виявлення взаємозв'язків та можливості оптимізації гіперпараметрів роблять її привабливим вибором для аналізу та прогнозування в бізнесі. Застосування цієї моделі може допомогти управляти ризиками, планувати виробництво та досягати оптимального розподілу ресурсів у портфелі.

1.5 Теоретична складова XGBoost для прогнозування часових рядів

XGBoost - це потужна бібліотека для градієнтного бустингу, яка широко використовується для задач машинного навчання, включаючи прогнозування часових рядів. Теоретична складова XGBoost включає кілька ключових понять та алгоритмів, які лежать в основі його роботи. Основні складові XGBoost для прогнозування часових рядів включають:

1) Градієнтний бустинг: Градієнтний бустинг є основною концепцією за алгоритмом XGBoost. Він базується на ідеї об'єднання слабких моделей для створення потужної ансамбльної моделі. Градієнтний бустинг покроково навчає модель, додаючи нову модель для виправлення помилок попередньої моделі. Цей процес продовжується доти, поки не буде досягнуто задовільного рівня точності. Градієнтний бустинг визначається наступним чином:

    - Функція втрати (Loss function):

    L(y, F) - функція втрати, де y - справжні значення, F - прогнозовані значення

    - Перший порядок апроксимації (First-order approximation):

    r = -[∂L(y, F) / ∂F] - градієнт функції втрати по прогнозованим значенням

- Другий порядок апроксимації (Second-order approximation):

h = $\partial^2 L(y, F) / \partial F^2$ - другий похідний функції втрати по прогнозованим значенням

- Функція об'єднання (Aggregation function):

T = ∑t=1 to T γ_t h_t - функція об'єднання, де γ_t - швидкість навчання (learning rate), h_t - окрема модель у бустингу

- Оновлення прогнозу (Update step):

F_t(x) = F_{t-1}(x) + T(x) - оновлення прогнозу на кожному кроці бустингу, де F_t(x) - прогноз після t-го кроку, F_{t-1}(x) - прогноз після (t-1)-го кроку, T(x) - внесок окремої моделі на кожному кроці

Ці рівняння використовуються для обчислення градієнту, другої похідної та оновлення прогнозу на кожному кроці градієнтного бустингу в моделі XGBoost.

2) Дерев'яні ансамблі: XGBoost використовує ансамбль рішень, зокрема, дерев'яні моделі, як базові моделі. Дерев'яні моделі є гнучкими, вони можуть виявити складні шаблони в часових рядах, а також автоматично вирішувати проблему важливості ознак.

Основні рішення для градієнтних дерев'яних моделей в XGBoost включають:

- Об'єктивна функція (Objective function):

Obj = L + Ω - об'єктивна функція, яка включає в себе функцію втрати L та регуляризаційний член Ω

- Функція втрати (Loss function):

L = ∑i=1 to n l(yi, yî) - функція втрати, яка вимірює різницю між справжніми значеннями yi та прогнозованими значеннями yî для кожного прикладу даних (індекс i)

- Регуляризаційний член (Regularization term):

Ω = γT + 1/2λ∑t=1 to T wt² - регуляризаційний член, який складається з двох компонент: перша компонента - штраф за кількість листків T з використанням параметру γ (gamma), а друга компонента - штраф за ваги листків wt з використанням параметру λ (lambda)

- Правило розбиття (Splitting rule):

$Gj = \sum i \in Ij \, \partial L / \partial yî + \lambda wj$, $Hj = \sum i \in Ij \, \partial^2 L / \partial yî^2 + \lambda$ - градієнти та другі похідні функції втрати для листка j, де Ij - множина прикладів даних, що попадають в листок j, Gj - сума градієнтів, Hj - сума других похідних

- Внесок листка (Leaf contribution):

$wj = -Gj / (Hj + \lambda)$ - внесок листка j, який визначається врахуванням градієнту та другої похідної функції втрати

- Перебудова дерева (Tree building):

$Obj = \sum j=1 \text{ to } T \, (Gj^2 / (Hj + \lambda) + \gamma T) + \Omega(t)$ - об'єктивна функція для побудови дерева, де $\Omega(t)$ - додатковий штраф за структуру дерева, який враховується на кожному кроці побудови

Ці формули використовуються для побудови та оптимізації градієнтних дерев'яних моделей в XGBoost, включаючи визначення функції втрати, регуляризаційного члена, правил розбиття, внеску листка та об'єктивної функції для побудови дерева.

3) Функція втрати: Функція втрати визначає спосіб оцінки помилки моделі та визначення напрямку оптимізації. У випадку прогнозування часових рядів зазвичай використовуються функції втрати, такі як середньоквадратична помилка (Mean Squared Error) або середня абсолютна помилка (Mean Absolute Error).

Для прогнозування часових рядів в моделі XGBoost застосовуються різні функції втрати. Основні з них включають:

- Середньоквадратична помилка (Mean Squared Error, MSE):

$L(y, F) = (1/n)\sum(y_i - F_i)^2$ - середньоквадратична помилка, де y_i - справжнє значення, F_i - прогнозоване значення, n - кількість прикладів даних

- Середня абсолютна помилка (Mean Absolute Error, MAE):

$L(y, F) = (1/n)\sum|y_i - F_i|$ - середня абсолютна помилка, де y_i - справжнє значення, F_i - прогнозоване значення, n - кількість прикладів даних

- Хи-квадрат (Chi-square):

$L(y, F) = 2 * [(F - y) / (F + y)]$ - хи-квадрат, де y - справжнє значення, F - прогнозоване значення

- Логарифмічна функція втрати (Log Loss):

$L(y, F) = -(y \log(F) + (1 - y) \log(1 - F))$ - логарифмічна функція втрати, де y - справжнє бінарне значення (0 або 1), F - ймовірність класу 1

- Експоненційна функція втрати (Exponential Loss):

L(y, F) = exp(-yF) - експоненційна функція втрати, де y - справжнє значення, F - прогнозоване значення

Ці формули використовуються для визначення помилки моделі та визначення напрямку оптимізації в XGBoost для прогнозування часових рядів. Залежно від конкретного випадку може бути вибрана певна функція втрати для досягнення найкращих результатів.

4) Регуляризація: XGBoost надає декілька параметрів регуляризації, які допомагають уникнути перенавчання моделі та забезпечити кращу узагальнюючу здатність. Деякі з популярних параметрів регуляризації XGBoost включають штраф за складність моделі (L1 або L2 регуляризація), штраф за кількість листків дерева (глибина дерева) та штраф за розподіл ваг між листками.

- L1 регуляризація (Lasso):

$\Omega_1 = \gamma \sum|w|$ - L1 регуляризація, де w - ваги моделі, γ - параметр регуляризації

- L2 регуляризація (Ridge):

$\Omega_2 = \gamma \sum w^2$ - L2 регуляризація, де w - ваги моделі, γ - параметр регуляризації

- Штраф за кількість листків дерева:

$\Omega_3 = \gamma T$ - штраф за кількість листків T, де γ - параметр регуляризації

- Штраф за розподіл ваг між листками:

$\Omega_4 = \lambda \sum wt^2$ - штраф за розподіл ваг wt між листками, де λ - параметр регуляризації

- Загальна функція регуляризації:

$\Omega = \lambda \sum wt^2 + \gamma T$ - загальна функція регуляризації, яка включає L2 регуляризацію за вагами та L1 регуляризацію за кількістю листків

Ці формули використовуються для визначення штрафів та функцій регуляризації в XGBoost. Залежно від потреби можна встановлювати параметри регуляризації для контролю перенавчання моделі та поліпшення її узагальнюючої здатності.

5) Функції важливості ознак: XGBoost надає можливість оцінити важливість кожної ознаки в прогнозуванні часового ряду. Це допомагає зрозуміти, які ознаки мають найбільший вплив на прогноз.

- Важливість ознаки на основі приросту вдосконалення (Gain-based feature importance):

$\text{Importance}(j) = \sum(\Delta\text{Gain}(t, j)) / \sum(\text{Gain}(t, j))$ - важливість ознаки j, яка вимірюється як відношення суми змін у функції втрати (Gain) внаслідок розбиття на ознаку j до загальної суми Gain по всіх розбиттях на ознаку j

- Важливість ознаки на основі покращення критерію (Weight-based feature importance):

Importance(j) = ∑(w(t) * I(t, j)) / ∑(w(t)) - важливість ознаки j, де w(t) - вага прикладу даних t, I(t, j) - індикатор розбиття на ознаку j для прикладу t

- Важливість ознаки на основі покращення шорт-цілі (Cover-based feature importance):

- Importance(j) = ∑(w(t) * C(t, j)) / ∑(w(t)) - важливість ознаки j, де w(t) - вага прикладу даних t, C(t, j) - кількість разів, коли ознака j була використана для розбиття прикладу t

Ці формули використовуються для визначення важливості ознак в XGBoost. Вони базуються на аналізі внеску кожної ознаки до покращення функції втрати, критерію або шорт-цілі. Залежно від використаної метрики можна отримати інформацію про вплив кожної ознаки на прогноз часового ряду.

