МАГІСТЕРСЬКА РОБОТА

МР. ІПм - 13.00.00.000 ПЗ

Група ІПм-24-3

Симчич Андрій

2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Симчич Андрій Мирославович

(прізвище, ім'я, по батькові)

УДК 004.9

(індекс)

МАГІСТЕРСЬКА РОБОТА

Моделі та методи прогнозування продажів на основі технологій машинного навчання

(назва роботи)

Інженерія програмного забезпечення

(назва освітньої програми)

121 - Інженерія програмного забезпечення

(шифр і назва спеціальності)

Симчич А.М.

(підпис, ініціали та прізвище здобувача освітнього ступеня)

Науковий керівник Шекета Василь Іванович, д.т.н., професор

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Допущено до захисту

Завідувач кафедри

доц. Бандура В.В.

(посада) (підпис) (дата) (ініціали та прізвище)

Нормоконтроль

доц. Вовк Р.Б.

(посада) (підпис) (дата) (ініціали та прізвище)

Робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

Івано-Франківськ – 2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Освітній рівень магістр

Спеціальність 121 – Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:
Зав. кафедрою **ІПЗ**

доц. В.В. Бандура

“ 04 ” вересня 2025 р.

ЗАВДАННЯ

НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Симчичу Андрію Мирославовичу

(прізвище, ім’я, по-батькові)

1. Тема магістерської роботи “ Моделі та методи прогнозування продажів на основі технологій машинного навчання”

керівник проекту (роботи) Шекета Василь Іванович, д.т.н., професор

затверджені наказом закладу вищої освіти від “ 05 ” листопада 2025 р. № 695/7

2. Строк подання студентом проекту (роботи) 15 грудня 2025 р.

3. Вихідні дані до проекту (роботи) Теоретичні концепції та формальні моделі побудови та функціонування інформаційних технологій машинного навчання

4. Зміст розрахунково - пояснювальної записки(перелік питань, які потрібно розробити)

1. Дослідження предметної області використання машинного навчання для продажу товарів
2. Моделі та архітектура системи прогнозування продажів на основі машинного навчання
3. Програмна реалізація методів прогнозування продажів на основі машинного навчання
4. Представлення архітектура та опис інформаційної панелі прогнозування продажів

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень)

1. Парадигми машинного навчання (рис. 1.1)
2. Загальна схема (High-level overview) представленої інтелектуальної системи (рис. 1.2)
3. Система прогнозування вартості на основі алгоритму Gradient Boost (рис. 1.3)
4. Алгоритм легковагового ансамблевого навчання для прогнозування продажів (рис. 1.4)
5. Дворівнева статистична модель (рис. 1.5)

6. Консультанти розділів проекту (роботи)

Розділ	Консультант	Підпис, дата
Перевірка на плагіат	доц., к.т.н. Вовк Р.Б.	

7. Дата видачі завдання 04 вересня 2025 р.

Керівник _____
(підпис)

Завдання прийняв до виконання_____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Назви етапів магістерської роботи	Строк виконання етапів роботи	Примітка
1	Підбір і вивчення літератури по темі магістерської роботи	15.09.2025	виконано
2	Дослідження предметної області використання машинного навчання для продажу товарів	29.09.2025	виконано
3	Моделі та архітектура системи прогнозування продажів на основі машинного навчання	17.10.2025	виконано
4	Програмна реалізація методів прогнозування продажів на основі машинного навчання	08.11.2025	виконано
5	Представлення архітектура та опис інформаційної панелі прогнозування продажів	15.11.2025	виконано
6	Затвердження пояснювальної записки роботи завідувачем кафедри	15.12.2025	виконано

Студент – магістр _____
(підпис)

Керівник роботи _____
(підпис)

АНОТАЦІЯ

Магістерська робота: 81 с., 36 рис., 5 табл., 36 джерел.

Тема: Моделі та методи прогнозування продажів на основі технологій машинного навчання

Мета роботи - розроблення моделей, методів та програмної системи прогнозування продажів на основі технологій машинного навчання.

Об'єкт дослідження - процес прогнозування продажів у системах бізнес-аналітики з використанням алгоритмів машинного навчання.

Предметом дослідження - моделі, методи та алгоритми машинного навчання, а також архітектурні рішення програмних систем, що застосовуються для прогнозування продажів.

Результати дослідження

В роботі запропонуванні архітектури системи прогнозування, побудованої відповідно до процесної моделі CRISP-DM, що забезпечує інтеграцію прогнозних моделей у бізнес-аналітичне середовище.

Висновок

Розроблено методику прогнозування продажів на основі ансамблевих моделей машинного навчання з урахуванням особливостей бізнес-середовища, а також створено інтерактивну систему візуалізації результатів прогнозування

МАШИННЕ НАВЧАННЯ; ПРОГНОЗУВАННЯ ПРОДАЖІВ; БІЗНЕС-АНАЛІТИКА; АНСАМБЛЕВІ МЕТОДИ; РЕГРЕСІЙНЕ МОДЕЛЮВАННЯ; CRISP-DM; АНАЛІТИКА ДАНИХ; ІНФОРМАЦІЙНА ПАНЕЛЬ.

ABSTRACT

Master Thesis: 81 pp., 36 fig., 5 tab., 36 sources.

Topic: Sales forecasting models and methods based on machine learning technologies

The purpose of the work is to develop models, methods and a software system for sales forecasting based on machine learning technologies.

The object of the study is the process of sales forecasting in business analytics systems using machine learning algorithms.

The subject of the study is machine learning models, methods and algorithms, as well as architectural solutions of software systems used for sales forecasting.

Research results

The work proposes an architecture of a forecasting system built according to the CRISP-DM process model, which ensures the integration of forecasting models into the business analytics environment.

Conclusion

A sales forecasting methodology based on ensemble machine learning models has been developed, taking into account the peculiarities of the business environment, and an interactive system for visualizing forecasting results has been created

MACHINE LEARNING; SALES FORECASTING; BUSINESS ANALYTICS; ENSEMBLE METHODS; REGRESSION MODELING; CRISP-DM; DATA ANALYTICS; DASHBOARD.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

BDA - Big Data Analytics

CART - Classification and Regression Tree Algorithm

CV - Cross Validation

DT - Decision Tree

GBDT - Gradient-boosted decision tree

MSE - Root Mean Squared Error

RMSE - Root Mean Squared Error

SCM - Supply Chain Management

XGBOOST - eXtreme Gradient Boost

ВСТУП

Актуальність теми.

У сучасних умовах цифрової трансформації бізнесу процеси прийняття управлінських рішень дедалі більше ґрунтуються на обробці великих обсягів даних. Прогнозування продажів є одним із ключових напрямів бізнес-аналітики, адже від точності прогнозів залежить ефективність виробничого планування, управління запасами, фінансових потоків і маркетингових стратегій підприємства. Зростання складності ринкових процесів, динаміка споживчого попиту та багатофакторність впливів зумовлюють потребу у використанні інтелектуальних технологій, здатних виявляти закономірності у великих обсягах даних і забезпечувати адаптивне моделювання.

Методи машинного навчання останніми роками продемонстрували значний потенціал у вирішенні задач прогнозування в економіці, фінансах та торгівлі. Їх застосування дає змогу формувати точні прогностичні моделі, які враховують нелінійні взаємозв’язки, сезонність, трендові компоненти та приховані залежності. Проте ефективне використання таких технологій потребує комплексного підходу — від якісної підготовки даних і вибору оптимальних алгоритмів до побудови архітектури програмної системи, здатної інтегрувати прогнозні моделі в реальні бізнес-процеси.

У межах магістерської роботи досліджено сучасні моделі та алгоритми машинного навчання, проаналізовано їх ефективність у задачах прогнозування продажів і розроблено програмну систему, що реалізує практичну імплементацію розробленої методики. Результати дослідження спрямовані на підвищення точності прогнозів, зниження ризиків прийняття управлінських рішень і підвищення ефективності використання аналітичних ресурсів підприємства.

Актуальність теми зумовлена швидким розвитком цифрової економіки, у межах якої дані стали одним із ключових стратегічних ресурсів підприємства. У сучасних умовах конкуренції точність прогнозування

продажів безпосередньо впливає на прибутковість бізнесу, рівень обслуговування клієнтів та ефективність управління логістичними і маркетинговими процесами. Традиційні статистичні методи прогнозування, які базуються на лінійних припущеннях, дедалі частіше виявляються неефективними у середовищах із великою кількістю взаємопов'язаних факторів і динамічними змінами попиту.

Застосування технологій машинного навчання дозволяє підвищити точність прогнозів за рахунок автоматичного виявлення закономірностей у даних, самоадаптації моделей до нових умов і врахування складних нелінійних залежностей. Інтелектуальні методи, такі як дерева рішень, випадкові ліси, градієнтний бустинг та нейронні мережі, відкривають нові можливості для побудови гнучких систем прогнозування, інтегрованих у бізнес-аналітичні платформи.

Крім того, актуальність теми зумовлена потребою українських підприємств у впровадженні інноваційних цифрових інструментів, які підвищують ефективність планування продажів, оптимізують запаси та забезпечують адаптивне реагування на зміни ринкової ситуації. Таким чином, дослідження моделей і методів прогнозування продажів на основі машинного навчання є важливим і своєчасним завданням як із наукової, так і з практичної точки зору.

Метою магістерської роботи є розроблення моделей, методів та програмної системи прогнозування продажів на основі технологій машинного навчання.

Об'єктом дослідження є процес прогнозування продажів у системах бізнес-аналітики з використанням алгоритмів машинного навчання.

Предметом дослідження є моделі, методи та алгоритми машинного навчання, а також архітектурні рішення програмних систем, що застосовуються для прогнозування продажів.

Для досягнення поставленої мети у роботі вирішено такі **основні завдання**:

1. Проаналізувати сучасний стан проблеми прогнозування продажів і визначити напрями використання технологій машинного навчання у цій сфері.

2. Дослідити теоретичні основи та класифікацію методів машинного навчання, застосовних до прогнозних задач.

3. Розробити методику побудови моделей прогнозування, що враховує етапи підготовки даних, вибору алгоритмів і оцінки ефективності.

4. Реалізувати архітектуру системи прогнозування продажів на основі CRISP-DM-підходу.

5. Розробити програмний прототип системи та створити інформаційну панель для візуалізації результатів прогнозів.

У роботі використано комплекс методів, серед яких:

- аналітичні методи — для огляду наукових джерел і визначення сучасних тенденцій у сфері машинного навчання;

- методи математичного моделювання — для побудови регресійних і ансамблевих моделей прогнозування;

- методи статистичного аналізу — для оцінки точності моделей і перевірки гіпотез;

- алгоритмічні методи машинного навчання — для реалізації моделей лінійної регресії, дерева рішень, Random Forest і XGBoost;

- методи програмної інженерії — для проєктування архітектури системи, обробки даних і створення інтерфейсу користувача.

Наукова новизна роботи полягає у розробленні узагальненої методики прогнозування продажів на основі ансамблевих моделей машинного навчання, що враховує етапи підготовки даних, вибору алгоритмів та оцінки результатів.

Практичне застосування результатів

Розроблену методику та створений програмний прототип можна застосовувати в аналітичних системах підприємств роздрібної торгівлі, електронної комерції, маркетингу та логістики. Реалізовані алгоритми

прогнозування дозволяють підвищити точність оцінки попиту, оптимізувати рівень запасів і планування продажів, а також забезпечують швидку адаптацію до змін ринкових умов.

Структура магістерської роботи. Робота складається зі вступу, трьох розділів та висновків. Загальний обсяг роботи становить 81 сторінку, і містить 36 рисунків, 5 таблиць, список використаних джерел із 36 найменувань.

РОЗДІЛ 1. ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ ВИКОРИСТАННЯ ТЕХНІК МАШИННОГО НАВЧАННЯ ДЛЯ ПРОДАЖУ ТОВАРІВ

1.1. Актуальність та застосування методів машинного навчання в прогнозуванні продажів

Застосування методів машинного навчання (МН) та інтелектуального аналізу даних (ІАД) набуває критичної значущості в сучасному технологічному ландшафті. Інноваційний потенціал цих підходів забезпечує суттєві переваги в широкому спектрі секторів, включаючи освіту, охорону здоров'я, інженерію, торгівлю, індустрію розваг та транспорт. Систематична екстракція цінної інформації з масивів даних стала ключовим фактором конкурентоспроможності в умовах глобалізованої економіки. Це вимагає розробки та імплементації високоточних прогностичних моделей. Експоненційне зростання обсягів даних, що генеруються комерційними транзакціями, створює значні виклики для бізнес-середовища щодо вибору точних методик та розробки ефективних стратегій прогнозування.

1.1.1. Проблематика та необхідність інноваційних підходів

Традиційні методи досягнення цілей у сфері продажів та маркетингу є недостатніми для підтримки конкурентоспроможності компаній на динамічному ринку. Їхня обмеженість полягає у відсутності глибокого розуміння покупкових звичок клієнтів. Прогрес у галузі машинного навчання зумовлює кардинальні трансформації в стратегіях продажів та маркетингу, де більшість комерційних структур покладаються на прогнозування попиту та ідентифікацію ринкових трендів. Для підвищення точності прогнозів, методи інтелектуального аналізу даних та техніки машинного навчання слугують потужними інструментами, які дозволяють виявляти приховані знання у великих масивах даних.

1.1.2. Мета та методологія проєкту

Метою даного проєкту є розробка програмного прототипу у форматі веб-сервісу, призначеного для прогнозування продажів товарів у торгових точках компаній.

У межах цього дослідження для прогнозування продажів будуть застосовані методи інтелектуального аналізу даних у поєднанні з моделями машинного навчання регресійного типу, зокрема:

- Лінійна регресія
- Дерево рішень
- Випадковий ліс
- XGBoost регресор

На основі аналізу отриманих результатів буде надана рекомендація щодо найкращої моделі для досягнення максимальної точності прогнозування.

1.1.3. Функціональність прототипу та очікувані результати

Розроблений прототип не обмежуватиметься лише прогнозуванням. Він також забезпечить графічне представлення впливу та кореляції між змінними, а також візуалізацію результатів роботи моделей із прогнозованими значеннями.

Очікується, що результати цієї роботи нададуть компаніям системне уявлення про оптимальні стратегії розміщення товарів та організації торгових точок. Це сприятиме формуванню позитивного досвіду клієнтів, що, у свою чергу, призведе до зростання обсягів продажів та доходів.

1.2. Значимість бізнес-аналітики та проблематика прогнозування

Прориви у сфері машинного навчання (МН) та аналізу даних зробили бізнес-аналітику критично важливою складовою сучасних систем підтримки прийняття рішень. У цьому контексті, прогнозування продажів і попиту

набуває вирішального значення для розробки ефективних аналітичних рішень. Фундаментальними концепціями, що лежать в основі комерційної діяльності, є попит та пропозиція. З огляду на стрімке поширення торговельних центрів та онлайн-платформ, конкуренція між різними комерційними суб'єктами щоденно посилюється.

Для залучення споживачів та управління запасами, логістикою та транспортним обслуговуванням, торговельні мережі регулярно застосовують унікальні акційні пропозиції, що вимагає точного прогнозування обсягів продажів за одиницю товару. З цією метою компанії здійснюють збір та моніторинг даних про продажі кожного окремого товару, що формує великі сховища даних, які містять значний обсяг інформації про клієнтів та специфічні атрибути товарів.

Ручна обробка та аналіз цих масивів даних є неефективними, схильними до суттєвих помилок і витратними з точки зору часу, що є неприпустимим у сучасному високошвидкісному бізнес-середовищі. Незважаючи на важливість прогнозування продажів і попиту, відсутність надійного прогностичного рішення призводить до неточних оцінок, які можуть негативно впливати на операційну діяльність і продажі організації. Переоцінка обсягу продажів призводить до невідповідності фактичних показників очікуваним, тоді як недооцінка може спричинити збільшення рекламних витрат та зниження прибутковості (Bajaj et al., 2020).