Ці складові разом допомагають XGBoost стати потужним інструментом для прогнозування часових рядів, здатним моделювати складні шаблони, використовуючи ансамбль дерев'яних моделей та оптимізуючи функцію втрати з використанням градієнтного бустингу.результаті аналізу літератури було виявлено, що для досягнення точних та надійних прогнозів часових рядів в оптимізації портфоліо варто використовувати ансамбль моделей. Цей підхід дозволяє комбінувати прогнози, отримані з різних моделей, з метою отримання більш об'єктивних та стійких результатів. Ансамбль моделей дозволяє скоригувати можливі недоліки окремих моделей та використовувати їх переваги для підвищення точності прогнозів.

1.6 Висновки розділу та подальші задачі

В результаті аналізу літератури було виявлено, що для досягнення точних та надійних прогнозів часових рядів в оптимізації портфоліо варто використовувати ансамбль моделей. Цей підхід дозволяє комбінувати прогнози, отримані з різних моделей, з метою отримання більш об'єктивних та стійких результатів. Ансамбль моделей дозволяє скоригувати можливі недоліки окремих моделей та використовувати їх переваги для підвищення точності прогнозів.

Крім того, використання інформаційних технологій, зокрема автоматизованого підбору гіперпараметрів, може значно полегшити процес аналізу та прогнозування часових рядів. Автоматизований підбір гіперпараметрів використовує алгоритми оптимізації для автоматичного визначення оптимальних значень гіперпараметрів моделі. Це дозволяє знайти найкращі значення гіперпараметрів шляхом систематичного пошуку та оцінки їх впливу на точність та ефективність моделі. Використання цієї інформаційної технології дозволяє знизити вплив людини на процес прогнозування та забезпечити більш ефективну оптимізацію портфоліо на основі аналізу часових рядів.

Отже, використання ансамблю моделей та інформаційних технологій, таких як автоматизований підбір гіперпараметрів, може покращити якість прогнозів та ефективність оптимізації портфоліо на основі аналізу часових рядів. Дані методи та підходи можуть бути використані в магістерському дослідженні з метою досягнення кращих результатів.

РОЗДІЛ 2. ПРАКТИЧНЕ ЗАСТОСУВАННЯ ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ ДЛЯ ОПТИМІЗАЦІЇ ПОРТФОЛІО

2.1 Модель XGBoost часових рядів при прогнозуванні

XGBoost є потужним алгоритмом для прогнозування часових рядів з кількох причин:

Обробка нелінійних зв'язків: Однією з ключових переваг моделі XGBoost при прогнозуванні часових рядів є її здатність фіксувати нелінійні залежності між змінними у даних. У багатьох випадках, особливо в складних бізнес-сценаріях, залежності між факторами та вихідними значеннями можуть бути нелінійними. XGBoost виявляється ефективним інструментом для моделювання та урахування цих складних нелінійних залежностей.

Інтеграція цих нелінійних зв'язків у модель дозволяє виявити та врахувати складні закономірності, які присутні в даних часових рядів. Наприклад, можуть існувати неочікувані взаємозв'язки між часом, погодою, економічними факторами та продажами, які можуть бути недосяжні для простих лінійних моделей. XGBoost може автоматично виявляти та використовувати ці нелінійні закономірності, що дозволяє точніше прогнозувати майбутні значення часового ряду.

Ця здатність моделі XGBoost до фіксації нелінійних зв'язків робить її особливо корисною в ситуаціях, де існує велика кількість факторів, що впливають на динаміку часового ряду. Вона дозволяє виявити складні взаємозв'язки між цими факторами і точніше передбачити майбутні зміни. Наприклад, у сфері фінансів XGBoost може враховувати взаємозв'язки між ринковими трендами, валютними курсами, економічними показниками та цінами акцій, що дозволяє зробити більш точні прогнози для оптимізації портфеліо та прийняття рішень щодо інвестування.

Важливість функцій: Одна зі суттєвих переваг моделі XGBoost при прогнозуванні часових рядів полягає в її здатності визначати важливість різних функцій або змінних у даних. Ця інформація допомагає виокремити найбільш впливові фактори, які визначають поведінку часового ряду та впливають на його майбутні значення.

XGBoost використовує алгоритми градієнтного бустінгу для поступового покращення моделі шляхом додавання нових дерев рішень. Під час цього процесу модель оцінює важливість кожної змінної, враховуючи її вплив на точність прогнозування. Чим більше змінна сприяє зменшенню помилки моделі, тим вищий її рейтинг важливості.

Ця можливість ідентифікувати найбільш важливі змінні в даних часових рядів має велике значення для бізнесу. Наприклад, в контексті прогнозування продажів, XGBoost може виявити, що певні фактори, такі як ціна товару, рекламні витрати або погодні умови, мають найбільший вплив на зміни у попиті. Це дає можливість бізнесу зосередити свої зусилля на цих ключових факторах, оптимізувати ресурси та приймати більш обґрунтовані рішення.

Більш того, знання про важливість функцій також сприяє відбору оптимального набору змінних для моделювання. Це дозволяє позбутися від непотрібних або слабо впливових факторів, що може покращити ефективність моделі та зменшити обчислювальну складність. В результаті, модель XGBoost здатна створювати більш точні та ефективні прогнози для часових рядів, сприяючи покращенню управління ризиками та оптимізації бізнес-процесів.

Контроль регулярізації та надмірного оснащення: XGBoost включає в себе ряд методів регуляризації, які є важливими для покращення якості прогнозування часових рядів. Одним з таких методів є усадка, яка вводить штраф для складних моделей, що мають багато розгалужень і можуть потрапити в пастку переобладнання. Це допомагає зменшити складність моделі та покращити її загальну здатність до узагальнення.

Крім того, XGBoost використовує підвибірку функцій, що означає випадкове включення певної підмножини змінних у кожному дереві. Це дозволяє зменшити кореляцію між деревами та зробити модель менш чутливою до конкретних шумів чи аномалій у даних. Цей підхід є ефективним способом контролювати надмірне оснащення, оскільки він дозволяє моделі зосередитися на найбільш інформативних функціях, запобігаючи перенапруженню моделі непотрібними змінними.

Важливою перевагою контролю регуляризації та надмірного оснащення є покращення стійкості та надійності прогнозів для часових рядів. Надмірне оснащення може спричинити значні відхилення між прогнозами та фактичними значеннями в нових, невідомих даних, що ускладнює точне прогнозування та прийняття рішень. Завдяки використанню методів регуляризації, XGBoost забезпечує більш узгоджені та надійні результати, зменшуючи ризик неправильних прогнозів та забезпечуючи більш точну передбачувану динаміку часових рядів.

Обробка відсутніх значень: В контексті аналізу часових рядів, наявність відсутніх значень є частим явищем, зокрема у реальних даних. XGBoost надає вбудовані механізми для ефективної обробки цих відсутніх даних, що сприяє покращенню точності та достовірності прогнозів.

XGBoost має розуміння про те, що відсутні значення можуть вплинути на точність моделі, тому він пропонує кілька підходів для їх обробки. Один із способів - це заміна відсутніх значень на попереднє відоме значення, що дозволяє зберегти структуру часового ряду та зменшити вплив пропусків на прогнозування. Іншим підходом є використання інтерполяційних методів, таких як лінійна або кубічна інтерполяція, для заповнення відсутніх значень на основі наявних даних.

Крім того, XGBoost дозволяє включати в модель додаткові ознаки, які можуть вказувати на наявність або відсутність значень у вхідних даних. Наприклад, створення бінарної ознаки, що вказує, чи є значення відсутнім, може допомогти моделі усвідомити цю особливість та адекватно обробити такі випадки.

Здатність XGBoost ефективно працювати з відсутніми даними важлива в реальних ситуаціях, де збір даних може бути неповним або незавершеним. Використання цього алгоритму дозволяє використовувати наявну інформацію максимально ефективно, зменшуючи вплив відсутності даних на результати прогнозування та забезпечуючи більш точні та надійні прогнози часових рядів.

Паралельна обробка та масштабованість - ще одна важлива перевага XGBoost при роботі з часовими рядами. Завдяки своїй високій масштабованості, XGBoost може ефективно використовувати паралельну обробку для навчання моделей на великих наборах даних.

Використання паралельної обробки дозволяє розподілити обчислювальні завдання між багатьма ядрами процесора або навіть між різними вузлами в розподіленій обчислювальній системі. Це забезпечує прискорення процесу навчання моделі і дозволяє швидше здійснювати експерименти та ітерацію моделі. За допомогою паралельної обробки XGBoost може швидко аналізувати великі обсяги даних часових рядів, що є важливим у сучасному аналізі даних.

Крім того, XGBoost може ефективно використовувати переваги розподілених обчислювальних інфраструктур, таких як кластери або хмарні середовища. Завдяки цьому, можна розподіляти навантаження обробки даних між різними вузлами, що дозволяє ефективно працювати з великими обсягами даних та прискорювати процес навчання моделі.

Застосування паралельної обробки та масштабованості XGBoost у контексті часових рядів робить його потужним інструментом для розв'язання складних завдань прогнозування. Він дозволяє аналізувати великі обсяги даних, виявляти складні залежності та швидко навчати моделі, що робить його незамінним інструментом для роботи з вимогливими за обчислювальними ресурсами задачами прогнозування часових рядів.

Ансамблеве навчання - ще одна потужна функція, яка відрізняє XGBoost в контексті прогнозування часових рядів. Замість використання однієї моделі, XGBoost використовує ансамбль моделей для отримання кращих прогнозів.

Ансамбль моделей означає, що XGBoost комбінує прогнози кількох слабких моделей, таких як дерева рішень, і об'єднує їх для отримання більш точного та надійного прогнозу. Кожна окрема модель може мати свої обмеження та недоліки, але колективне рішення, яке забезпечує ансамбль, може компенсувати ці обмеження та покращити загальну точність прогнозування.

Ансамбль моделей в XGBoost будується шляхом поєднання багатьох дерев рішень, які взаємодіють між собою. Кожне дерево рішень вирішує лише частину задачі прогнозування, і їх прогнози об'єднуються в кінцевий прогноз. Цей процес агрегування прогнозів дозволяє XGBoost виявляти та моделювати складні залежності у даних часових рядів.

Ансамблеве навчання з XGBoost забезпечує не лише покращення точності прогнозування, але й стабільність результатів. Завдяки агрегації прогнозів з кількох моделей, XGBoost може знизити вплив шуму або неправильних прогнозів окремих моделей і забезпечити більш надійний прогноз на основі колективного рішення.

Використання ансамблевого навчання з XGBoost у задачах прогнозування часових рядів дозволяє отримати більш точні та надійні результати, підвищуючи якість прогнозів і сприяючи кращому розумінню динаміки часових рядів.

Ці переваги роблять XGBoost популярним вибором для завдань прогнозування часових рядів, пропонуючи як точність, так і можливість інтерпретації, а також дозволяючи приймати рішення на основі даних у різних областях.

2.2 Прогнозування часових рядів як контрольоване навчання

Прогнозування часових рядів як контрольоване навчання є одним із підходів, що використовуються для моделювання та прогнозування динаміки в часових рядах. У цьому підході модель навчається за допомогою доступних даних, що включають історичні значення часового ряду та можливі вхідні фактори, які впливають на ряд. Суть контрольованого навчання полягає в тому, що модель побудована таким чином, щоб відповідати цілям прогнозування на основі наявної інформації.

Один з основних аспектів контрольованого навчання в прогнозуванні часових рядів - це використання історичних даних як навчального набору для моделювання. Історичні значення часового ряду використовуються для навчання моделі відтворювати залежності та закономірності, що спостерігаються в ряді. Це дозволяє моделі розуміти тенденції, сезонні зміни, аномалії та інші фактори, які впливають на розвиток часового ряду.

Крім того, в контрольованому навчанні можуть використовуватись додаткові вхідні фактори або змінні, які мають вплив на часовий ряд. Це можуть бути екзогенні змінні, які не залежать від самого ряду, але мають прогнозний потенціал. Наприклад, в прогнозуванні продажів можуть використовуватись економічні показники, погодні умови, маркетингові активності тощо. Включення таких факторів допомагає моделі узгоджувати прогноз з реальними змінами, що спостерігаються у вхідних факторах.

Прогнозування часових рядів як контрольоване навчання також може використовувати різні алгоритми та моделі, залежно від характеристик ряду та поставленої задачі. Наприклад, можуть бути використані методи на основі статистичних моделей, авторегресійні моделі, методи машинного навчання, такі як нейронні мережі, дерева рішень або ансамблеві методи.

Контрольоване навчання в прогнозуванні часових рядів дозволяє побудувати моделі, які враховують складні залежності та підвищують точність прогнозування. Використання різноманітних методів та підходів може допомогти в розв'язанні різних видів задач, від короткострокового прогнозування до довгострокового планування. Контрольоване навчання стає потужним інструментом у сфері прогнозування часових рядів, що дозволяє зробити більш точні та надійні прогнози для підтримки прийняття рішень в різних галузях.

Ковзне вікно (sliding window) є одним із підходів, який використовується при прогнозуванні часових рядів як контрольованого навчання. Цей підхід дозволяє моделі використовувати історичні дані для навчання та прогнозування майбутніх значень ряду. Принцип роботи полягає в тому, що для кожної точки прогнозу модель використовує певну кількість попередніх значень ряду як вхідних даних.

Ковзне вікно можна уявити як віртуальне вікно, яке рухається по часовому ряду, при цьому оновлюючи навчальний набір даних для кожної нової точки прогнозу. Наприклад, якщо ми маємо часовий ряд з даними за останні 12 місяців, і вибираємо ковзне вікно шириною 6 місяців, то для кожної точки прогнозу модель використовує дані з попередніх 6 місяців для навчання та прогнозування.

Використання ковзного вікна має кілька переваг. По-перше, воно дозволяє моделі використовувати більше історичних даних для навчання, що може покращити точність прогнозів. По-друге, воно дозволяє моделі враховувати динаміку та зміну залежностей в часовому ряді, оскільки кожне нове вікно оновлюється з новими даними. Це особливо корисно для рядів з сезонністю або тенденціями, де залежності можуть змінюватись з часом.

Проте, при використанні ковзного вікна також існують деякі виклики. Перш за все, потрібно визначити оптимальну ширину вікна, яка забезпечить належну балансу між збереженням достатньої історичної інформації та забезпеченням достатньої гнучкості для виявлення змін. Крім того, при великих обсягах даних використання ковзного вікна може призвести до складних обчислювальних вимог.

У підсумку, ковзне вікно є корисним підходом для прогнозування часових рядів як контрольованого навчання. Воно дозволяє моделі використовувати історичні дані для прогнозування майбутніх значень та враховувати динаміку ряду. Правильне використання та налаштування ковзного вікна може покращити точність та надійність прогнозів.

Розсувне вікно з багатокроковим прогнозуванням є підходом, який дозволяє прогнозувати часові ряди як контрольоване навчання, використовуючи інформацію з попередніх кроків часу. Цей підхід використовується, коли метою є прогнозування не лише наступного кроку ряду, але й кількох майбутніх значень.

Замість того, щоб використовувати лише одне попереднє значення для прогнозу наступного кроку, розсувне вікно з багатокроковим прогнозуванням використовує декілька попередніх значень як вхідні дані для прогнозування майбутніх значень. Наприклад, якщо ми маємо часовий ряд з місячними даними продажів за останні два роки і хочемо зробити прогноз наступних шести місяців, розсувне вікно може використовувати 12 попередніх значень (дані за попередні 12 місяців) для прогнозування наступних 6 місяців.

Використання розсувного вікна з багатокроковим прогнозуванням має свої переваги. По-перше, воно дозволяє моделі враховувати довгострокові залежності і тренди в часовому ряді, що може покращити точність прогнозів на більш віддалені майбутні кроки. Крім того, такий підхід дозволяє враховувати сезонність та циклічність у даних.

Проте, використання розсувного вікна з багатокроковим прогнозуванням також має свої виклики. При великій кількості майбутніх кроків для прогнозу модель може зіткнутися зі зростаючою невизначеністю та похибками. Крім того, врахування більшої кількості попередніх значень може призвести до більш складної обробки даних та складніших моделей.

Загалом, розсувне вікно з багатокроковим прогнозуванням є потужним підходом для прогнозування часових рядів як контрольованого навчання, оскільки воно дозволяє враховувати залежності в часі та прогнозувати багато майбутніх значень. Однак, використання цього підходу вимагає ретельного аналізу та налаштування моделі, а також обізнаності щодо викликів, пов'язаних з довготривалими прогнозами.