Точний прогноз продажів є критичним для планування операційної та збутової діяльності організацій. Чітке передбачення потенційних продажів продукту дозволяє роздрібним продавцям і виробникам більш ефективно планувати маркетингові, виробничі та закупівельні стратегії. Прогнозування продажів є ключовим елементом для виробників, оптових та роздрібних торговців, а також важливим завданням у діяльності, пов'язаній із ланцюгом поставок. Цей процес є складним, і його труднощі зростають за наявності дефіциту даних, пропущених значень або викидів.

Сучасні алгоритми машинного навчання пропонують методи для оцінки потенційного попиту та прогнозування доходів компанії. Різноманітні методики МН використовуються для прогнозування обсягу продажів, що сприяє розробці та вдосконаленню ринкових бізнес-стратегій та підвищенню обізнаності споживачів. Прогнози допомагають підприємствам аналізувати минулі тенденції, виявляти потенційні фінансові проблеми та оптимізувати планування. Ретельна розробка стратегії підвищує ймовірність успіху.

Для виявлення складних закономірностей у динаміці продажів, які включають численні фактори ризику, можуть бути застосовані методи контрольованого машинного навчання.

1.3. Обсяг та цілі проєкту

Сфера застосування даного проєкту обмежена аналізом даних про продажі товарів виключно для компанії, що працює у секторі харчових продуктів та напоїв (Foods and Beverages Company).

Методологія проєкту передбачає використання даних про поведінку клієнтів при купівлі та агрегованих даних про продажі товарів у торгових точках (аутлетах), отриманих з відповідних звітних документів. Крім того, до обсягу проєкту включено:

- Застосування методів інтелектуального аналізу даних (ІАД) для всебічного аналізу інформації.

- Розробка прогностичної моделі у формі прикладного програмного забезпечення (application software), що базується на найбільш ефективній моделі машинного навчання (МН).

Основною метою даного дослідження є створення надійної прогностичної моделі для прогнозування обсягів продажів товарів у торгових точках.

Головні завдання (задачі) проєкту:

а) Розробка програмного забезпечення.

Створити прикладне програмне забезпечення, що забезпечує функціонал прогнозування продажів на основі введених користувачем параметрів.

б) Імплементація алгоритмів МН - реалізувати обрані алгоритми машинного навчання на даних про продажі компанії у сфері харчових продуктів та напоїв за допомогою розробленого програмного інструменту.

в) Аналіз продуктивності - провести детальний аналіз продуктивності та точності імплементованих прогностичних моделей для визначення моделі з максимальною точністю прогнозування.

1.4. Концептуальні основи машинного навчання та прогнозування продажів

Даний проєкт зосереджує свою увагу на застосуванні машинного навчання (МН) для прогнозування продажів на основі історичних даних, зібраних у численних торгових точках та по різних категоріях товарів у галузі харчових продуктів та напоїв.

1.4.1. Сутність машинного навчання

Масштабні обсяги необроблених даних, що генеруються у сучасному світі, вимагають ретельного аналізу для отримання високоінформативних та достовірних результатів у відповідних галузях. Машинне навчання є підгалуззю штучного інтелекту, яка займається розробкою комп'ютерних програм, здатних природно покращувати свою продуктивність з часом.

Машинне навчання — це «галузь досліджень, яка надає комп’ютерам здатність до навчання без явного програмування». Інакше кажучи, МН займається створенням алгоритмів, що генерують прогнози на основі даних. Це є завданням штучного інтелекту, що полягає у виявленні (навчанні) функції $f: X \rightarrow Y$, яка відображає вхідну область X (даних) у вихідну область Y (можливих прогнозів).

Методи машинного навчання дозволяють навчати дані розпізнавати патерни та оцінювати їх для прийняття рішень з мінімальною участю людини.

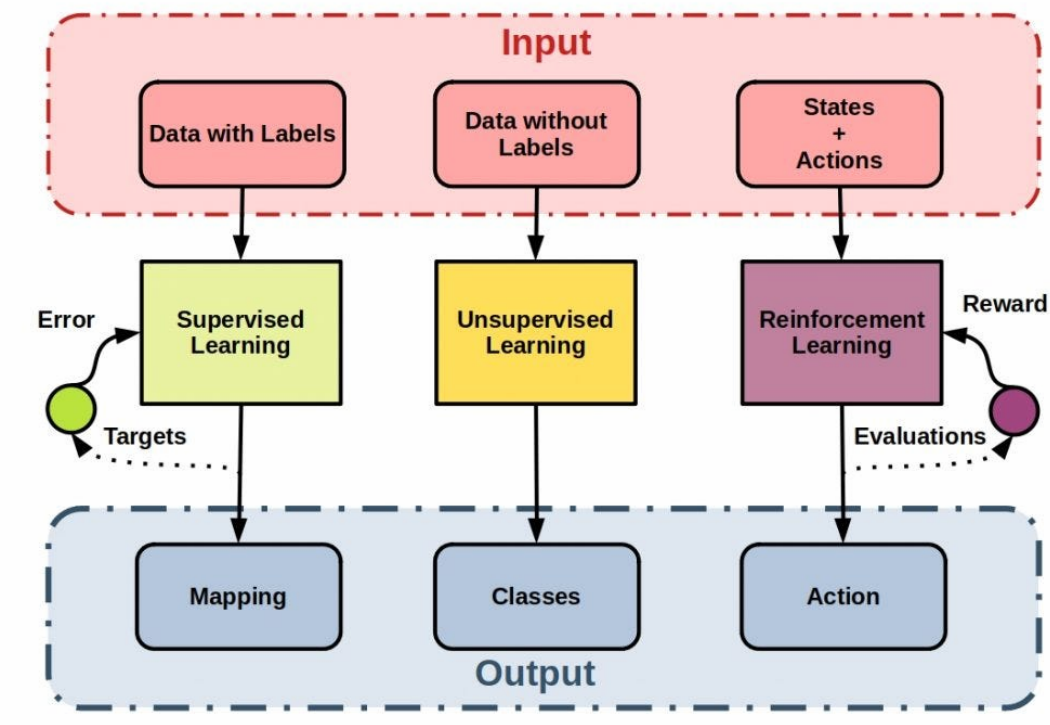


Рис. 1.1. Парадигми машинного навчання

Машинне навчання традиційно категоризується за трьома основними парадигмами навчання:

1. Контрольоване навчання (Supervised Learning) - модель тренується за допомогою алгоритмів для виявлення зв'язків (патернів) у наборі даних, що містить ознаки та відповідні мітки (відповіді). Надалі ця модель використовується для прогнозування міток для ознак у новому наборі даних.

2. Неконтрольоване навчання (Unsupervised Learning) - алгоритм тренується на некласифікованих і немаркованих даних, що дозволяє йому діяти на цих даних без зовнішнього нагляду. У цьому випадку, за відсутності попереднього навчального набору даних, метою системи є класифікація несортованих даних на основі схожості, патернів та відмінностей.

Неконтрольоване навчання охоплює такі групи алгоритмів, як кластеризація та асоціація.

3. Навчання з підкріпленням (Reinforcement Learning) - фокусується на визначенні того, як інтелектуальні агенти повинні поводитися у певному середовищі для максимізації сукупної винагороди (cumulative reward). Воно застосовується для визначення оптимального курсу дій у заданій ситуації. На відміну від контрольованого навчання, де навчальні дані містять ключ із відповіддю, навчання з підкріпленням покладається на агента, який навчається на основі власного досвіду та взаємодії з середовищем, а не на заздалегідь відомих відповідях.

1.4.2. Аналіз даних та прогностичні можливості

З огляду на зростаючу цінність даних, аналіз та інтерпретація для отримання ефективних результатів розвиваються паралельно з технологічним прогресом. У машинному навчанні вирішуються як завдання контрольованого, так і неконтрольованого типу. Часто проблеми класифікації слугують джерелом для виявлення знань (knowledge discovery). Основний акцент робиться на розробці самодостатньої системи, здатної виконувати обчислення та аналіз для отримання більш точних і прецизійних результатів. Це передбачає створення ресурсів та використання регресії для формування точних прогнозів щодо майбутнього. Шляхом застосування статистичних та імовірнісних алгоритмів дані можуть бути трансформовані у знання. Концептуальною основою для статистики слугують вибіркові розподіли (sampling distributions).

1.5. Роль та значення прогнозування продажів у бізнес-плануванні

Прогнозування продажів являє собою процес оцінки майбутнього обсягу реалізації товарів або послуг торговим підрозділом з метою передбачення потенційного доходу. Цей процес має критичне значення для

численних галузей, оскільки сприяє збільшенню прибутків компанії через формування ефективних стратегічних планів.

Експоненційне зростання обсягів даних, що використовуються в комерційних транзакціях (особливо в електронній комерції), створює для бізнесу значні виклики щодо вибору точних методик інтелектуального аналізу даних та розробки успішної стратегії прогнозування [3]. Здатність точно передбачати продажі є фундаментальною вимогою для ефективного корпоративного планування та прийняття рішень, що дозволяє підприємствам оптимізувати свою операційну діяльність.

Прогнозування продажів суттєво впливає на планування відділів, включаючи відділи продажів, маркетингу та складського господарства, зокрема, на визначення оптимального розташування складських приміщень. Відповідно, дані про минулі продажі дозволяють більш точно передбачати майбутні ринкові тенденції.

На базовому рівні, прогноз продажів є оцінкою того, як ринок відреагує на маркетингові ініціативи компанії. Завдання прогнозування полягає в оцінці майбутніх обсягів продажів для широкого спектру комерційних суб'єктів, таких як супермаркети, продовольчі магазини, заклади громадського харчування (ресторани, пекарні, кондитерські тощо).

Важливість прогнозування продажів простягається на всі функціональні рівні підприємства:

- Прогнозування допомагає підприємству зменшувати запаси товарів з очікуваним зниженням продажів і збільшувати їх для товарів з прогнозованим зростанням, що безпосередньо призводить до підвищення загальних продажів.

- Фінансові відділи використовують прогнози продажів для формування бюджетів, планування набору персоналу та розробки графіків виробництва й операційних циклів.

Звіт про аналіз продажів відображає динаміку обсягу продажів у часі, сигналізуючи про їхнє зростання чи падіння. Цей звіт дозволяє компанії

регулярно оцінювати тенденції для вибору найбільш адекватного курсу дій, ідентифікувати ринкові можливості та потенційні зони зростання. Для великих організацій аналіз продажів може бути сегментований за регіонами, підрозділами або філіями. Порівняння фактичних та прогнозованих продажів у звіті є ключовим.

Основна мета прогнозування продажів полягає у наданні компанії інструменту для визначення своїх стратегічних цілей та коригування підходів з метою підвищення продуктивності у майбутніх періодах.

1.6. Аналіз існуючих досліджень та методів прогнозного моделювання роздрібних продажів

У цьому проєкті було застосовано низку методів дослідження даних (data mining) та машинного навчання (ML) для прогнозування обсягів роздрібних продажів. Метою було оцінювання продажів для будь-якого роздрібного продавця на задану дату. У цьому розділі наведено огляд релевантних наукових праць, присвячених прогнозуванню та аналізу продажів за допомогою алгоритмів машинного навчання, що слугує основою для подальшого дослідження.

Дослідження демонструють стійку тенденцію до використання передових ML-алгоритмів для підвищення точності прогнозів:

1.6.1. Застосування алгоритмів бустингу та ансамблевих методів

В роботі [3] порівняли ефективність ML-алгоритмів із традиційними методами дослідження даних, використовуючи історичні дані про продажі. Їхні результати підтвердили надійність та точність ML-систем, при цьому алгоритм Gradient Boost показав найкращу підгонку (fit) та найвищу точність у метриках класифікації.

В [4] дослідили проблеми прогнозування в електронній комерції та запропонували метод стекованої генералізації (stacked generalization) з

використанням регресорів підрівня. Цей підхід, хоча і не мав суттєвої статистичної переваги над Random Forest, продемонстрував високу точність прогнозування попиту навіть при обмеженій кількості даних. Також було відзначено використання регресора XGBoost та градієнтних алгоритмів для прогнозування продажів великих роздрібних компаній.

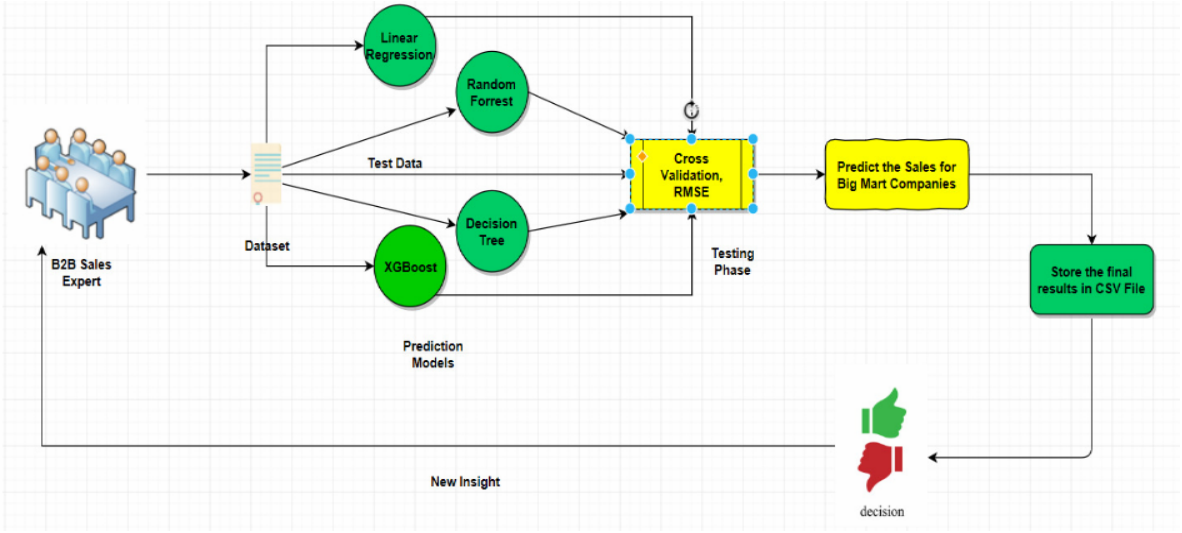


Рис. 1.2. Загальна схема (High-level overview) представленої інтелектуальної системи

У цій моделі для розв'язання задачі прогнозування доходу від продажів (Sales revenue) різних продуктів у різних торгових точках компанії використовується п'ятиетапна процедура.

На першому етапі здійснюється збір, отримання та поділ даних на навчальну (training) та тестову (test) вибірки. Ці дані проходять попередній аналіз, який включає одновимірний (univariate) та двовимірний (bivariate) аналіз.

На другому етапі виконується попередня обробка даних (data pre-processing), яка спрямована на усунення пропущених та помилкових значень у наборі даних.

На третьому етапі здійснюється відбір та модифікація ознак (features) для досягнення найкращих результатів.

На четвертому етапі застосовується трансформація ознак (feature transformation) для перетворення категоріальних ознак на числові.

На п'ятому етапі за допомогою різних алгоритмічних методів будуються моделі та оцінюються результати.

Ці результати передаються компанії, і після їх затвердження фірма застосовує їх для формування бізнес-моделі на наступний рік. Використання цього методу гарантує точніші та кращі результати.

В дослідженні [5] зосередились на прогнозуванні майбутньої вартості дорогоцінних металів, зокрема алмазів, використовуючи алгоритм Gradient Boost для досягнення максимальної точності. Вони підкреслили технічні переваги XGBoost, включаючи наближений метод для жадібних алгоритмів, паралельне навчання з внутрішньопам'ятним зберіганням даних, попереднє вибирання (prefetching) з урахуванням кешу та обробку даних поза межами оперативної пам'яті (out-of-core processing), що дозволяє ефективно працювати з більшими наборами даних.