2.3 Параметри моделі XGBoost при прогнозуванні

Параметри моделі XGBoost відіграють ключову роль у прогнозуванні часових рядів. XGBoost є потужним і популярним алгоритмом градієнтного бустінгу, який заснований на ансамблю дерев рішень і здатний до точного та надійного прогнозування. Деякі з параметрів моделі XGBoost мають особливе значення при роботі з часовими рядами і можуть бути налаштовані для досягнення кращих результатів.

- Одним з важливих параметрів є "learning rate" або крок навчання. Цей параметр визначає швидкість збіжності моделі під час навчання. Вибір оптимального значення кроку навчання може бути критичним, оскільки недостатньо велике значення може призвести до повільної збіжності, а занадто велике - до перенавчання моделі. Тому рекомендується експериментувати з різними значеннями та знаходити оптимальну величину для конкретного часового ряду.

  Приріст (Increment) для кожного дерева обчислюється шляхом множення швидкості (learning rate) на прогнозовану величину для кожного листка дерева.

  Припустимо, що ми маємо поточний прогноз моделі для конкретного прикладу після обробки попередніх дерев: F(t-1)(x), де t - номер дерева, x - приклад даних.

  Тоді приріст (increment) для поточного дерева буде:

  Increment(t) = learning_rate * h(t)(x),

де h(t)(x) - прогнозована величина для листка дерева t та прикладу x.

Таким чином, learning rate контролює внесок кожного дерева в загальний прогноз моделі. Менше значення learning rate зменшує внесок кожного дерева, тоді як більше значення learning rate збільшує його внесок.

При великому значенні learning rate може статися перенавчання, оскільки кожне нове дерево намагатиметься повністю скоригувати помилки попередніх дерев. З іншого боку, надто мале значення learning rate може призвести до повільної збіжності моделі, оскільки внесок кожного дерева буде незначним.

Тому, при підборі оптимального значення learning rate, рекомендується експериментувати з різними значеннями та спостерігати за впливом на швидкість збіжності та уникнення перенавчання моделі.

- При побудові дерева рішень у XGBoost, використовується алгоритм розділення, який базується на обчисленні приросту інформації (Information Gain) або виграшу Джині (Gini Gain). Ці метрики оцінюють, як добре розбиття вузла впливає на покращення прогнозної точності моделі.

Максимальна глибина дерева (max_depth) визначає кількість шарів дерева рішень. Чим більше значення max_depth, тим більше вузлів і розбиттів може містити дерево, що дозволяє моделі навчатися складнішим шаблонам у даних. Однак, занадто велика глибина може призвести до перенавчання, коли модель "запам'ятовує" тренувальні дані, включаючи шум та непотрібні особливості, що погіршує її узагальнюючу здатність.

Математично, глибина дерева визначається шляхом обмеження кількості рівнів (L) у дереві. Кожен рівень додає нові вузли та розбиття, що дозволяє моделі навчатися більш складним шаблонам. Зазвичай, максимальна глибина (max_depth) визначається цілим числом, наприклад:

max_depth = L

Таким чином, параметр max_depth в XGBoost контролює складність моделі та рівень деталізації, які можуть бути захоплені деревом. Підбір оптимального значення max_depth відбувається шляхом експериментування та спостереження за впливом на точність та узагальнюючу здатність моделі..

- Параметр У XGBoost, ансамбль складається з декількох дерев рішень, які послідовно додаються для отримання кінцевого прогнозу моделі. Параметр "n_estimators" визначає кількість дерев, які будуть створені в ансамблі.

Приріст прогнозу для кожного нового дерева додається до загального прогнозу моделі. Чим більше дерев, тим складніша і потужніша модель, оскільки вона може взяти до уваги більше шаблонів та взаємодій у даних. Однак, занадто велика кількість дерев може призвести до перенавчання, коли модель надмірно адаптується до тренувальних даних і не узагальнюється на нові дані.

Математично, параметр "n_estimators" визначається як кількість дерев, які будуть додані до ансамблю. Зазвичай, цей параметр встановлюється цілим числом, наприклад:

n_estimators = N,

де N - кількість дерев.

Таким чином, параметр "n_estimators" в XGBoost визначає кількість дерев, які будуть використовуватись у моделі. Підбір оптимального значення "n_estimators" відбувається шляхом експериментування та спостереження за впливом на точність моделі та обчислювальні ресурси.

• У XGBoost, параметр "subsample" визначає частку вибірки даних, яка буде використовуватися для навчання кожного дерева. Це дозволяє використовувати стохастичний градієнтний спуск, де не всі доступні дані використовуються при кожній ітерації.

Параметр "subsample" контролює випадкове вибір підмножини даних для кожного дерева. Зазвичай, значення параметра "subsample" встановлюється в діапазоні від 0.5 до 1.0. Наприклад, якщо значення "subsample" дорівнює 0.8, це означає, що для кожного дерева буде використовуватись випадково обрана 80% вибірка з наявних даних.

Математично, можна представити параметр "subsample" за допомогою виразу:

subsample = f,

де f - частка вибірки даних, яка буде використовуватися.

Таким чином, параметр "subsample" в XGBoost дозволяє контролювати розмір вибірки даних, яка буде використовуватись для навчання кожного дерева. Це допомагає уникнути перенавчання та покращує узагальнюючу здатність моделі. Підбір оптимального значення "subsample" відбувається шляхом експериментування та спостереження за впливом на точність моделі та здатність уникнути перенавчання."subsample" визначає частку вибірки даних, яка буде використовуватися для навчання кожного дерева. Вибір значення менше 1.0 дозволяє використовувати стохастичний градієнтний спуск і запобігає перенавчанню. Зазвичай рекомендується значення в діапазоні від 0.5 до 1.0.

- Параметр "colsample_bytree" в XGBoost визначає частку ознак (стовпців) дерева, яка буде використовуватися для навчання кожного дерева. Це дозволяє контролювати випадковий вибір підмножини ознак для кожного дерева.

Математично, параметр "colsample_bytree" можна представити за допомогою виразу:

colsample_bytree = f,

де f - частка ознак (стовпців) дерева, яка буде використовуватися.

Наприклад, якщо значення "colsample_bytree" дорівнює 0.8, це означає, що для кожного дерева буде використовуватись випадково обрана 80% ознак з наявних даних.

Застосування параметра "colsample_bytree" дозволяє використовувати лише підмножину ознак для кожного дерева, що може допомогти уникнути перенавчання та забезпечити більшу різноманітність дерев у моделі.

Отже, параметр "colsample_bytree" дозволяє контролювати кількість ознак, які будуть використовуватись для навчання кожного дерева у XGBoost. Підбір оптимального значення "colsample_bytree" відбувається шляхом експериментування та спостереження за впливом на точність моделі та уникнення перенавчання.

- Параметр "gamma" в XGBoost визначає мінімальну зміну в градієнті, яка потрібна для створення нового розбиття у дереві. Цей параметр використовується для контролювання зростання моделі шляхом відкидання розбиттів, які не приносять значного покращення.

Математично, параметр "gamma" можна представити за допомогою виразу:

gamma = g,

де g - мінімальна зміна в градієнті для створення нового розбиття.

При використанні алгоритму XGBoost, для кожного розбиття у дереві обчислюється величина, відома як "gain". Вона визначає, наскільки значно покращення принесе нове розбиття. Якщо "gain" менше за значення "gamma", то розбиття не буде виконано, і це допомагає уникнути незначних розбиттів, які не покращують якість моделі.

Застосування параметра "gamma" дозволяє контролювати складність моделі і запобігати перенавчанню. Значення "gamma" визначає поріг зміни градієнту, нижче якого розбиття не виконується.

Отже, параметр "gamma" в XGBoost допомагає контролювати рост моделі шляхом відкидання незначних розбиттів на основі мінімальної зміни градієнту. Підбір оптимального значення "gamma" відбувається шляхом експериментування та спостереження за впливом на точність моделі та здатність уникнути перенавчання.

- Параметр "reg_alpha" в XGBoost використовується для регуляризації моделі шляхом додавання штрафу на суму абсолютних значень ваг ознак. Цей параметр допомагає контролювати складність моделі та уникнути перенавчання.

  Математично, параметр "reg_alpha" можна представити за допомогою виразу:

  reg_alpha = λ,

  де λ - коефіцієнт регуляризації, який визначає силу штрафу на суму абсолютних значень ваг ознак.

  Параметр "reg_alpha" додає штраф до функції втрати, який залежить від суми абсолютних значень ваг ознак. Цей штраф допомагає убезпечити модель від надмірно великих ваг, що може призвести до перенавчання та залежності від конкретних даних.

  Застосування параметра "reg_alpha" дозволяє контролювати рівень регуляризації моделі. Значення "reg_alpha" визначає силу штрафу, де більше значення веде до сильнішої регуляризації.

Отже, параметр "reg_alpha" в XGBoost допомагає контролювати регуляризацію моделі шляхом додавання штрафу на суму абсолютних значень ваг ознак. Підбір оптимального значення "reg_alpha" відбувається шляхом експериментування та спостереження за впливом на точність моделі та здатність уникнути перенавчання.

- Параметр "reg_lambda" в XGBoost використовується для регуляризації моделі шляхом додавання штрафу на суму квадратів ваг ознак. Цей параметр допомагає контролювати складність моделі та уникнути перенавчання.

  Математично, параметр "reg_lambda" можна представити за допомогою виразу:

  reg_lambda = λ,

  де λ - коефіцієнт регуляризації, який визначає силу штрафу на суму квадратів ваг ознак.

  Параметр "reg_lambda" додає штраф до функції втрати, який залежить від суми квадратів ваг ознак. Цей штраф допомагає контролювати величину ваг, зменшує вплив незначних ознак та допомагає уникнути перенавчання.

  Застосування параметра "reg_lambda" дозволяє контролювати рівень регуляризації моделі. Значення "reg_lambda" визначає силу штрафу, де більше значення веде до сильнішої регуляризації.

Отже, параметр "reg_lambda" в XGBoost допомагає контролювати регуляризацію моделі шляхом додавання штрафу на суму квадратів ваг ознак. Підбір оптимального значення "reg_lambda" відбувається шляхом експериментування та спостереження за впливом на точність моделі та здатність уникнути перенавчання.

Оптимальний вибір параметрів визначається експериментуванням та залежить від конкретної задачі та даних. Важливо проводити аналіз і налаштування параметрів для досягнення оптимальних результатів прогнозування часових рядів з використанням моделі XGBoost.

2.4 Висновки другого розділу

Таким чином, цей розділ зосереджувався на глибокому вивченні та використанні моделі XGBoost для прогнозування часових рядів. В рамках цього підрозділу ми представили детальний опис моделі XGBoost, звернули увагу на її особливості та робочі принципи, а також висвітлили її конкретне застосування для прогнозування часових рядів.

Окрім того, ми провели аналіз впливу різних параметрів моделі XGBoost на результати прогнозування. Зокрема, ми досліджували вплив глибини дерева, швидкості навчання, кількості дерев та інших параметрів на точність і якість прогнозів. Це дозволило нам розуміти, як варіювання цих параметрів впливає на ефективність моделі XGBoost при прогнозуванні часових рядів.

Отримані результати і аналіз дозволили нам зробити висновок, що правильний вибір та налаштування параметрів моделі XGBoost має велике значення для досягнення точних та надійних прогнозів часових рядів. Крім того, наші дослідження вказують на потенціал моделі XGBoost як потужного інструменту для прогнозування часових рядів у порівнянні з іншими методами та моделями, що використовуються в цій сфері.

РОЗДІЛ 3. РЕАЛІЗАЦІЯ МЕТОДІВ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ПРОДАЖІВ

3.1 Моделювання, інформаційні та програмні технології

Прогнозування реалізовано за допомогою алгоритму XGBoost. Математична реалізація представлена функцією оптимізації градієнтного бустингу:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

Ключові компоненти алгоритму XGBoost:

- Функція втрат (Loss Function):

Функція втрат вимірює різницю між спостережуваними значеннями (y) і передбаченими значеннями (y_hat). Зазвичай використовується середньоквадратична помилка (Mean Squared Error, MSE). Формула для функції втрат:

Loss(y, y_hat) = 0.5 * (y - y_hat)^2

- Об'єктивна функція (Objective Function):

Об'єктивна функція XGBoost об'єднує функцію втрат та регуляризаційний член. Вона визначається наступним чином:

Objective = Loss(y, y_hat) + λ * Regularization

де:

Loss(y, y_hat) - функція втрат, яка вимірює різницю між спостережуваними значеннями (y) і передбаченими значеннями (y_hat).

λ - параметр регуляризації, який контролює рівень складності моделі.

Regularization - регуляризаційний член, який допомагає уникнути перенавчання та поліпшує узагальнюючу здатність моделі.

- Градієнт та гессіан (Gradient and Hessian):

XGBoost використовує градієнтний бустинг, тому потрібно обчислити градієнт (перший похідний) та гессіан (другий похідний) функції втрат для кожного прикладу. Формули для градієнта та гессіана:

Gradient = ∂Loss(y, y_hat) / ∂y_hat

Hessian = ∂^2Loss(y, y_hat) / ∂(y_hat)^2

- Прогнозування (Prediction):

Прогнозування в XGBoost здійснюється шляхом агрегування прогнозів базових моделей. Загальне прогнозоване значення (y_hat) обчислюється як сума прогнозів базових моделей, помножених на коефіцієнти (learning rate):

y_hat = Σ(base_model_prediction * learning_rate)

Для кожної базової моделі у формулі використовується додатковий член (штрафний член), що залежить від градієнта та гессіана для кожного прикладу. Це дозволяє враховувати розбіжності між прогнозованими значеннями та спостережуваними значеннями.

Для створення інформаційного забезпечення та автоматизованої системи програмування потрібно використовувати різноманітні інформаційні та програмні технології. Деякі з таких технологій включають:

- Мови програмування: Для розробки інформаційного забезпечення та автоматизованих систем програмування можуть бути використані різні мови програмування, наприклад, Python, C++ та інші. Ці мови надають потужні інструменти для обробки даних, реалізації алгоритмів та розробки користувацьких інтерфейсів.

- Бібліотеки для роботи з часовими рядами: Важливо вибрати відповідні бібліотеки, які надають функціонал та інструменти для маніпулювання, аналізу та візуалізації часових рядів. Деякі популярні бібліотеки для роботи з часовими рядами в Python включають Pandas, NumPy, Matplotlib, XGBoost, Tkinter та інші.

- Інтерактивні інструменти та візуалізація: Для зручної роботи з часовими рядами рекомендується використовувати інтерактивні інструменти та засоби візуалізації. Jupyter Notebook або JupyterLab забезпечують інтерактивне середовище для аналізу даних та створення звітів з кодом, текстом і графіками.

- Середовище розробки: Рекомендується використовувати зручне середовище розробки, яке надає можливості для зручної роботи з кодом, налагодження та керування проектом. Одним з популярних середовищ розробки для Python є Visual Studio, яка має широкий спектр функцій та інструментів для розробки та налагодження програм. Також рекомендується використовувати хмарні рішення, наприклад, Azure.

Важливо враховувати особливості завдання і вимоги проєкту під час вибору інформаційних і програмних технологій. Комбінація Python, бібліотек для роботи з часовими рядами, Visual Studio як середовища розроблення та інтерактивних інструментів, як-от Jupyter Notebook, забезпечить ефективне розроблення та реалізацію інформаційного забезпечення.

3.2 Вхідні дані проекту аналізу та прогнозування продажів

У цьому розділі буде представлено розроблену систему прогнозування продажів з використанням моделі XGBoost. Система надає можливість автоматизовано прогнозувати майбутні продажі на основі вхідних даних.

Перш за все, будуть завантажені тестові дані, що включають історичні дані про продажі протягом певного періоду часу. Далі, будуть застосовані різні методи прогнозування, доступні у системі, зокрема модель XGBoost.

Метод буде застосовано до тестових даних, і результати прогнозів будуть візуалізовані у вигляді графіків. Це дозволить наочно оцінити ефективність та точність роботи системи.

У результаті експериментів та аналізу буде продемонстровано, що розроблена система здатна ефективно прогнозувати майбутні продажі за допомогою різних методів, використовуючи модель XGBoost. Цей функціонал надає підприємствам і організаціям засоби для прийняття обґрунтованих рішень на основі надійних прогнозів, що є важливим інструментом управління та планування в сфері бізнесу.

Дані для подальшого аналізу включають наступні таблиці:

Таблиця "Продажі": Ця таблиця містить дані про продажі, такі як дата продажу, кількість проданих одиниць, ціна продажу, ідентифікатор продукту або послуги, ідентифікатор кастомера та інші відповідні атрибути. Вона відображає історичні дані про продажі компанії.

Таблиця "Постачальники": Ця таблиця містить дані про постачальників, такі як їх ідентифікатор, назву, контактну інформацію, адресу тощо. Вона використовується для зберігання інформації про постачальників, які постачають матеріали або компоненти для компанії.

Таблиця "Матеріали": Ця таблиця містить дані про матеріали, які використовуються в виробництві або продуктах компанії. Кожен матеріал має свій ідентифікатор, назву, опис, вартість, посилання на постачальника і інші відповідні атрибути.

Таблиця "Компоненти": Ця таблиця містить дані про компоненти, з яких складаються матеріали. Кожен компонент має свій ідентифікатор, назву, опис, посилання на матеріал, до якого він належить, і інші відповідні атрибути.