Метод XGBoost навчає дерева рішень, використовуючи фреймворк градієнтного бустингу (gradient boosting), який передбачає мінімізацію функції втрат (loss function) шляхом ітеративного додавання дерев рішень до ансамблю.

Функція втрат обчислює розбіжність між фактичними та прогнозованими значеннями цільової змінної. Градієнтний бустинг ітеративно додає дерева рішень, де кожне нове дерево коригує недоліки (помилки) попередніх дерев.

На кожній ітерації метод створює нове дерево рішень і підганяє його до залишків (residuals) попередніх дерев. Залишки — це різниці між фактичними та прогнозованими значеннями.

Для зменшення перенавчання (overfitting) та підвищення узагальнюючої здатності (generalization capabilities) моделі XGBoost використовує регуляризовану цільову функцію (regularized objective function).

Цільова функція складається з двох компонентів:

- Функція втрат оцінює розбіжність між прогнозованими та фактичними значеннями.

- Член регуляризації (regularization term) накладає штраф на складність моделі.

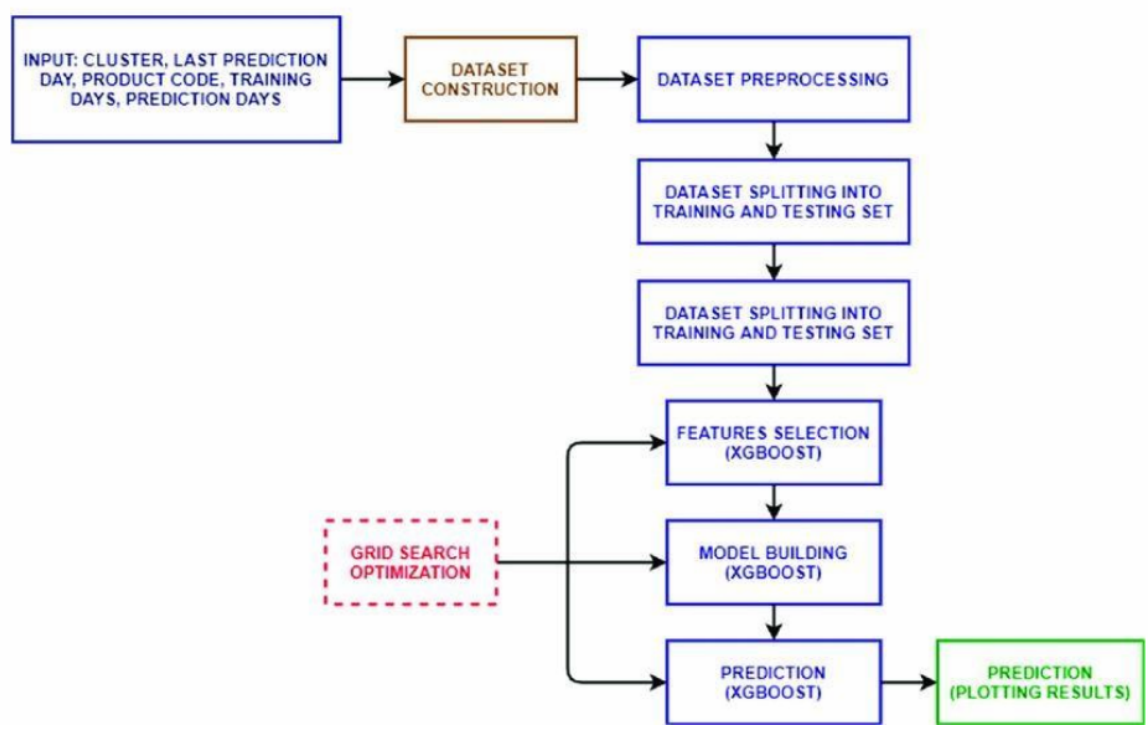


Рис. 1.3. Система прогнозування вартості на основі алгоритму Gradient Boost

Алгоритм XGBoost застосовує метод оптимізації градієнтного спуску (gradient descent optimization) для мінімізації цільової функції. На кожній ітерації алгоритм обчислює градієнти цільової функції відносно параметрів моделі та оновлює параметри в напрямку негативного градієнта.

Для використання XGBoost необхідно враховувати та налаштовувати наступні параметри регуляризації:

- Gamma (γ): Збільшення значення γ призводить до створення меншої кількості розщеплень у деревах рішень (контролює мінімальну втрату зниження, необхідну для подальшого розділення вузла).

- Alpha (α): L1-регуляризація ваг листків. Зі збільшенням α зростає регуляризація, що може обнулити деякі ваги.

- Lambda (λ): L2-регуляризація ваг листків. Вона допомагає плавно зменшувати ваги листків до нуля, на відміну від сильних обмежень L1-регуляризації.

1.6.2. Використання Random Forest та інших моделей

В дослідженні [9] запропонували ML-алгоритм для прогнозування структури продажів та обсягів товарів, визначивши модель Random Forest як найбільш відповідну для їхнього дослідження на основі даних про продажі за попередні роки.

Алгоритм LELSF (Lightweight Ensemble Learning for Sales Forecasting) — Легковагове ансамблеве навчання для прогнозування продажів є простим, ефективним та точним методом прогнозування. Його реалізація включає наступні ключові етапи:

1. Попередня обробка даних.

Процес починається з очищення та попередньої обробки даних про продажі, вилучення місячних трендів з поля "Order Date" (Дата Замовлення), а також нормалізації показників продажів за допомогою Min-Max масштабування.

2. Розробка ознак (Feature Engineering).

Створюються ключові прогностичні ознаки, такі як лагові значення (lag values) та ковзні середні (rolling means), призначені для виявлення як короткострокових, так і довгострокових трендів продажів.

3. Ансамблеве моделювання.

Тренуються три легковагові моделі регресії: Лінійна Регресія (Linear Regression), Регресор на Основі Дерев Рішень (Decision Tree Regressor) та Метод K-Найближчих Сусідів (K-Nearest Neighbors).

4. Комбінування.

Ці моделі об'єднуються за допомогою Voting Regressor (Голосуючий Регресор) для формування ансамблю. Цей ансамбль усереднює прогнози кожної окремої моделі для отримання фінального прогнозу продажів.

5. Оцінка.

Модель навчається на 70% даних і тестується на решті 30%. Продуктивність оцінюється за метриками: RMSE (Root Mean Square Error), MAE (Mean Absolute Error) та MAPE (Mean Absolute Percentage Error).

Такий підхід забезпечує точні, інтерпретовані та ресурсоефективні прогнози продажів.

Рисунок 1.4 ілюструє блок-схему алгоритму LELSF.

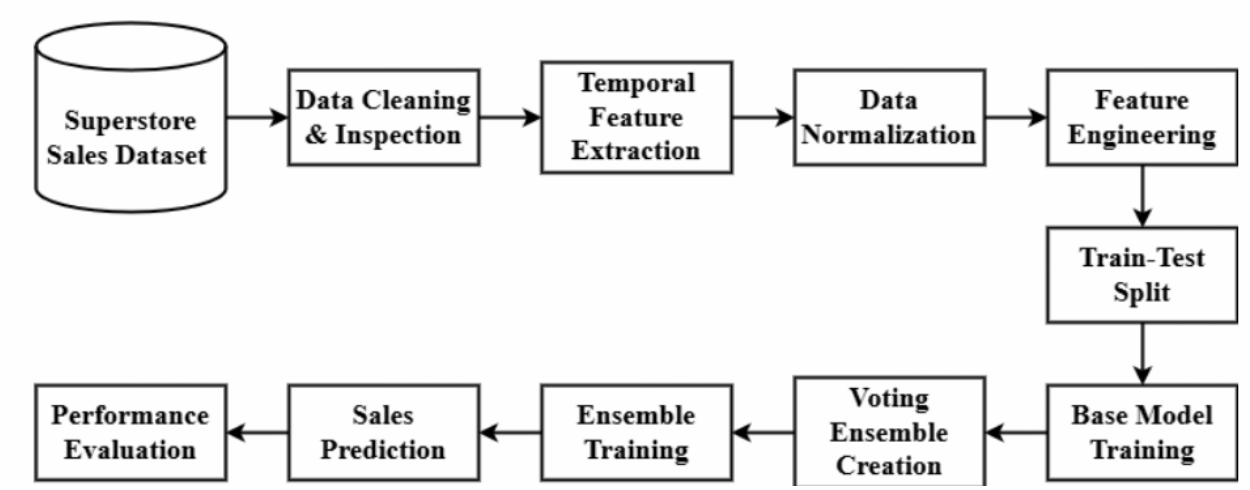


Рис. 1.4. Алгоритм легковагового ансамблевого навчання для прогнозування продажів

В роботі [10] використовували Random Forest та XGBoost для попередньої обробки сирих даних (включаючи заповнення пропущених значень, виявлення аномалій та викидів) та прогнозування продажів великого магазину. Порівняльний аналіз показав, що обидва методи є найкращими моделями для прогнозування продажів.

В роботі [11] досліджували прогнозування фондового ринку. Обидва дослідження використовували Support Vector Machine (SVM) та Random

Forest, виявивши, що алгоритм Random Forest показав вищу точність у прогнозуванні ринкових цін.

В дослідженні [12] представили дворівневий статистичний підхід для прогнозування продажів товарів, який використовує метрику MAE (Mean Absolute Error). Для оцінки Squared Error Value вони застосували різноманітні алгоритми, зокрема KNN (K-Nearest Neighbors), Support Vector Regression, Linear Regression та Regression Tree.

У цій роботі побудова моделі прогнозування відбувається у два етапи:

1. Побудова одиничних моделей

На першому етапі будуються одиничні моделі популярних прогностичних методів, таких як:

- Лінійна регресія (linear regression)
- Дерево регресії (regression tree)
- Опорно-векторна регресія (support vector regression)
- Метод k-найближчих сусідів (k-nearest neighbor)

2. Побудова дворівневої статистичної моделі

На другому етапі була побудована дворівнева статистична модель. Вона складалася з методів машинного навчання, таких як лінійна регресія, опорно-векторна регресія. Ці алгоритми машинного навчання були об'єднані для здійснення фінального прогнозу.

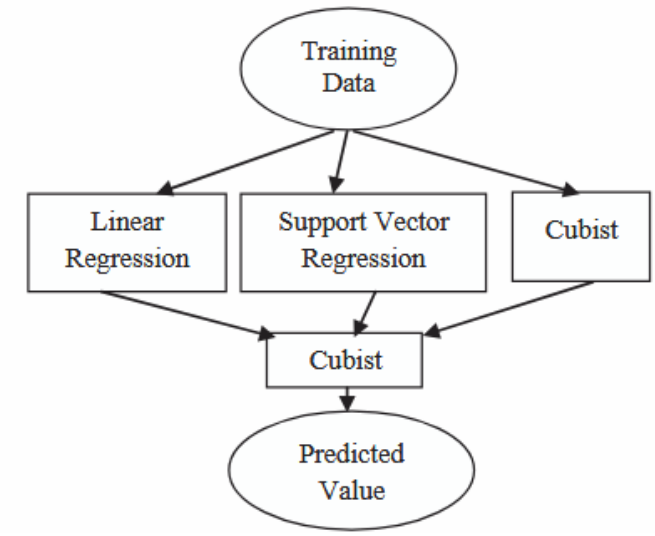


Рис. 1.5. Дворівнева статистична модель

Стеккування (Stacking) є типом ансамблевого методу, який зазвичай використовується для комбінування методів машинного навчання з метою підвищення точності прогностичних моделей. По суті, це комбінація різних моделей, які розглядаються як єдиний блок. Стеккування може мати більше двох шарів; це збільшує складність моделі, але може бути корисним для точних прогнозів.

Згідно з рис. 1.6, у дворівневій статистичній моделі:

- Лінійна регресія, опорно-векторна регресія виступають як моделі нижнього шару (bottom layer models). Вони приймають оригінальні ознаки набору даних як вхідні дані.

Потім Cubist виступає як модель верхнього шару (top layer model). Вона приймає прогнози моделей нижнього шару як свої вхідні дані та формує фінальний прогноз.

В роботі [13] запропонували програмне рішення для прогнозування майбутніх продажів на основі історичних даних, використовуючи моделі Multiple Linear Regression та Random Forest.

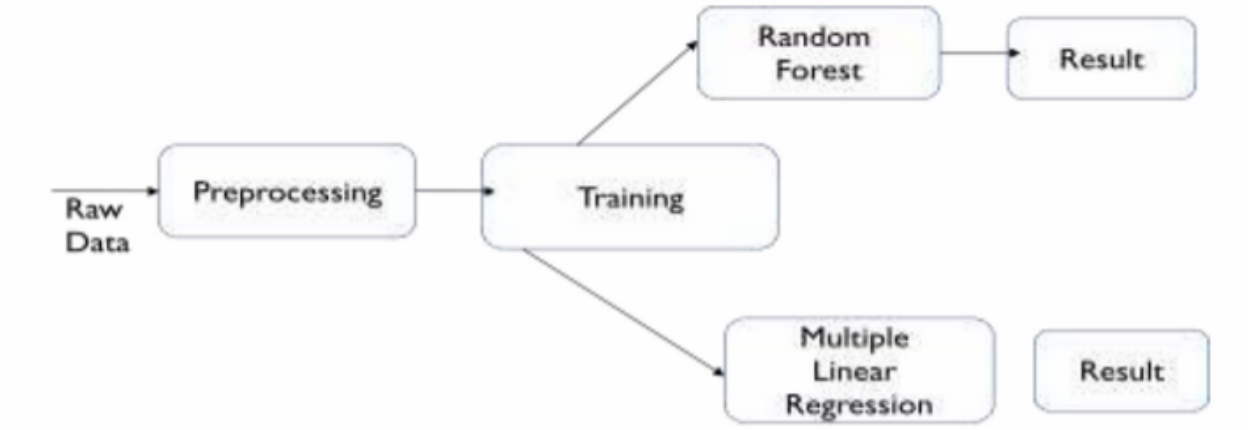


Рис. 1.6. Алгоритм роботи системи прогнозування майбутніх продажів на основі історичних даних

В роботі [18] критично обговорюєnmcz застосування аналітики великих даних (BDA) в управлінні ланцюгами постачання (SCM), висвітлюючи

поточні проблеми та прогалини в прогнозуванні попиту та продажів за допомогою BDA.

В [19] проаналізували тенденції використання машинного навчання у прогнозуванні попиту та продажів у період з 2009 по 2017 рік, наголошуючи на необхідності інвестування в ці підходи на противагу традиційним методам.

Автори в [20] представили дослідження моделювання та прогнозування поведінки споживачів на основі даних про продажі супермаркетів, запропонувавши новий підхід до рекомендацій продуктів, а в [21] зосередилися на огляді репрезентативних ML-додатків у SCM, зокрема у сфері прогнозування попиту та продажів.

В дослідженні [22] автори надали стислий огляд та таксономію методологій прогнозування фондового ринку у фінансовій сфері.

Отже, аналіз літератури підтверджує домінуючу роль алгоритмів машинного навчання, зокрема ансамблевих методів як-от Gradient Boosting (зокрема XGBoost) та Random Forest, у досягненні високої точності при прогнозуванні роздрібних продажів. Ці методи демонструють ефективність як для обробки та очищення даних, так і для побудови кінцевої прогностичної моделі.

Висновки до розділу

У першому розділі розглянуто теоретичні та методологічні основи використання технологій машинного навчання для прогнозування продажів. Проаналізовано сучасні підходи до побудови моделей прогнозування та виявлено, що традиційні статистичні методи не забезпечують належної точності у динамічному бізнес-середовищі. Визначено, що алгоритми машинного навчання, зокрема ансамблеві моделі, дозволяють враховувати складні нелінійні взаємозв’язки між факторами попиту. Досліджено роль бізнес-аналітики та обґрунтовано необхідність інтеграції прогнозних моделей

у процеси планування продажів. Отримані результати стали методологічним підґрунтям для подальшого моделювання та побудови системи прогнозування на основі ML-технологій.

Проведене дослідження предметної області дозволило обґрунтувати вибір напрямів моделювання, визначити вимоги до системи та сформувати структуру аналітичного рішення, орієнтованого на практичне використання у сфері бізнес-аналітики.

РОЗДІЛ 2. МОДЕЛІ, АЛГОРИТМИ ТА АРХІТЕКТУРА СИСТЕМИ ПРОГНОЗУВАННЯ ПРОДАЖІВ НА ОСНОВІ ТЕХНОЛОГІЙ МАШИННОГО НАВЧАННЯ

2.1. Дослідження алгоритмів машинного навчання та використання регресійних моделей для прогнозування продажів

Машинне навчання (МН) має потенціал для обробки складних типів даних з різних аспектів, включаючи часові ряди, категоріальні змінні, текстові дані, зображення, нечіткі (fuzzy) елементи та інші змінні. У контексті прогнозування продажів перевага надається регресійним методам, оскільки застосування алгоритмів регресії на основі МН дозволяє отримувати результати, що перевершують показники традиційних методів аналізу часових рядів.

Такі алгоритми, як лінійна регресія, регресія на основі дерева рішень, регресія на основі випадкового лісу та XGBoost, демонструють вищу продуктивність у порівнянні з класичними аналітичними техніками часових рядів.

2.1.1. Алгоритм лінійної регресії

Лінійна регресія є алгоритмом машинного навчання, що належить до категорії контрольованого навчання (supervised learning) і виконує регресійний аналіз. Регресія використовується для моделювання заздалегідь визначеного прогностичного значення (залежної змінної) на основі незалежних змінних. Вона переважно застосовується для встановлення зв'язку між змінними та їхнього впливу на прогнозування.

Це параметричний метод, який використовує набір незалежних змінних для прогнозування безперервної (залежної) змінної. Метод називається параметричним, оскільки його застосування ґрунтується на низці припущень щодо розподілу даних.

Лінійна регресія виконує завдання прогнозування значення залежної або цільової змінної (y) на основі наданої незалежної змінної (x). Таким чином, цей регресійний метод встановлює лінійну залежність між вхідною (x) та вихідною (y) змінними.

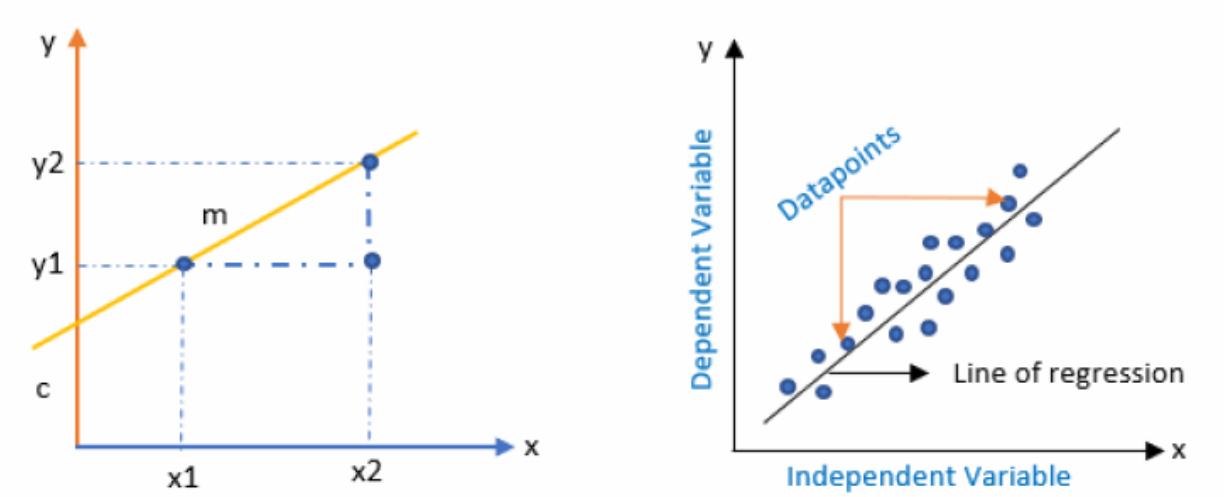


Рис. 2.1. Лінійна регресія

Просте лінійне рівняння регресії з однією залежною та однією незалежною змінною, що є найпростішим представленням прямої лінії, описується формулою:

$$y = m \cdot x + c$$

де:

y — залежна змінна (цільовий показник);

x — незалежна змінна (наприклад, продаж певного продукту);

m — коефіцієнт нахилу (slope) лінії;

c — перетин з віссю Y (коефіцієнт).

Для випадку множинної регресії (з кількома незалежними змінними) рівняння має вигляд:

$$y = m_1 \cdot x_1 + m_2 \cdot x_2 + m_3 \cdot x_3 + \cdots + m_n \cdot x_n + C$$

Основною метою цього підходу є знаходження прямої (або гіперплощини для множинної регресії), яка найкраще апроксимує залежну (цільову) змінну та незалежні змінні у наборі даних. Це досягається шляхом ідентифікації найбільш оптимальних значень усіх коефіцієнтів mi. "Найкраща відповідність" (Best fit) означає, що прогнозоване значення є максимально наближеним до фактичних даних і має мінімальну похибку.

2.1.2. Принципи та застосування дерева рішень

Алгоритм дерева рішень (Decision Tree, DT) належить до сімейства алгоритмів контрольованого навчання (supervised learning). Хоча його можна застосовувати як для задач класифікації, так і для регресії, частіше він використовується саме для класифікаційних задач.

Дерево рішень є інтуїтивно зрозумілою моделлю з низьким рівнем упередженості (bias). Воно використовується для побудови дерева класифікації або регресії у підході "зверху-вниз", де першим вузлом, що розглядається, є кореневий вузол (root node).

Структура класифікатора-дерева складається з:

- Внутрішніх вузлів, які відображають ознаки (features) набору даних.
- Гілок, які представляють правила прийняття рішень (decision rules).
- Листових вузлів (leaf nodes), які відповідають кінцевим результатам (прогнозованому класу або значенню).

Метод DT створює навчальну модель шляхом вивчення простих правил прийняття рішень, виведених з минулих даних, і використовує цю модель для прогнозування класу або значення цільової змінної.

Процес прийняття рішень і тестування в деревах рішень ґрунтується на ознаках набору даних. Побудова дерева здійснюється за допомогою алгоритму CART (Classification and Regression Tree) .

Процес прогнозування для певного набору даних починається з кореневого вузла дерева. На цьому етапі відбувається порівняння значень кореневого атрибута зі значеннями атрибутів запису/набору даних.

Подальший поділ дерева на піддерева (subtrees) здійснюється на основі результатів порівняння або відповідей (наприклад, "ТАК"/"НІ").

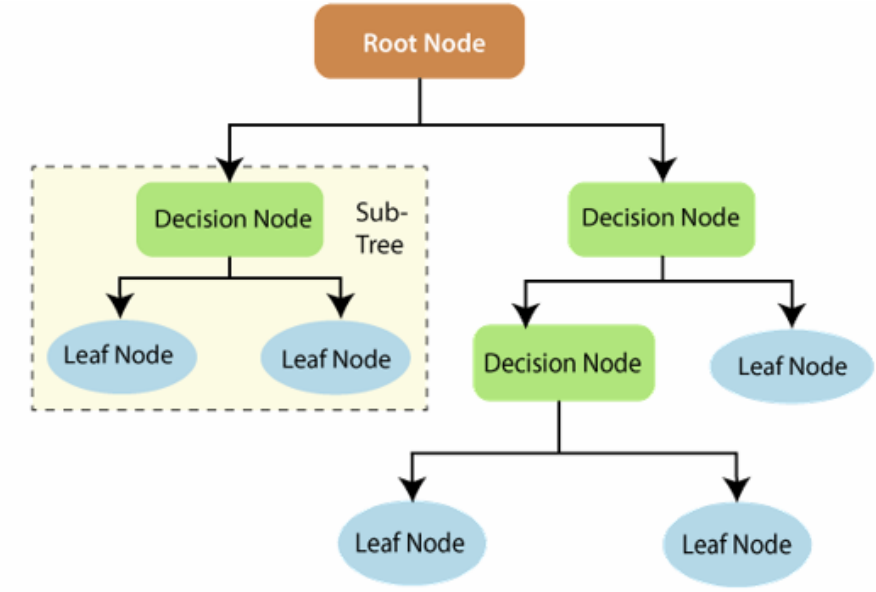


Рис. 2.2. Структура дерева рішень

Алгоритм переходить до наступного вузла, дотримуючись гілки, що відповідає отриманому значенню. Для наступного вузла алгоритм повторно перевіряє значення атрибута з іншими підвузлами й продовжує цей ітеративний процес, доки не досягне листового вузла дерева, який містить кінцевий прогноз.

2.1.3. Ансамблевий метод регресії випадкового лісу

Регресія випадкового лісу (Random Forest Regression, RF) є широко застосовуваним статистичним ансамблевим методом, який поєднує прогнози, отримані з множини дерев рішень, збудованих на різних вибірках початкового набору даних.

Алгоритм RF використовує усереднене значення (mean) індивідуальних прогнозів для задач регресії (або моду для класифікації), що дозволяє використовувати його для класифікації, регресії та інших ансамблевих завдань машинного навчання. Його робота полягає у формуванні значної

кількості дерев рішень на етапі тренування, після чого кінцевий прогноз генерується як середнє значення (для регресії) або мода класу (для класифікації) окремих дерев. Використання випадкових лісів є ефективним рішенням для мінімізації проблеми перенавчання (overfitting), характерної для окремих дерев рішень.

Випадковий ліс можна визначити як специфічний тип адитивної моделі, яка агрегує рішення від серії базових моделей для отримання фінального прогнозу.

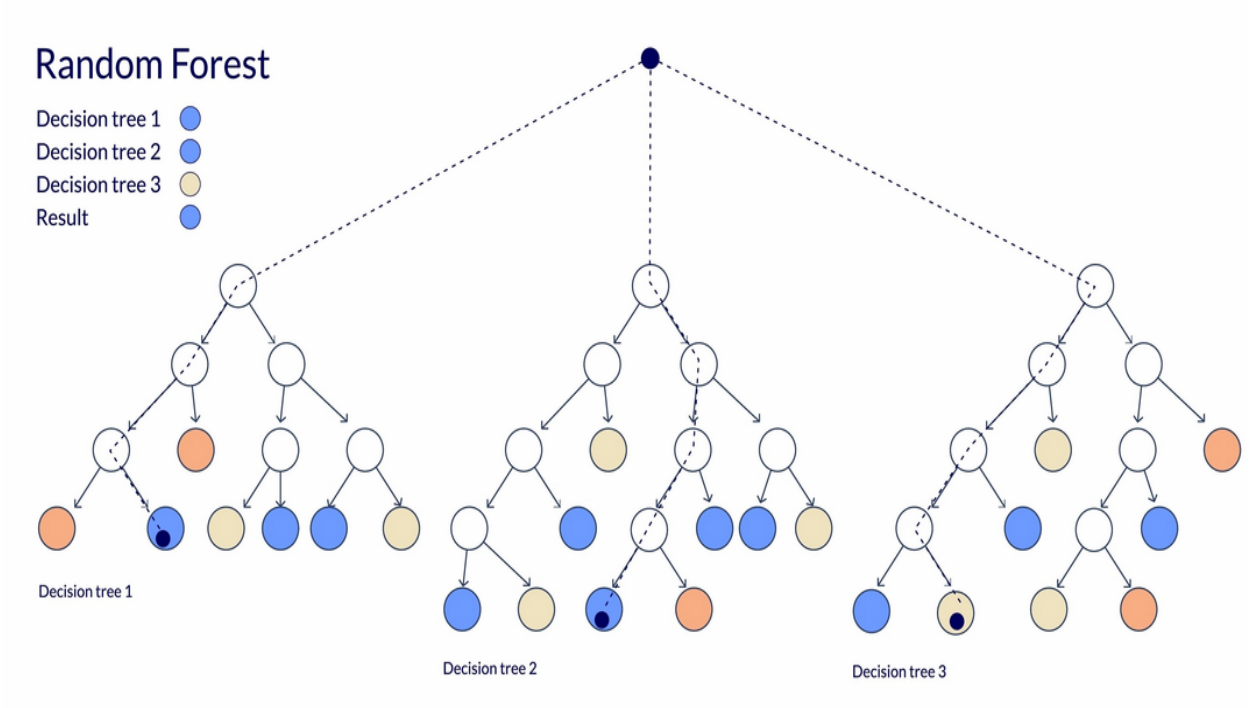


Рис. 2.3. Регресія випадкового лісу

Для забезпечення більш надійного та стійкого прогнозу, випадковий ліс будує численні дерева рішень та об'єднує їх. При цьому RF вводить додатковий рівень рандомізації у процес побудови моделі.

На відміну від стандартного дерева рішень, де на кожному вузлі для розбиття обирається найкраща ознака з усього набору, випадковий ліс шукає найкращу ознаку лише в випадково обраній підмножині ознак. Це створює високу різноманітність серед дерев, що, як правило, призводить до побудови кращої ансамблевої моделі.

Завдяки тому, що на кожному вузлі для розбиття обирається лише частина ознак, випадковий ліс цінується за його швидшу продуктивність прогнозування та знижене використання пам'яті порівняно з методами, що розглядають увесь простір ознак.

2.1.4. Оптимізована імплементація градієнтного бустингу (XGBoost)

Бібліотека XGBoost (Extreme Gradient Boosting) є оптимізованою, високоточною реалізацією розподіленого градієнтного бустингу (Gradient Boosting), розробленою переважно для підвищення продуктивності та обчислювальної швидкості моделей машинного навчання. Її архітектура спрямована на максимально ефективне використання обчислювальних ресурсів для методів бустингу дерев рішень. XGBoost імплементує алгоритми машинного навчання в рамках загальної парадигми градієнтного бустингу.

Подібно до загальних методів градієнтного бустингу на деревах рішень (GBDT/GBM), XGBoost пропонує паралельний бустинг дерев. Для оцінки якості розщеплень у навчальному наборі даних XGBoost застосовує рівневий підхід (level-wise). Це передбачає сканування градієнтних значень та використання часткових сум для обчислення потенційної вигоди від розщеплення на кожній можливій точці.

Градієнтний бустинг є вдосконаленою варіацією концепції "бустингу", де нові моделі послідовно конструюються для прогнозування залишків (residuals) або помилок попередніх моделей. Фінальне передбачення формується шляхом комбінування прогнозів усіх побудованих моделей.

Цей метод є розширенням бустингу, оскільки для мінімізації функції втрат (loss function) при додаванні нових моделей використовується підхід градієнтного спуску (gradient descent). Стратегія градієнтного бустингу ефективна для вирішення задач прогнозного моделювання як у регресії, так і в класифікації.

Градієнтний бустинг належить до ансамблевих методів навчання (ensemble learning), які інтегрують численні дерева рішень для формування остаточної прогностичної моделі. Фундаментальна гіпотеза цього підходу полягає в тому, що група слабких класифікаторів (weak learners) може бути об'єднана за допомогою процесу бустингу для створення сильного класифікатора (strong learner).

Алгоритми GBDT (Gradient Boosting Decision Trees) ітеративно навчають набір глибоких дерев рішень. На кожній ітерації залишки помилок, отримані від моделі з попередньої ітерації, використовуються для підгонки (fitting) наступної моделі. Кінцевий прогноз являє собою зважене середнє прогнозів, згенерованих усіма деревами в ансамблі.