Таблиця "Клієнти": Ця таблиця містить дані про клієнтів, які придбали продукти компанії. Кожен клієнт має свій ідентифікатор, назву, контактну інформацію, адресу тощо.

Таблиця "Зв'язки": Ця таблиця містить зв'язки між різними таблицями. Наприклад, вона може містити ідентифікатори продукту та матеріалу, щоб показати, які матеріали використовуються в конкретному продукті.

На основі таблиці "Продажі" ми виконаємо прогнозування продуктів. Система дає можливість аналізувати історичні дані продажів продуктів і використовувати їх для прогнозування майбутнього попиту на ці матеріали. Наприклад, використовуючи моделі прогнозування, такі як XGBoost або інші методи, можна передбачити, які матеріали будуть потрібні в майбутньому на основі попиту на продукти або послуги, що засновані на даних про продажі. Такий прогноз може допомогти в управлінні запасами, плануванні виробництва та забезпеченні належного постачання матеріалів для виробництва.

Однак, прогноз можна зробити на основі будь-яких даних, залежно від конкретної потреби. Наприклад, система дозволяє прогнозувати попит на певні товари зі сторони клієнта.

3.3 Система прогнозування продажів для оптимізації портфоліо

Для демонстрації роботи системи прогнозування на основі продажів, ми використовували дані про окремий продукт протягом 5 років, зібрані для кожного місяця. Цей часовий ряд представляється на рисунку 3.1.

Цей часовий ряд продажів окремого продукту протягом п'ятирічного періоду має виражену сезонність, що означає, що продажі показують повторюваність або регулярні коливання в певних місяцях. Це може бути пов'язано з факторами, такими як сезонність попиту на продукт, святкові періоди або інші циклічні впливи.

Зауважимо, що в цьому часовому ряді відсутній загальний тренд між роками. Тобто, немає чіткого зростання або спаду продажів з року в рік. Продажі змінюються на більш менш випадковий спосіб, зберігаючи сезонність в межах кожного року.

Ця інформація про сезонність та відсутність загального тренду є важливою при аналізі та прогнозуванні майбутніх продажів на основі цього часового ряду.

Наша програма використовує ці дані для аналізу та прогнозування майбутнього попиту на цей продукт.

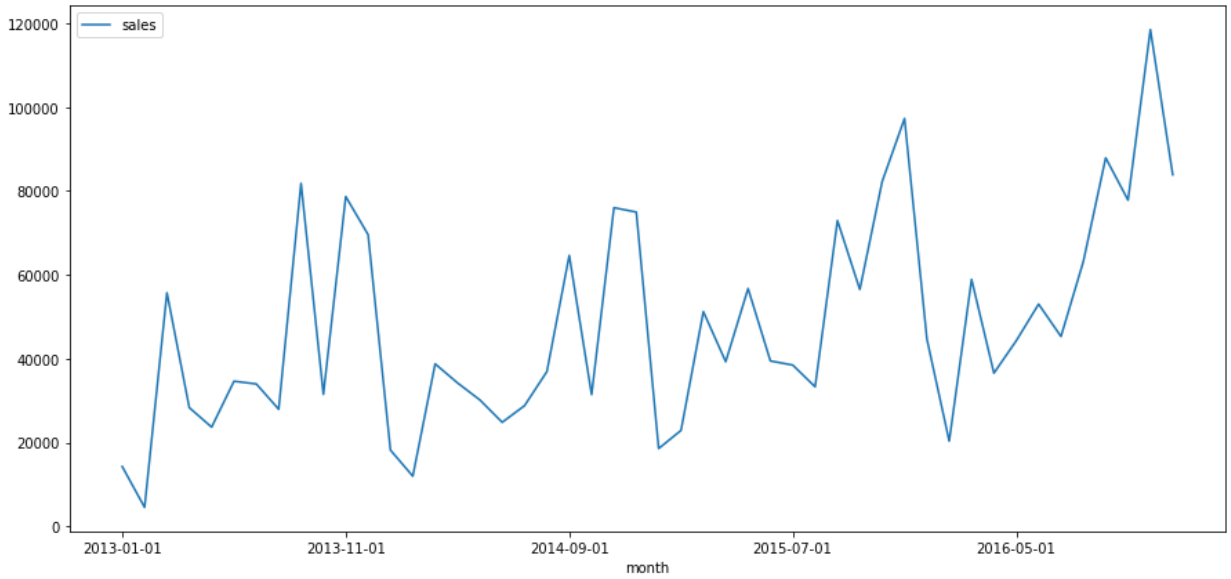


Рисунок 3.1 - Візуалізація вхідного часового ряду

Рисунок 3.2 демонструє інтерфейс програми, який забезпечує зручну взаємодію з користувачем під час проведення аналізу. Цей базовий інтерфейс містить ряд кнопок, що спрощують роботу з програмою та проведення аналізу даних. Користувач може використовувати такі кнопки:

- "Open file" - ця кнопка дозволяє вибрати файл з даними для подальшого аналізу. Користувач може обрати потрібний файл, що містить дані про продажі, і програма використає ці дані для прогнозування.
- "Run analysis" - натискання на цю кнопку запускає процес прогнозування продажів на основі вибраних даних. Після натискання програма аналізує дані і надає користувачеві прогнозні результати.
- "Exit" - ця кнопка дозволяє вийти з програми після завершення аналізу та прогнозування.
- "Get visualization" - натискання на цю кнопку дозволяє отримати візуалізацію поточних продажів або прогнозованих значень. Користувач

може побачити графіки, діаграми або інші візуальні зображення, які надають зрозумілу візуальну інформацію про дані.

- "Get Dates" - ця кнопка дозволяє користувачеві вибрати дати для проведення аналізу. Користувач може вказати потрібний період часу, для якого він бажає отримати прогноз.
- "Select file separator" - ця кнопка дозволяє користувачу вибрати роздільник, який використовується в файлі з даними. Вибір правильного роздільника допомагає програмі правильно інтерпретувати дані з файлу.
- "Select pipeline" - ця кнопка дозволяє користувачу вибрати тип пайплайну для аналізу та прогнозування. Користувач може вибрати "classic" для використання наявних попередньо натренованих моделей для прогнозування або "New Platform", який включає в себе агрегацію даних з різних таблиць і вимагає додаткових виборів.

Ці кнопки та інтерфейс допомагають забезпечити зручну та ефективну роботу з програмою, а також дозволяють користувачеві налаштувати аналіз та прогнозування відповідно до його потреб.

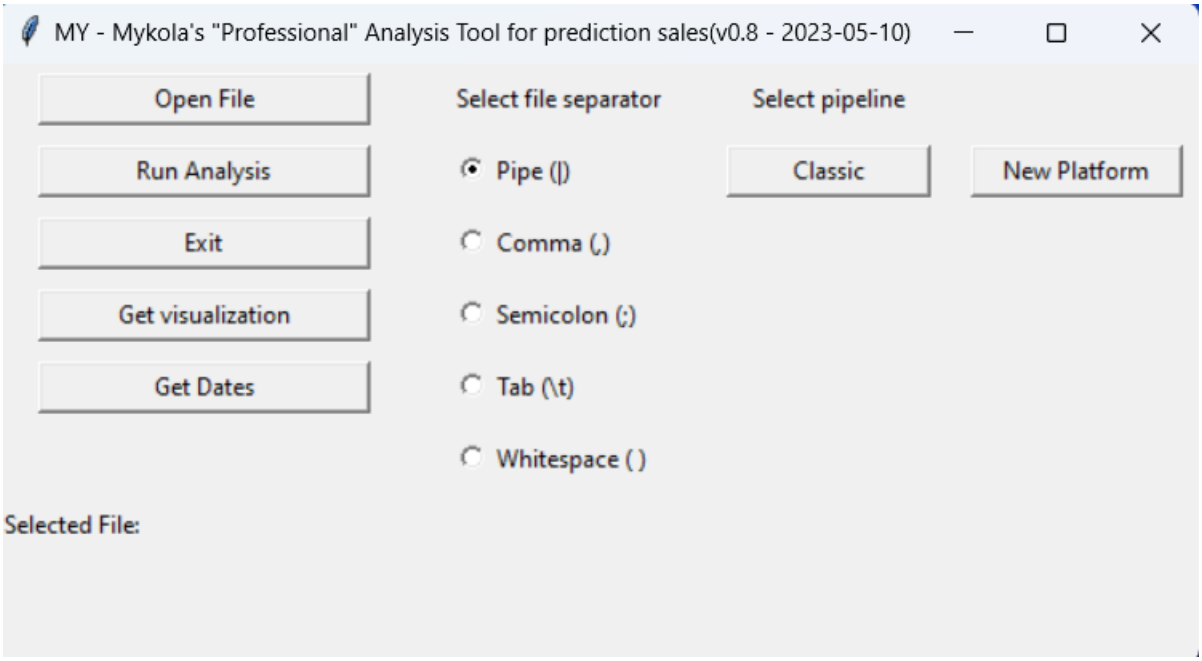


Рисунок 3.2 – Інтерфейс користувача для моделювання

На рисунку 3.3 представлена візуалізація, яка демонструє якість роботи алгоритму прогнозування продажів. Графік відображає існуючі дані про продажі протягом останнього року, а також прогнозовані значення на цей період.