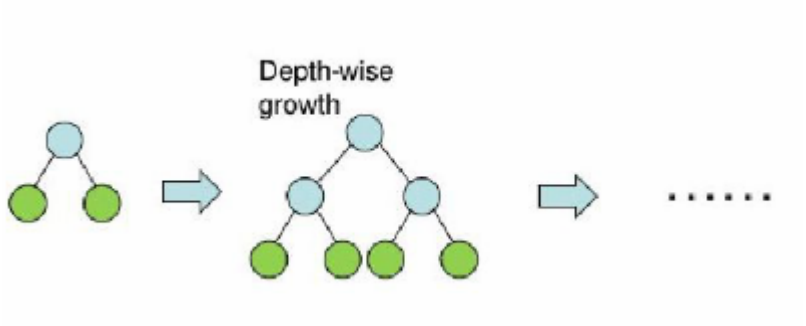


Рис. 2.4. XGBoost архітектура

Ітеративний процес моделювання може бути формалізований таким чином. Нехай $F_{t-1}(x)$ позначає повну модель після ітерації t−1, а h(x) — нове дерево, що додається на поточному кроці:

$$F_0 = 0$$
$$F_t(x) = F_{t-1}(x) + h(x)$$

Кожна нова функція h(x) прагне скоригувати помилки, допущені моделлю на попередніх ітераціях. Таким чином, нова функція h(x) має прогнозувати залишок (residual), який визначається як різниця між цільовим значенням та прогнозом попередньої моделі $F_{t-1}(x)$.

2.2. Методологічний підхід до створення моделей прогнозування

У цьому розділі представлено методологію та запропоновану архітектуру системи, які були використані для розробки прогнозних моделей для прогнозування продажів (Sales Forecasting).

На відміну від чітко структурованого життєвого циклу розробки програмного забезпечення (Software Development Life Cycle, SDLC), проєкти в галузі науки про дані (Data Science), зокрема створення прогнозних моделей, зазвичай характеризуються ітеративним та нелінійним робочим процесом із частими поверненнями та затримками. Цей процес є вкрай ресурсомістким з точки зору часу. У реальних бізнес-сценаріях може знадобитися від кількох місяців до років, перш ніж розроблена модель почне демонструвати значущі результати.

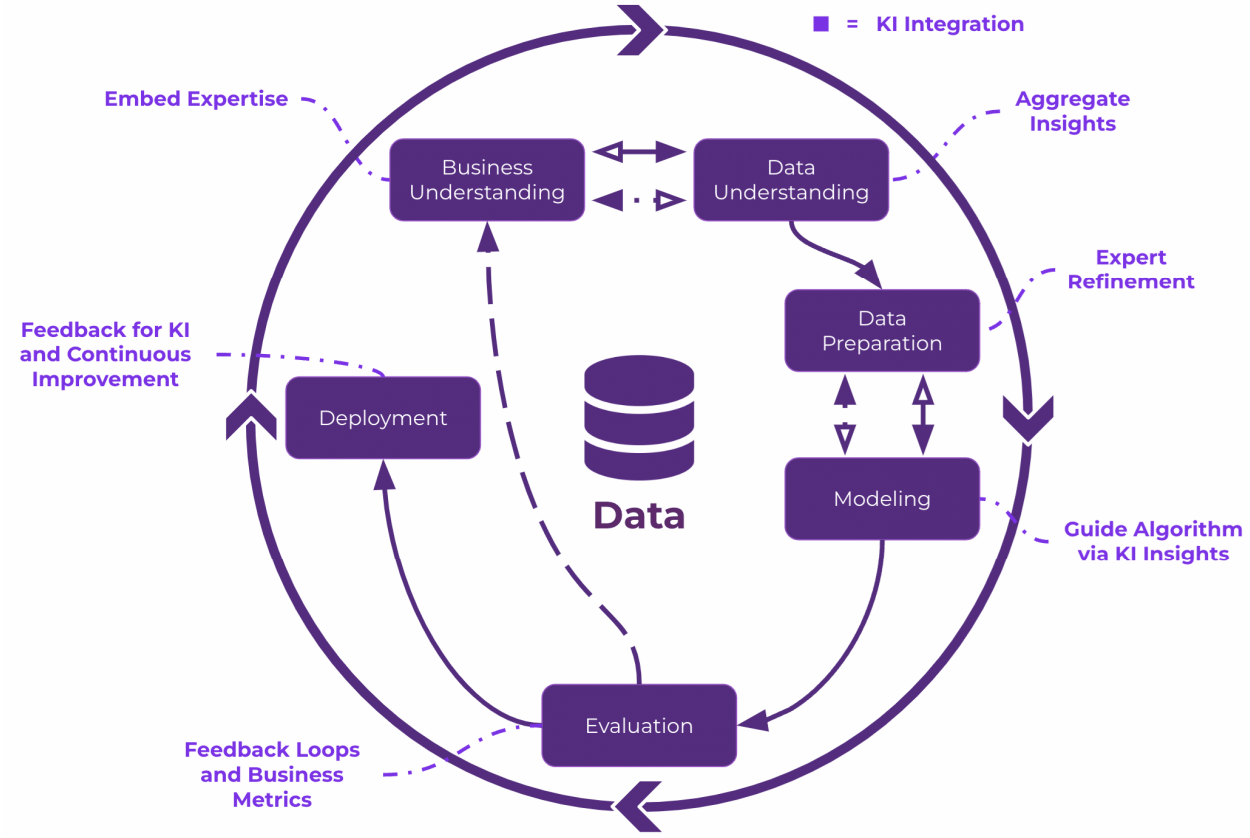


Рис. 2.5. Модель процесу добування даних із використанням методології CRISP-DM

Для реалізації цього проєкту було застосовано стандартизований робочий процес, що базується на одній із найстаріших та найбільш визнаних методологій — CRISP-DM (CRoss-Industry Standard Process for Data Mining). Методологія CRISP-DM забезпечує структурований підхід до планування проєктів добування даних (Data Mining), який сьогодні широко адаптований і використовується у сфері науки про дані та машинного навчання.

Використовуючи методологію CRISP-DM, було спроєктовано архітектурну діаграму запропонованої системи з метою створення ефективного продукту даних. Цей продукт даних являє собою інформаційну панель (dashboard), призначену для сприяння прийняттю рішень та розв'язання відповідної бізнес-проблеми. Для досягнення кінцевої мети — створення продуктів даних — робочий процес проілюстровано на рис. 2.5.

2.3. Опис пропонованої архітектура системи

Дотримуючись описаної методології CRISP-DM, для прогнозування продажів різноманітних товарів у різних торгових точках було запропоновано архітектуру системи, представлену на рисунку 2.6.

2.3.1. Етапи реалізації та робочий процес

Реалізація проєкту, згідно із запропонованою архітектурою, розпочинається з розуміння бізнес-контексту та ідентифікації вимог до даних.

1. Збір даних.

Відповідні дані, включаючи навчальні (train) та тестові (test) набори, будуть отримані з репозиторію Kaggle.

2. Попередній аналіз даних.

Отримані набори даних проходять етап попереднього розвідувального аналізу даних (Exploratory Data Analysis, EDA), що охоплює уніваріантний та біваріантний аналізи.

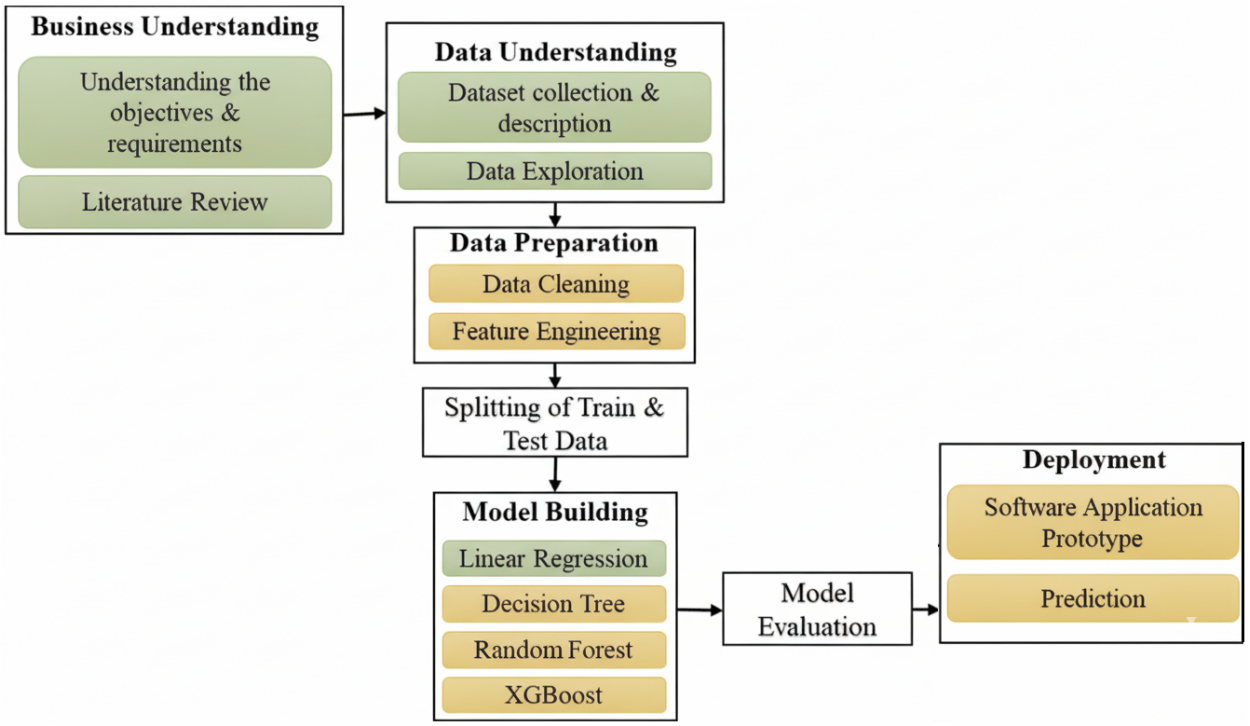


Рис. 2.6. Архітектура запропонованої системи

3. Підготовка даних.

На наступному етапі виконується попередня обробка даних (data preprocessing), яка включає методи корекції пропущених значень, помилкових записів та викидів (outliers). На цьому ж етапі здійснюється модифікація та відбір ознак (feature selection), а також трансформація ознак (feature transformation) для створення нових змінних на основі наявних. Завершення цього етапу забезпечує готовність даних для застосування алгоритмів машинного навчання (МН).

4. Розробка моделей.

Чотири попередньо обрані алгоритми МН будуть застосовані та навчені з використанням навчального набору даних для розробки прогнозних моделей.

5. Оцінка ефективності.

Ефективність розроблених моделей буде перевірена за допомогою тестового набору даних для визначення моделі з найвищою точністю прогнозування.

6. Впровадження (Deployment). На фінальному етапі впровадження буде створено програмний прототип для інтеграції з найкращою моделлю, обраною для розгортання в системі. За допомогою інтерфейсу користувачі зможуть візуалізувати дані про продажі та отримувати прогнози обсягів продажів певного товару, вводячи його ідентифікатор та ідентифікатор торгової точки.

2.3.2. Визначення цілей та вимог проекту для бізнесу (Business Understanding)

Фаза розуміння бізнесу зосереджується на визначенні цілей та вимог проєкту. Ключові дії цього етапу включають: ідентифікацію бізнес-областей для дослідження, розробку бізнес-моделі (за необхідності), окреслення очікуваних результатів аналізу, формулювання мети проєкту, висунення гіпотез та пошук ресурсів для збору даних.

Мета цього проєкту полягає у розробці прогнозної моделі для визначення обсягу продажів кожного продукту в конкретній торговій точці. Областю дослідження є аналіз даних про продажі, що вимагає розуміння характеристик товарів і закладів, які суттєво впливають на збільшення продажів.

На цьому етапі розробляється низка гіпотез, які можуть бути класифіковані на рівні продукту та торгової точки.

Гіпотези на рівні продукту можуть включати такі параметри, як бренд продукту, наявність акцій/пропозицій, рекламна підтримка, спосіб використання, процес пакування.

Гіпотези на рівні торгової точки можуть стосуватися таких факторів, як місце розташування магазину, демографічні показники (населення) в районі розташування, тип локації магазину, його розмір, взаємовідносини з клієнтами, реклама, відповідність локальним потребам, конкурентне середовище тощо.

Наприклад, гіпотеза може припускати, що брендований продукт може мати вищий обсяг продажів у певній точці, що залежить від потреб споживачів у цій локації, проте це припущення не може бути узагальнено на всі торгові точки. Генерація таких гіпотез є критично важливою для кращого розуміння бізнес-вимог.

2.3.3. Фаза розуміння даних (Data Understanding)

Фаза розуміння даних ґрунтується на результатах фази розуміння бізнесу та акцентує увагу на ідентифікації, зборі та аналізі наборів даних, релевантних для проєкту. Необхідні дані отримуються з авторитетних джерел, таких як Kaggle або GitHub.

Вибір відповідних даних є вирішальним, оскільки він дозволяє підтвердити або спростувати початково висунуті гіпотези. Ця фаза включає:

- визначення характеристик даних, таких як їхній формат, кількість записів та ідентифікатори полів.

- розвідувальний аналіз даних (EDA), тобто аналіз, візуалізація та ідентифікація кореляцій між змінними.

Крім того, на цьому етапі проводиться перевірка достовірності даних, отриманих з різних джерел, та їхня інтерпретація з урахуванням бізнес-цілей проєкту.

Для забезпечення чіткого розуміння природи обраного набору даних необхідно провести розвідувальний аналіз даних (EDA). EDA є стратегією аналізу, яка використовує статистичну графіку та інші методи візуалізації даних з метою узагальнення основних характеристик або ознак даних, змінних та відповідних взаємозв'язків.

Розвідувальний аналіз включає два основні типи аналізу, які проводяться для узагальнення та виявлення закономірностей у цільовій змінній (Target Variable):

- Уніваріантний аналіз, що оперує лише одним атрибутом.

- Біваріантний аналіз, що досліджує взаємозв'язок між двома атрибутами.

Уніваріантний аналіз є найпростішим типом статистичного аналізу, зосередженим на дослідженні однієї змінної (стовпця даних) за раз. У межах уніваріантного аналізу для візуалізації даних та отримання базових інсайтів щодо окремих змінних використовуються, зокрема, графіки частот (count plots).

Біваріантний аналіз визначається як дослідження будь-якого одночасного взаємозв'язку між двома змінними або атрибутами. Він спрямований на вивчення того, як дві змінні взаємодіють, а також на виявлення відмінностей між ними та їхніх потенційних причинно-наслідкових зв'язків.

Детальне обговорення уніваріантного та біваріантного аналізу обраного набору даних, разом із відповідними візуалізаціями, буде представлено в третьому розділі.

2.3.4. Підготовка даних (Data Preparation)

Фаза підготовки даних має на меті формування фінальних наборів даних, придатних для моделювання, і включає три послідовні під-етапи: відбір (selection), попередня обробка (pre-processing) та трансформація (transformation).

- Відбір. Етап розпочинається з вибору даних для аналізу, при цьому оцінюється їхня однорідність, обґрунтованість та консистентність (узгодженість).

- Попередня обробка. На цьому етапі дані оцінюються, і залежно від їхньої значущості, вони або включаються, або виключаються з основного масиву даних, що підлягає дослідженню.

Необхідність обробки. Необроблені (сирі) дані (Raw data) є первинним джерелом для прийняття рішень, заснованих на даних. Однак, через особливості їх збору, ці дані, як правило, не можуть бути використані у своїй

початковій формі для алгоритмів машинного навчання. Обробка сирих даних є необхідною умовою для візуалізації та формування аналітичних висновків.

Мета етапу полягає у визначенні типу інформації, яка є необхідною, перш ніж приймати будь-які рішення щодо її остаточного використання. Для побудови моделі, здатної точно прогнозувати результати, обраний набір даних про продажі повинен пройти низку процесів обробки, кожен з яких є критично важливим для забезпечення точності прогнозу продажів.

2.3.5. Очищення даних (Data Cleaning)

Для успішної розробки прогнозних моделей критично важливим є належна підготовка та очищення зібраних даних. Процес очищення даних (Data Cleaning) забезпечує коректну категоризацію даних та заповнення виявлених інформаційних прогалин відповідними значеннями.

Очищення даних по суті включає корекцію або видалення неточних, пошкоджених, неформатованих, надлишкових або неповних даних з набору. Цей процес очищення може бути відображений у чотирьох основних кроках, як показано на рисунку 2.7.

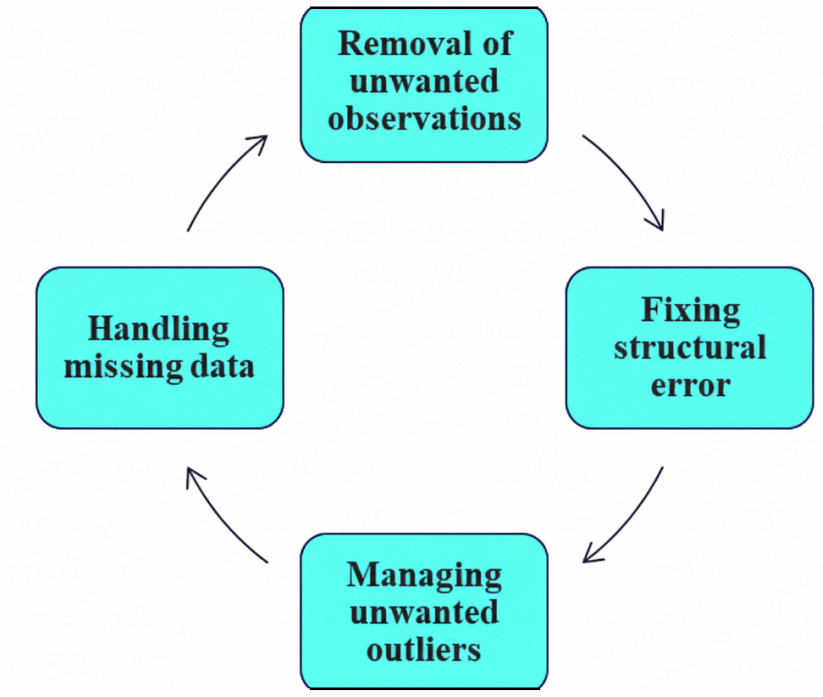


Рис. 2.7. Процес очищення даних

Проблема пропущених даних (Missing Data) може суттєво перешкоджати розробці моделей, оскільки ці значення можуть містити важливу інформацію для точніших прогнозів. Отже, виконання імпутації пропущених даних стає необхідним етапом. Залежно від конкретної задачі та характеру даних, для обробки пропущених значень можуть застосовуватися різні підходи. Нижче наведено деякі стандартні методи:

а) Видалення рядків (Row Deletion). Виключає будь-які спостереження з набору даних, які містять пропущені значення для будь-якої змінної. Недоліками цього методу є потенційна втрата інформації та зниження прогностичної потужності моделі.

б) Імпутація середнім/медіаною/модою (Mean/Median/Mode Imputation). Якщо змінна є неперервною, пропущені значення можуть бути замінені на середнє арифметичне або медіану всіх наявних значень цієї змінної. У випадку категоріальних змінних, для заміни пропущених значень може використовуватися мода (найбільш часте значення).

в) Створення прогностичної моделі (Prediction Model Imputation). Для імпутації пропущених значень у змінній також може бути розроблена прогностична модель. У цьому випадку змінна з пропущеними даними використовується як цільова змінна (target variable), а інші змінні виступають як предиктори. Набір даних поділяється на дві частини: одна без пропущених значень для цільової змінної, інша — з пропущеними значеннями. Прогностична модель, навчена на першому наборі, застосовується до другого для прогнозування відсутніх значень.

2.3.6 Розробка ознак (Feature Engineering)

Процес розробки ознак є ключовим компонентом архітектури всіх моделей машинного навчання (МН). Під час попередньої обробки сирі дані трансформуються в ознаки (features), придатні для використання алгоритмами МН, зокрема у прогностичних моделях. Оскільки прогностичні моделі складаються зі змінних-предикторів та цільових змінних-результатів,

на етапі розробки ознак генеруються та обираються найбільш доцільні змінні-предиктори.

Чотири основні процеси розробки ознак у МН включають: створення ознак (feature creation), трансформацію ознак (feature transformation), вилучення ознак (feature extraction) та відбір ознак (feature selection). Розробка ознак усуває варіації та флуктуації, які можуть виникнути в наборі даних під час фази розвідувального аналізу, тим самим покращуючи прогностичну потужність моделі.

2.3.7. Кодування категоріальних змінних (Encoding Categorical Values)

Для проведення кореляційного та регресійного аналізу вимагається перетворення категоріальних змінних у чисельні. Трансформація категоріальних даних у числовий формат є необхідною для підвищення ефективності моделі машинного навчання. Для цієї процедури будуть використані такі два методи:

а) Міткове кодування (Label Encoding). Цей метод присвоює числове значення кожній категорії в змінній. Він найкраще підходить для порядкових (ординальних) категоріальних змінних.

б) Одноразове кодування (One-Hot Encoding). Ця техніка передбачає перетворення кожної категорії змінної у новий бінарний формат (1/0) та додавання її як окремої ознаки. Це один із найбільш поширених методів, що дозволяє порівнювати кожен рівень числової змінної з попередньо визначеною відправною точкою.

Для кодування категоріальних значень у використовуваному наборі даних будуть застосовані обидва вищезазначені методи.

2.3.8. Кореляційний аналіз даних (Data Correlation)

Кореляційний аналіз даних є методом визначення зв'язку між змінними та прогнозування одного атрибута на основі іншого. Кореляція вважається позитивною, якщо збільшення однієї ознаки призводить до збільшення іншої.

Кореляція є негативною, якщо збільшення однієї ознаки призводить до зменшення іншої. Відсутність кореляції свідчить про відсутність зв'язку між двома атрибутами.

Атрибути з негативною кореляцією можуть бути виключені для підвищення ефективності моделі машинного навчання. Кореляція є статистичною мірою, яка кількісно оцінює лінійний зв'язок між двома змінними X та Y. Діапазон значень кореляції становить від −1 до 1.

Співвідношення значень кореляції:

а) Негативна кореляція: [0,−1]

б) Позитивна кореляція: (0,1]

в) Відсутність кореляції: 0

Кореляційний графік (correlation plot) візуалізує взаємозв'язки між усіма можливими парами змінних у даних. Чим сильніший зв'язок між змінними (тобто, чим ближче значення до 1 або −1), тим важливіше враховувати цей кореляційний зв'язок у процесі моделювання.

2.4. Розподіл наборів даних, побудова та оцінка моделі

Етап навчання алгоритму та його підгонка до розпізнавання патернів є критично важливим для створення прогностичної моделі. Після навчання модель має бути протестована на незалежному наборі даних для валідації її прогностичної здатності. З метою запобігання перенавчанню (overfitting), в даному дослідженні не використовуються два окремі імпортовані набори даних. Натомість, розподіл на тренувальний та тестовий набори здійснюється в межах єдиного вихідного набору даних. Співвідношення цього розподілу може бути оптимізовано для досягнення вищої точності моделі.

Фаза побудови моделі передбачає тренування алгоритмів машинного навчання, а потім їхню оцінку за допомогою відповідних даних. Після завершення підготовки набору даних створюється прогностична модель машинного навчання шляхом навчання на тренувальних даних.

Вибір типу моделі залежить від природи цільової змінної:

- Якщо цільова змінна є якісною (дискретною), створюється модель класифікації.

- Якщо цільова змінна є кількісною (неперервною), розробляється модель регресії.

Оцінка моделі є необхідною для визначення найбільш придатного алгоритму для даного набору даних у контексті конкретної задачі. Цей підхід дозволяє порівняти продуктивність різних моделей машинного навчання за умови використання ідентичного вхідного набору даних. Основний акцент методу оцінки зосереджений на точності прогнозування кінцевого результату моделлю.

2.4.1. Методи оцінки моделі

Існують два основні методи для оцінки продуктивності моделі: метод відкладеної вибірки (Holdout) та крос-валідація (Cross-Validation).

Метод відкладеної вибірки - набір даних випадковим чином поділяється на три взаємовиключні підмножини: тренувальний, валідаційний та тестовий набори даних.

Крос-валідація - початковий набір спостережень поділяється на тренувальний набір (використовується для навчання моделі) та незалежний набір (використовується для оцінки).

Найпоширенішим варіантом є k-кратна крос-валідація (k-fold cross-validation), яка передбачає поділ початкового набору даних на k рівних частин (фолдів). k є параметром, заданим дослідником, і зазвичай рекомендується встановлювати його значення на 5 або 10. Процес повторюється k разів: щоразу одна з k підмножин використовується як тестовий/валідаційний набір, а решта $k-1$ підмножин об'єднуються для формування тренувального набору. Для визначення загальної ефективності моделі обчислюється середнє значення оцінки помилки за всіма k ітераціями.

2.4.2. Метрики оцінки моделі

Для кількісного вимірювання продуктивності моделі необхідні метрики оцінки. Вибір метрик залежить від специфіки задачі машинного навчання (класифікація, регресія, кластеризація тощо).

Для завдання класифікації типові метрики включають точність класифікації, матрицю плутанини (confusion matrix), логарифмічні втрати, площу під кривою (AUC) та F-міру.

Для завдання регресії найчастіше використовуються середня абсолютна помилка (MAE) та середньоквадратична помилка (RMSE).

Оскільки сформульована проблема для цього проєкту є завданням регресії, оцінка моделі буде проводитися з використанням регресійних метрик.

MAE представляє середнє значення абсолютних різниць між прогнозованим значенням моделі (y^i) та фактичним значенням (yi).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Де N — загальна кількість точок даних, yi — фактичне значення, y^i — прогнозоване значення.

RMSE (Root Mean Squared Error) є квадратним коренем середнього значення квадратів різниць між очікуваними (y^j) та фактичними (yj) значеннями наданого набору даних. RMSE розраховується як квадратний корінь із MSE (середньоквадратичної помилки). Це найпоширеніший метод оцінки, який застосовується до задач регресії, заснований на припущенні, що помилки є незміщеними та мають нормальний розподіл. Вище значення RMSE свідчить про більшу розбіжність між прогнозованими та фактичними значеннями.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Де n — загальна кількість точок даних, yj — фактичне значення, y^j — прогнозоване значення.

Коефіцієнт детермінації (R2) обчислює частку дисперсії залежної змінної, яка пояснюється незалежними змінними. Це загальна метрика для оцінки точності моделі, яка вказує, наскільки тісно точки даних відповідають підігнаній регресійній лінії. Вище значення R2 (ближче до 1) свідчить про кращу відповідність.

$$R^2 = 1 - \frac{SSE}{SST}$$

де: SSE (сума квадратів помилок, Sum of Squares of Errors) — сума квадратів різниць між фактичним значенням (yi) та прогнозованим значенням (y^i). SST (загальна сума квадратів, Total Sum of Squares) — сума квадратів, що представляє різницю між фактичним значенням (yi) та його середнім ($\bar{y}$).

Значення R^2 коливаються в діапазоні [0,1]. Оптимально, щоб значення R^2 було максимально наближеним до 1. Якщо R^2=0, модель не перевершує випадкову модель (baseline model). Негативне значення R^2 свідчить про те, що регресійна модель є некоректною.

2.4.3. Важливість ознак (Feature Importance)

Концепція важливості ознак (Feature Importance) стосується процесу присвоєння вагових коефіцієнтів ознакам, які формують прогностичну модель. Ці коефіцієнти оцінюють корисність кожної окремої змінної для моделі та її прогностичного результату.

Оцінки важливості ознак надають інформацію про:

- Найбільш та найменш важливі ознаки для прогностичної здатності моделі.

- Можливість покращення прогностичної моделі шляхом відбору ознак (feature selection).

Використовуючи ці оцінки, можна вибрати ознаки з вищими оцінками для збереження та відкинути менш значущі. Такий відбір ознак сприяє спрощенню задачі моделювання, прискоренню процесу навчання та, в деяких випадках, підвищенню загальної продуктивності моделі.

2.5. Методика розгортання моделі і опис платформи для реалізації

Ефективність прогностичної моделі визначається її доступністю для кінцевого користувача та можливістю практичного застосування результатів. Отже, фаза розгортання (Deployment) повинна плануватися з урахуванням вимог до реального використання результатів проєкту.

На цьому етапі розробляється стратегія розгортання, яка може варіюватися від простого звіту до складної реалізації моделі прогнозування в режимі реального часу. Ключові дії включають:

- Складання фінальних звітів.

- Аналіз усього процесу для ідентифікації помилок та визначення необхідності ітеративного повторення попередніх етапів.

- Встановлення стратегії моніторингу та управління результатами моделі інтелектуального аналізу даних з метою оцінки їхньої корисності та підтримки актуальності.

Процес аналізу великих даних та вилучення знань з них є комплексним завданням. Це вимагає використання потужного та ефективного інструменту інтелектуального аналізу даних (Data Mining) для обробки складних наборів даних та підтримки прийняття рішень у майбутньому.

У цьому проєкті як основний інструмент використано R — потужне, безкоштовне, відкрите середовище та мова програмування для статистичного аналізу даних. R є незамінним елементом статистики та включає процес конвертації даних у знання, розуміння та інсайти.

R вирізняється широким спектром вбудованих методів статистичного моделювання та інструментів машинного навчання, що дозволяє розробляти

продукти даних та проводити відтворювані дослідження. Незважаючи на наявність інших інструментів для обробки великих даних, R забезпечує унікальні переваги завдяки великій кількості алгоритмів від сторонніх розробників та вбудованих статистичних формул.

Для роботи з R використовувалося інтегроване середовище розробки (IDE) RStudio. RStudio забезпечує комплексну підтримку робочого процесу аналізу даних, включаючи:

- Консоль та редактор із підсвічуванням синтаксису.

- Інструменти для побудови графіків, налагодження та управління робочим простором.

У рамках розробки програмного прототипу проєкту, графічний інтерфейс користувача (GUI) був реалізований за допомогою пакета R Shiny. Shiny — це відкритий пакет, розроблений RStudio, PBC, який надає елегантну та просту у використанні веб-платформу для створення інтерактивних веб-додатків та дашбордів без необхідності попереднього знання HTML, CSS або JavaScript.

Висновки до розділу

У другому розділі здійснено аналіз моделей та алгоритмів машинного навчання, придатних для прогнозування обсягів продажів. Порівняльне дослідження лінійної регресії, дерева рішень, випадкового лісу та XGBoost показало перевагу ансамблевих методів за точністю та стабільністю результатів. Розроблено архітектуру системи прогнозування, що базується на методології CRISP-DM та охоплює етапи збору, очищення, інженерії ознак і навчання моделей. Визначено основні метрики оцінювання ефективності моделей — MAE, RMSE, R^2 — та проведено аналіз важливості ознак. Запропонована методика забезпечує підвищення точності прогнозів і може бути використана як основа для створення аналітичних систем підтримки бізнес-рішень.

РОЗДІЛ 3. ПРОГРАМНА РЕАЛІЗАЦІЯ МЕТОДІВ ПРОГНОЗУВАННЯ ПРОДАЖІВ НА ОСНОВІ ТЕХНОЛОГІЙ МАШИННОГО НАВЧАННЯ

3.1. Опис набору даних

Цей розділ містить детальний аналіз набору даних із його графічним представленням, а також опис реалізації моделі та прототипу рішення. Тут послідовно висвітлюються структура даних, механізм проведення експериментів, аналіз даних у контексті завдань моделювання, графічний інтерфейс користувача (GUI) прототипу та оцінка продуктивності розроблених прогностичних моделей.