Цей графік надає користувачеві можливість порівняти фактичні продажі з прогнозованими значеннями та оцінити точність алгоритму прогнозування. Якщо прогнозовані значення добре узгоджуються з фактичними даними, графік візуалізує близькість прогнозу до реальності. Це може служити підтвердженням ефективності алгоритму та його здатності до надійних прогнозів продажів.

Така візуалізація дозволяє зрозуміти, наскільки точними є прогнози та чи можна їм довіряти. Вона є важливим інструментом для аналізу результатів прогнозування та прийняття обґрунтованих рішень на основі цих прогнозів.

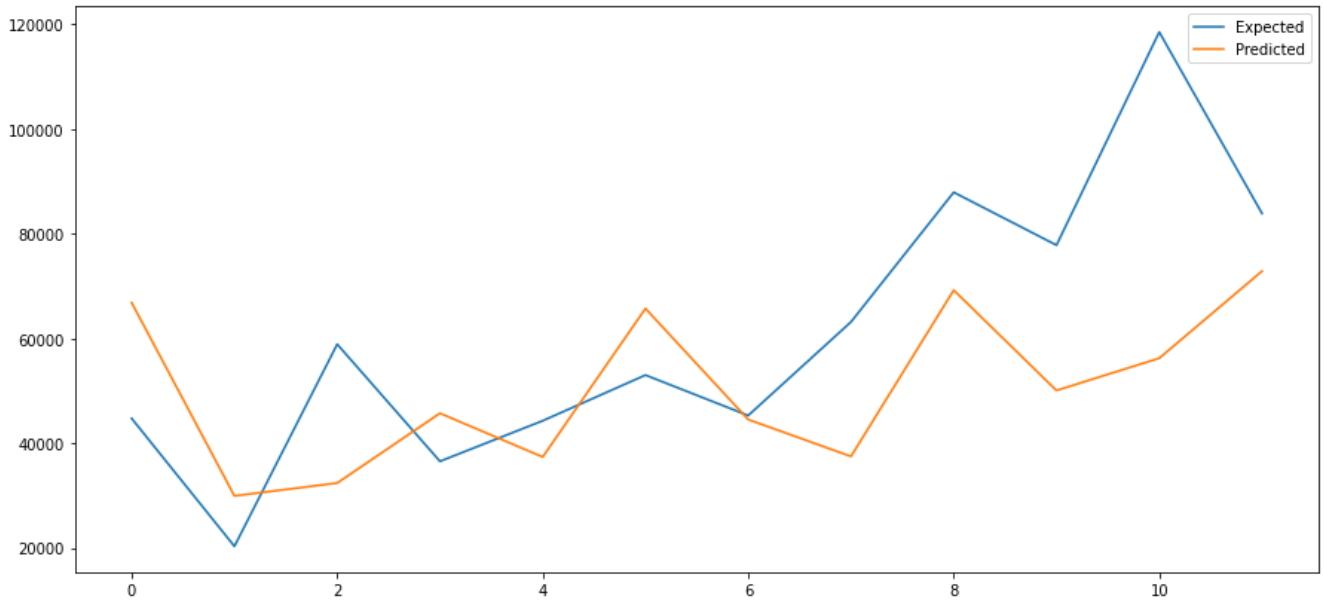


Рисунок 3.3 – Результати моделювання

На рисунку 3.4 показана таблиця, яка надає числовий варіант прогнозу продажів та реальних значень продажів для тестового періоду. Кожен рядок таблиці відповідає конкретному місяцю в тестовому періоді, і містить прогнозоване значення продажів і фактичне значення продажів.

Крім того, таблиця також включає значення метрики MAE (Mean Absolute Error) для тестового періоду. MAE є однією з найпоширеніших метрик оцінки точності прогнозів. Вона вимірює середню абсолютну різницю між прогнозованими значеннями і фактичними значеннями. Чим менше значення MAE, тим ближче прогнози до фактичних значень і тим вища точність прогнозування.

Аналізуючи таблицю та значення MAE, користувач може оцінити, наскільки добре прогнози відповідають реальним значенням продажів. Більш низьке значення MAE вказує на вищу точність прогнозування і показує, що алгоритм має добрі прогнозні можливості для даного часового ряду.

```
>expected=44703.1, predicted=66791.8
>expected=20301.1, predicted=29922.3
>expected=58872.3, predicted=32373.3
>expected=36521.5, predicted=45695.9
>expected=44261.1, predicted=37354.4
>expected=52981.7, predicted=65679.2
>expected=45264.4, predicted=44504.9
>expected=63120.9, predicted=37458.6
>expected=87866.6, predicted=69181.9
>expected=77776.9, predicted=50043.8
>expected=118447.8, predicted=56196.9
>expected=83829.3, predicted=72813.1
MAE: 19424.516
```

Рисунок 3.4 – Результати моделювання

3.4 Висновки третього розділу

У тексті представлена розроблена система прогнозування продажів з використанням моделі XGBoost. Система надає можливість автоматизовано прогнозувати майбутні продажі на основі вхідних даних. Користувач системи може вибрати бажаний метод прогнозування, і система забезпечить прогноз на основі цього методу. Розроблена система демонструє ефективність та точність прогнозування майбутніх продажів за допомогою різних методів, використовуючи модель XGBoost. Прогноз продажів може бути корисним для управління запасами, планування виробництва та забезпечення належного постачання матеріалів для виробництва. Система також може прогнозувати попит на певні товари зі сторони клієнта. Візуалізація результатів прогнозування дозволяє оцінити точність прогнозів та прийняти обґрунтовані рішення на основі цих прогнозів. Використання метрики MAE допомагає оцінити точність прогнозів шляхом порівняння прогнозованих значень з фактичними значеннями продажів. Система представлена у вигляді зручного інтерфейсу для користувача.

ВИСНОВКИ

В рамках даної кваліфікаційної роботи було створену систему для прогнозування продажів, яка є ключовим елементом оптимізації портфоліо компанії. Важливо, що було досліджено, що використання ансамблю моделей та інформаційних технологій, таких як автоматизований підбір гіперпараметрів, може покращити якість прогнозів та ефективність оптимізації портфоліо на основі аналізу часових рядів. Ансамбль моделей дозволяє поєднати прогнози з різних моделей для отримання більш об'єктивних результатів, а автоматизований підбір гіперпараметрів спрощує процес аналізу та прогнозування часових рядів шляхом визначення оптимальних значень гіперпараметрів моделі. Застосування цих методів може привести до досягнення кращих результатів в області прогнозування та оптимізації портфоліо.

Крім того, у магістерській роботі було проведено глибоке дослідження моделі XGBoost для прогнозування часових рядів. Виявлено, що правильний вибір та налаштування параметрів моделі мають велике значення для отримання точних та надійних прогнозів. Дослідження показали потенціал моделі XGBoost як потужного інструменту для прогнозування часових рядів.

У роботі також була розроблена система прогнозування продажів з використанням моделі XGBoost. Ця система демонструє ефективність та точність прогнозування майбутніх продажів і може бути корисною для оптимізації портлоліо. Система також може прогнозувати попит на товари зі сторони клієнтів. Використання метрики MAE дозволяє оцінити точність прогнозів, а візуалізація результатів сприяє прийняттю обґрунтованих рішень на основі цих прогнозів. Розроблена система має зручний інтерфейс для користувача, що спрощує використання.