Для цього проєкту використано набір даних, що містить інформацію про продажі товарів у різних супермаркетах. Дані було отримано з відкритого онлайн-репозиторію Kaggle.

Набір даних включає як категоріальні, так і числові вхідні змінні, а також неперервну вихідну змінну, що визначає задачу як завдання регресії. Набір поділено на тренувальний та тестовий набори.

Тренувальний набір містить 8523 унікальних спостереження (товари) з 12 атрибутами (11 вхідних та 1 вихідна змінна). Товари розподілені по різних локаціях та містах.

Тестовий набір містить 5681 спостереження з 11 атрибутами. Для цього набору необхідно спрогнозувати продажі.

3.1.1. Атрибути набору даних

Перелік та опис атрибутів, які формують набір даних, наведено нижче:

Item_Identifier - Унікальний ідентифікатор продукту.

Item_Weight - Вага продукту.

Item_Fat_Content - Вміст жиру в продукті.

Item_Visibility - Частка загальної площі вітрини магазину, виділена для продукту (видимість).

Item_Type - Категорія продукту.

Item_MRP - Рекомендована роздрібна ціна товару.

Outlet_Identifier - Унікальний ідентифікаційний номер магазину.

Outlet_Establishment_Year - Рік заснування магазину.

Outlet_Size - Розмір магазину (категоріальний).

Outlet_Location_Type - Класифікація міста, де розташований магазин.

Outlet_Type - Тип торгової точки (супермаркет або продуктовий магазин).

Item_Outlet_Sales - Обсяг продажів продукту в кожному магазині.

Атрибут Item_Outlet_Sales виступає як цільова змінна (змінна відгуку), яку необхідно прогнозувати, тоді як усі інші атрибути використовуються як предикторні змінні. Цей набір даних є цінним ресурсом, що містить широкий спектр прихованих патернів, які можуть бути виявлені для підвищення точності прогнозування.

```
$ Item_Identifier          : chr  "FDA15" "DRC01" "FDN15" "FDX07" ...
$ Item_Weight              : num  9.3 5.92 17.5 19.2 8.93 ...
$ Item_Fat_Content         : chr  "Low Fat" "Regular" "Low Fat" "Regular" ...
$ Item_Visibility          : num  0.016 0.0193 0.0168 0 0 ...
$ Item_Type                : chr  "Dairy" "Soft Drinks" "Meat" "Fruits and Vegetables" ...
$ Item_MRP                 : num  249.8 48.3 141.6 182.1 53.9 ...
$ Outlet_Identifier        : chr  "OUT049" "OUT018" "OUT049" "OUT010" ...
$ Outlet_Establishment_Year: int  1999 2009 1999 1998 1987 2009 1987 1985 2002 2007 ...
$ Outlet_Size              : chr  "Medium" "Medium" "Medium" "" ...
$ Outlet_Location_Type     : chr  "Tier 1" "Tier 3" "Tier 1" "Tier 3" ...
$ Outlet_Type              : chr  "Supermarket Type1" "Supermarket Type2" "Supermarket Type1" "Grocery Store" ...
$ Item_Outlet_Sales        : num  3735 443 2097 732 995 ...
```

Рис. 3.1. Структура даних

3.1.2. Попередній перегляд даних

Для ознайомлення зі структурою даних наведено попередній перегляд тренувального та тестового наборів у таблицях 3.1 та 3.2 відповідно.

Таблиця 3.1.

Попередній перегляд тренувального набору даних

Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
DRCO1	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228

Таблиця 3.2.

Попередній перегляд тестового набору даних

Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type
FDNY06	20.75	Low Fat	0.007564656	Snack Foods	107.8622	OUT049	1999	Medium	Tier 1	Supermarket Type1
FDNY14	8.3	reg	0.038437677	Dairy	87.3198	OUT017	2007		Tier 2	Supermarket Type1

У тестовому наборі наявні пропущені значення, що потребує подальшої обробки.

3.2. Аналіз даних та візуалізація цільової змінної

Усі графічні представлення в межах цього проєкту були реалізовані за допомогою інструментального пакета ggplot2 в середовищі R.

Цільова змінна Item_Outlet_Sales (Продажі товарів у точці продажу), будучи неперервною змінною, була візуалізована за допомогою гістограми (рис. 3.2).

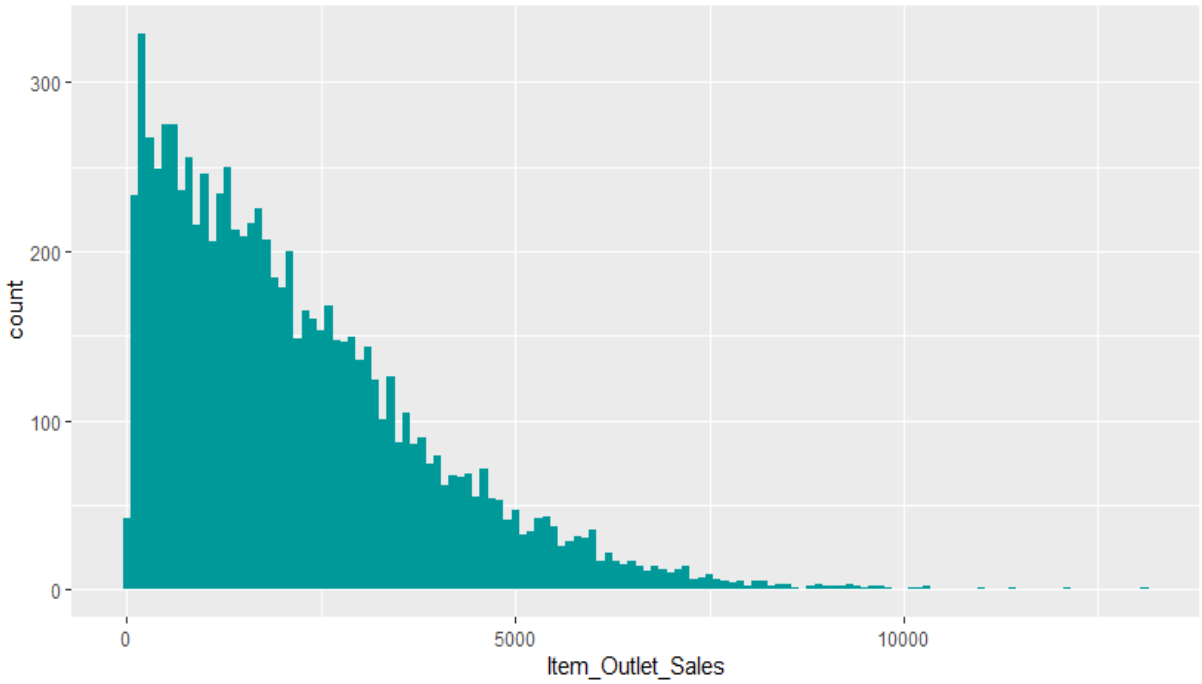


Рис. 3.2. Розподіл цільової змінної Item_Outlet_Sales

На основі гістограми встановлено, що розподіл Item_Outlet_Sales має виражену правосторонню асиметрію (positive skewness). Це свідчить про необхідність застосування методів трансформації даних (наприклад, логарифмування) для корекції асиметрії та забезпечення кращої відповідності припущенням прогностичних моделей.

Для відображення розподілу незалежних числових та категоріальних змінних було проведено одновимірний аналіз із застосуванням відповідних графічних інструментів (гістограм та діаграм частот).

3.2.1. Розподіл незалежних числових змінних

Аналіз розподілу незалежних числових змінних, представлений на рисунку 3.3, дозволяє сформулювати такі висновки.

- Item_Weight (вага товару) - розподіл не демонструє чітких закономірностей. Обсяги продажів (Item_Outlet_Sales) розподілені випадковим чином по всьому діапазону значень ваги.

- Item_MRP (рекомендована роздрібна ціна) - спостерігається наявність чотирьох виражених піків або кластерів, що вказує на дискретні рівні ціноутворення.

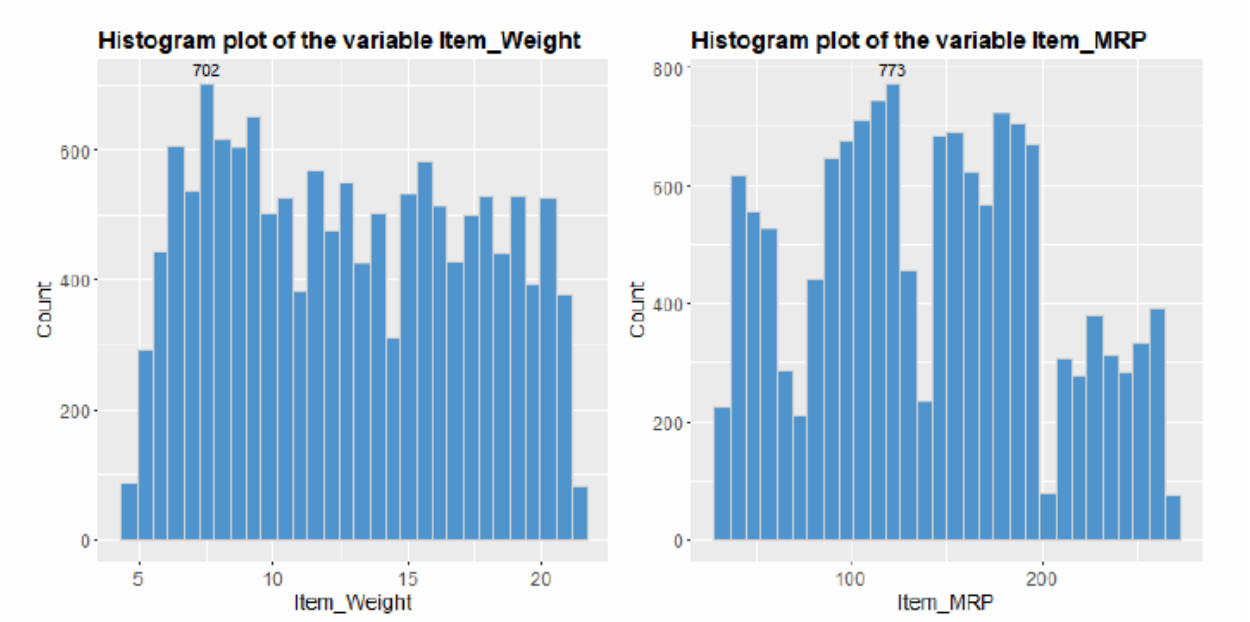


Рис. 3.3. Розподіл незалежних числових змінних Item_Weight, Item_MRP

3.2.2. Розподіл незалежних категоріальних змінних

Незалежні категоріальні змінні (ознаки) можуть набувати лише скінченного набору значень.

Діаграма частот для змінної Item_Fat_Content (рис. 3.4) спочатку виявила неконсистентність у маркуванні: дві основні категорії, Low Fat та Regular Fat, були представлені під різними ідентифікаторами ('low fat', 'LF' замість 'Low Fat'; 'reg' замість 'Regular'). Було встановлено, що більшість товарів містять низький вміст жиру.

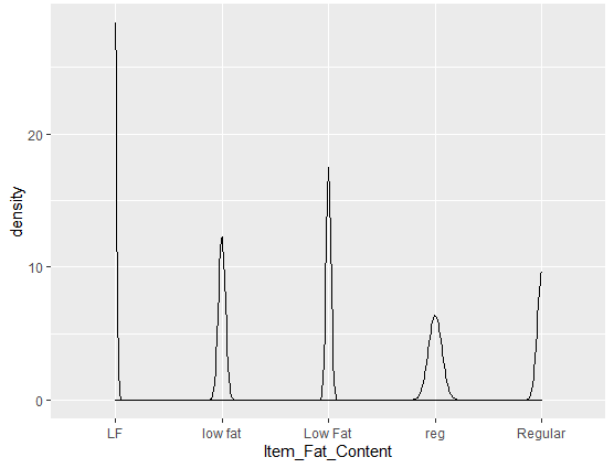


Рис. 3.4. Діаграма частот для Item_Fat_Content (вихідні дані)

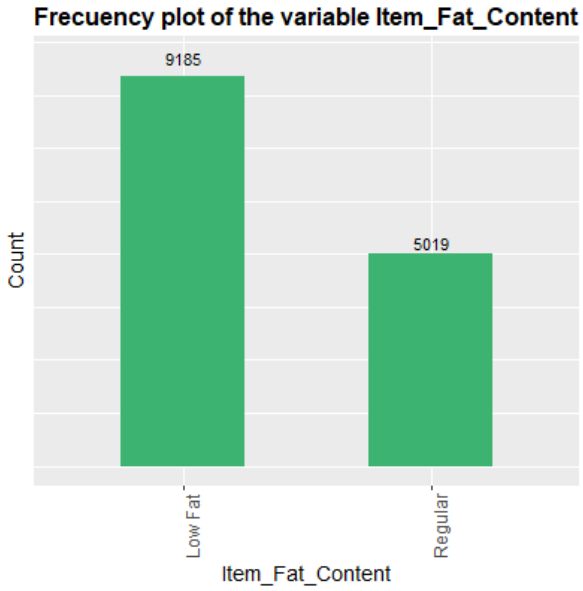


Рис. 3.5. Діаграма частот для Item_Fat_Content (після об'єднання категорій)

Для коректного аналізу було проведено уніфікацію категорій (Regular та Low Fat). На рисунку 3.5 представлено графік після об'єднання, який підтверджує, що кількість продуктів із низьким вмістом жиру суттєво переважає продукти зі звичайним вмістом жиру.

3.2.3. Комплексний аналіз категоріальних змінних

Діаграми частот для решти категоріальних змінних (рисунок 3.6) дозволяють зробити такі висновки:

Рис. 3.6. Розподіл незалежних категоріальних змінних Outlet_Size, Item_Type, Outlet_Identifier, Outlet_Type

- Outlet_Size (розмір торгової точки). Виявлено, що 4016 записів для цього атрибута є пропущеними або порожніми. Обробка цих пропущених значень вимагатиме двовимірного аналізу. Крім того, лише незначна частка торгових точок класифікована як Large (Великий) або Huge (Величезний).

- Item_Type (категорія товару). Найбільш поширеними категоріями товарів є ’Fruits and Vegetables’ (Фрукти та овочі), за якими йдуть ’Snack Foods’ (Снеки). Найменш представленою категорією є ’Seafood’ (Морепродукти).

- Outlet_Identifier (Ідентифікатор торгової точки). Найбільша кількість магазинів відповідає ідентифікаторам OUT027 та OUT013. Ідентифікатор OUT019 представляє найменшу кількість магазинів.

- Outlet_Type (тип торгової точки): Найбільш поширеним типом є Supermarket Type1.

3.3. Проведення двовимірного аналізу даних

Була візуалізована кореляція між цільовою змінною (Item_Outlet_Sales) та ключовими категоріальними змінними, що характеризують торгову точку (Outlet_Identifier, Outlet_Size, Outlet_Location_Type) і показано на рисунку 3.7.

Виявлено, що продукти з низьким вмістом жиру мають високий рівень продажів у магазинах, що відповідають таким характеристикам: середній розмір (Medium), розташовані у Tier 3 та належать до Supermarket Type 3. Крім того, магазини, засновані до 1990 року, також демонструють вищі продажі продуктів цієї категорії.

Розподіл продажів за категоріями товарів (Item_Type) є відносно однорідним і має схожі патерни з розподілом за вмістом жиру. Однак, помітно, що побутові товари (Household) демонструють стабільно високий рівень продажів у всіх категоріях змінних магазину (рис. 3.8).