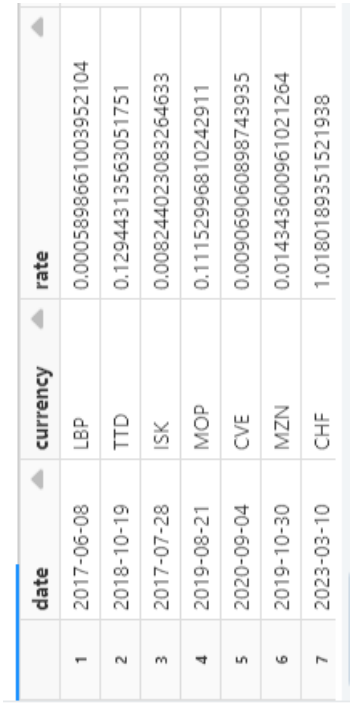
Отже, магістерська робота зосереджувалась на використанні ансамблю моделей та інформаційних технологій для покращення прогнозів та оптимізації портфоліо на основі часових рядів. Дослідження моделі XGBoost підтвердили її потенціал як ефективного інструменту для прогнозування часових рядів. Розроблена система прогнозування продажів з використанням моделі XGBoost демонструє точність та можливості прогнозування. Загалом, робота пропонує цінний внесок у розвиток області прогнозування та оптимізації портфоліо на основі часових рядів.

СПИСОК ВИКОРИСТАНИХ ІНФОРМАЦІЙНИХ ДЖЕРЕЛ

1. Zhang, J., Shi, T., King, I., & Yeung, D. Y. (2017). Deep learning for time series modeling: A comparative review
2. Sales Analytics: The Essential Guide to Using Data and Analytics to Improve Your Sales Performance" ,Cesar Brea
3. Khemphila, A., Sriyakul, T., & Jaroenwattana, A. (2019). Time series forecasting with xgboost in retail business. In 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 106-111)
4. "Selling with Analytics: The New Era of Predictive Sales" , Alexander Broun
5. "Practical Time Series Analysis: Prediction with Statistics and Machine Learning" ,Aileen Nielsen & Jason Brownlee
6. Вебсайт. URL: https://machinelearningmastery.com/xgboost-for-time-series-forecasting
7. Вебсайт. URL: https://365datascience.com/tutorials/python-tutorials/xgboost-lgbm/
8. Вебсайт. URL: https://forecastegy.com/posts/multiple-time-series-forecasting-with-xgboost-in-python/
9. Вебсайт. URL: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

ДОДАТОК

Додаток А. Огляд даних



	date	currency	rate
1	2017-06-08	LBP	0.00058986610039521 04
2	2018-10-19	TTD	0.12944313563051751
3	2017-07-28	ISK	0.00824402308326463 3
4	2019-08-21	MOP	0.11152996810242911
5	2020-09-04	CVE	0.00906906089874393 5
6	2019-10-30	MZN	0.01434360096102126 4
7	2023-03-10	CHF	1.01801893515219 38

Таблиця обміну валют

	id	productType	gearboxDesign	gearboxVariant	driveType	transmission	stageNumber	size	specification	lastDelivery	phase	name
1	10001093	ADP	NA	NA	NA	NA	NA	NA	S	9999-01-01	NA	ADP LP m045D 100 06 081 0475 S
2	10003030	ADP	NA	NA	NA	NA	NA	NA	S	9999-01-01	NA	ADP LP m066G 165 10 131 0425 S
3	10007506	ADP	NA	NA	NA	NA	NA	NA	S	9999-01-01	NA	ADP LP m045D 149 08 116 0478 S
4	10008376	DSH	NA	NA	NA	NA	NA	NA	NA	9999-01-01	NA	DSH LP 13.000-D S
5	10010016	ADP	NA	NA	NA	NA	NA	NA	S	9999-01-01	NA	ADP LP m030C 070 04 052 0345 S
6	10011825	LP	O	M	4	5	1	090	NA	2020-04-01	5	NA
7	10017575	LP	O	E	1	15	2	090	-	2024-05-31	2	LP 090-E02-15 -1X1-000

Таблиця продуктів

	id	name	materialType	materialClass	materialStatus	materialSalesStatus	whenCreated	family
1	10001726	Umlaufträger SP100	ZMAT	1101	99	VZ	1996-12-28	N/A
2	10034263	Ring LPB 120 1-stufig	ZMAT	150301	D	VM	2014-11-14	N/A
3	10040353	Gehaeuse LK 050 GK	ZKAU	1102	SZ	P1	2016-01-20	N/A
4	10045064	Fett PARALIQ GA 351	ZKAU	090101	S	VS	2017-03-09	N/A
5	10046050	Ring VZ NP m100 Z108 VD	ZMAT	11040203	S	VM	2017-05-30	N/A
6	10051086	Rd 50x115 EN10060-1.4104	ZKAU	140102	S	VS	2018-08-02	N/A
7	110917	Dichtsatz Ø63 mm Matra	ZERS	IDD	NA	NA	2018-03-28	N/A

Таблиця компонентів

	id	name
1	LPK	LPK
2	premo Plattform	premo Plattform
3	HG+	HG+
4	NVH	NVH
5	VDT+	VDT+
6	LPB+	LPB+
7	NPR	NPR

Таблиця категорії продуктів

	VBAP_VBELN	VBAP_MATNR	VBAP_NETWR	VBAP_KWMENG	VBAP_WERKS	VBAP_cost	VBAK_KUNNR	VBAK_VKORG	VBAK_AUDAT	VBAK_VTWEG	KNA1_KUNNR	KNA1_LAND1	VBAP_POSNR
1	0000000012	20053683	188.92	2	#	37.67	11016973	2900	2020-10-06	00	11016973	TW	1
2	0000000013	10017569	1897.59	4	#	1128.03	11002355	2900	2020-10-07	00	11002355	TW	1
3	0000000013	10032789	770.89	2	#	340.72	11002355	2900	2020-10-07	00	11002355	TW	2
4	0000000016	10017569	28665.58	60	#	16920.56	11002355	2900	2020-10-29	00	11002355	TW	1
5	0000000016	10032789	11643.4	30	#	5110.79	11002355	2900	2020-10-29	00	11002355	TW	2
6	0000000018	10049207	861	2	#	306.54	11002355	2900	2020-11-12	00	11002355	TW	1
7	0000000018	20083609	985.68	10	#	647.96	11002355	2900	2020-11-12	00	11002355	TW	2

Таблиця продажів

	from_id	to_id	revenue	sales_quantity	cost	dateOrder
1	0000000012_11016973_20201006	20053683	188.92	2	37.67	2020-10-06
2	0000000013_11002355_20201007	10017569	1897.59	4	1128.03	2020-10-07
3	0000000013_11002355_20201007	10032789	770.89	2	340.72	2020-10-07
4	0000000016_11002355_20201029	10017569	28665.58	60	16920.56	2020-10-29
5	0000000016_11002355_20201029	10032789	11645.4	30	5110.79	2020-10-29
6	0000000018_11002355_20201112	10049207	861	2	306.54	2020-11-12
7	0000000018_11002355_20201112	20083609	985.68	10	647.96	2020-11-12

Таблиця замовлень

	from_id	to_id
1	11007774	0000000030_11007774_20201203
2	11002439	0000003453_11002439_20201104
3	11004328	0000003571_11004328_20210128
4	10012120	0000044938_10012120_20190618
5	10012142	0000044967_10012142_20190620
6	10014781	0000045447_10014781_20190809
7	11013401	000004726_11013401_20220929

Таблиця виключень

	id	plant	inventoryValueCost	inventoryValueQuantity	date
1	500115730-00-0	0200	0	0	2022-09-01
2	10042554	0500	0	0	2022-07-01
3	20062534	0800	0	0	2021-08-01
4	10061812	0200	0	0	2022-08-01
5	40000268	0600	0	0	2020-02-01
6	20090645	0200	0	0	2022-09-01
7	20098197	0800	0	0	2022-10-01

Таблиця запасів

	id	inventoryValueQuantityAvg12m	inventoryValueCostAvg12m
1	10033089_01	4	472.12
2	20089257	0	0
3	A53-0003018	9.88	2086.04
4	20003284	12691.81	125.92
5	20036040	0.17	13.39
6	10025810	14.22	310.19
7	20076012	0	0

Таблиця запасів агрегована

	id	alternative	type	production_site
1	00194947	01	M	NA
2	00205835	01	M	NA
3	00209818	01	M	NA
4	00210912	01	M	NA
5	00220633	01	M	NA
6	00223234	01	M	NA
7	00223652	01	M	NA

Таблиця складу продуктів

	from_id	to_id
1	20026306	00198439
2	20061846	00257591
3	20073054	00232070
4	40022229	00197088
5	10047439	00186208
6	20088416	00168391
7	20089293	00173619

Таблиця належності продуктів

	from_id	to_id
1	00191898	10049418
2	00209864	20032545
3	00278859	20043383
4	00211728	20044217
5	00234996	20000110
6	00169563	20044528
7	00255691	40000458

Таблиця списку продуктів

	MATNR	plantId	procurementType
1	20075589	NA	NB
2	20013859	NA	NB
3	40002382	NA	NB
4	20024964	NA	NB
5	20013881	NA	NB
6	20058312	NA	NB
7	10000475	NA	NB

Таблиця закупівлі