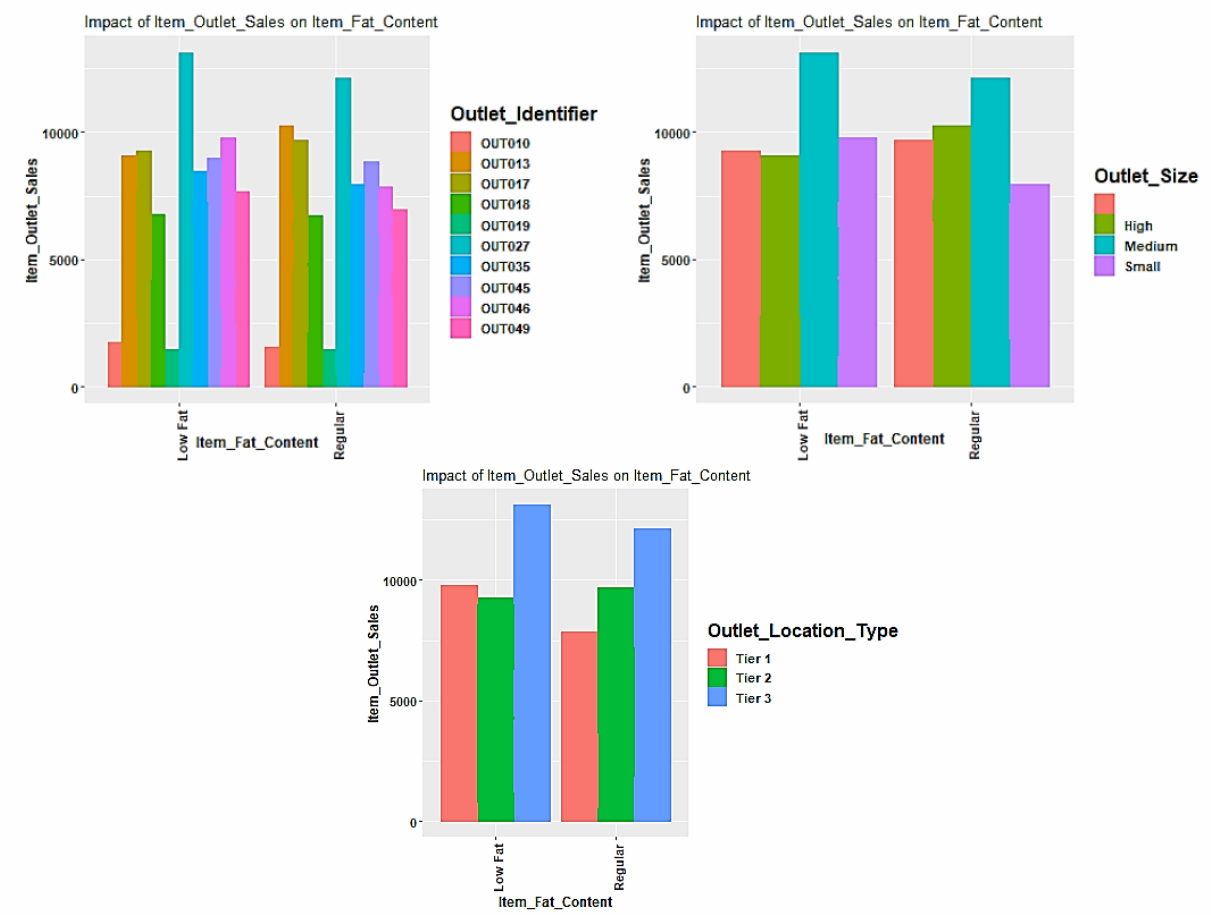


Рис. 3.7. Item_Outlet_Sales проти Item_Fat_Content, згруповані за змінними магазину

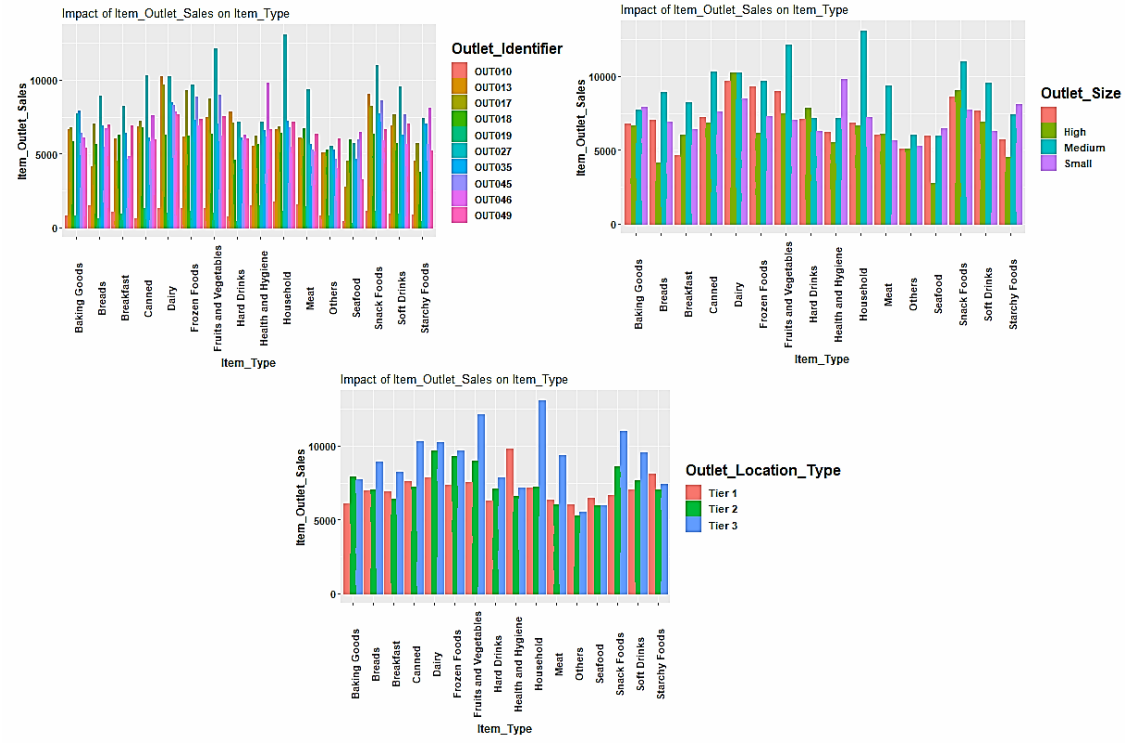


Рис. 3.8. Item_Outlet_Sales проти Item_Type, згруповані за змінними магазину

Аналіз ідентифікаторів точок продажу виявив, що магазини OUT010 та OUT019 характеризуються невеликим розміром і мають найнижчі показники продажів. Інші магазини демонструють значно вищі обсяги продажів. Магазин OUT027 виділяється найвищими продажами серед усіх точок. Цей магазин, розташований у Tier 3, має середній розмір (Medium) і є єдиним представником категорії Supermarket Type 3 у наборі даних.

Візуалізація підтвердила наявність пропущених даних у змінній Outlet_Size, що ускладнює повне розуміння впливу розміру на продажі. Магазин OUT013, хоча є єдиною точкою великого розміру (Large), не демонструє найвищих продажів. Grocery Store має найнижчі продажі порівняно з іншими типами.

Магазини середнього розміру (Medium) мають найбільшу частку товарів із високими продажами та демонструють зіставні результати з магазинами великого розміру.

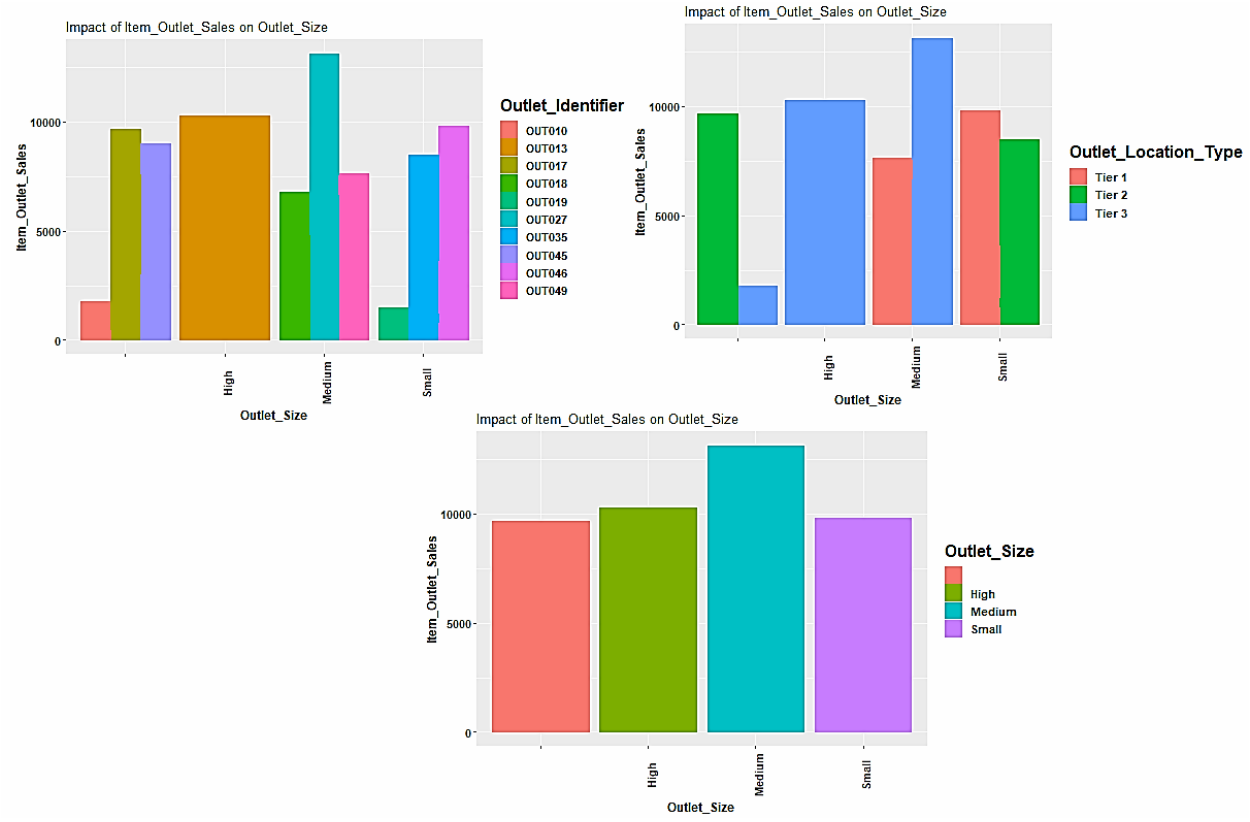


Рис. 3.9. Item_Outlet_Sales проти Outlet_Size, згруповані за змінними магазину

Аналіз року заснування магазинів показав, що найстаріший магазин (OUT027), заснований у 1985 році, має найбільший позитивний вплив на загальний обсяг продажів. Встановлено, що локації Tier 3 мають найвищі загальні продажі. Продажі у локаціях Tier 1 та Tier 2 є порівняно схожими за обсягами.

3.4. Виконання очищення даних та інженерія ознак

Після завершення розвідувального аналізу даних (EDA) та візуалізації необхідно перейти до етапу підготовки даних, ураховуючи висновки, отримані на попередній фазі. Цей процес вимагає обробки викидів та імпутації пропущених значень.

Було ідентифіковано, що пропущені дані наявні у двох атрибутах: Item_Weight (числова змінна, 2439 спостережень) та Outlet_Size (категоріальна змінна, 4016 спостережень).

Стратегії:

- Item_Weight: Відсутні значення були імпутовані середнім значенням ваги, розрахованим для кожного Item_Identifier (унікального ідентифікатора продукту).

- Outlet_Size: Оскільки ця змінна є категоріальною, імпутація середнім значенням є неприйнятною. Замість цього було використано модальне значення (найчастіше значення) розміру, розраховане для кожного Outlet_Identifier (ідентифікатора торгової точки).

- Item_Visibility: Нульові значення (0.0) у полі Item_Visibility було інтерпретовано як пропущені значення, оскільки нульова видимість для продаваного товару є аномалією. Ці нульові значення були замінені середнім значенням видимості, розрахованим для відповідного Item_Identifier.

- Гістограма Item_Visibility до та після корекції (рисунок 3.10) демонструє, що після заміни нульових значень на середні, розподіл видимості став більш реалістичним і менш зміщеним.

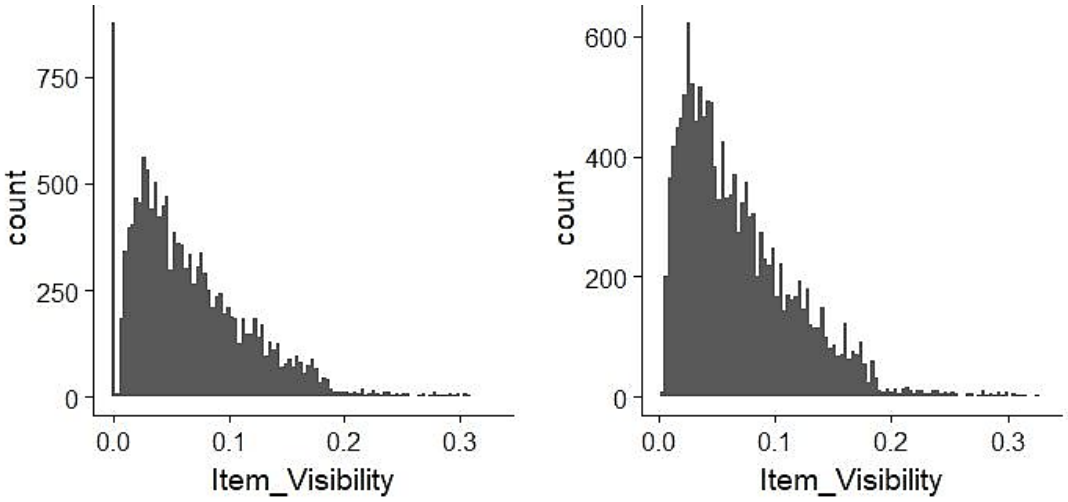


Рис. 3.10. Гістограма Item_Visibility до та після заміни нульових значень

Інженерія ознак є критичним етапом, спрямованим на створення нових ознак та трансформацію наявних для підвищення прогностичної потужності моделі.

3.4.1. Вибір ознак для трансформації

З числових змінних було встановлено, що Outlet_Establishment_Year має низький прогностичний вплив у вихідній формі. Серед категоріальних змінних було ідентифіковано 7 стовпців для модифікації або кодування (таблиця 3.3).

Таблиця 4.4.

Вибір ознак для інженерії ознак

Категорійні значення		**Числові значення**
Порядкові змінні:	Номінальні змінні:	
а) Item_Fat_Content	а) Item_Identifier	а) Outlet_Establishment_Year
б) Outlet_Size	б) Item_Type	
в) Outlet_Location_Type	в) Outlet_Identifier	
	г) Outlet_Type	

3.4.2. Стратегія кодування та видалення

На основі аналізу, проведеного в попередньому підрозділі, атрибути Outlet_Establishment_Year, Item_Identifier та Outlet_Identifier були визнані недостатньо значущими у їхній поточній формі та підлягають видаленню.

Порядкові змінні (таблиця 3.3) будуть закодовані за допомогою Міткового Кодування (Label Encoding). Номінальні змінні Outlet_Type та Item_Type будуть закодовані за допомогою одноразового кодування (One Hot Encoding).

3.4.3. Створення нових прогностичних ознак

Оскільки наявні ознаки виявилися недостатніми для задовільного прогнозування, було застосовано два методи створення нових ознак:

1) Видобування прихованих ознак (з існуючих даних),

2) Додавання зовнішніх ознак (додаткових пояснювальних факторів).

Категорії нових ознак, розроблених у рамках першої методології:

а) Item_category: Створена категоріальна змінна на основі префіксів Item_Identifier. Оскільки кожен унікальний ідентифікатор починається з префіксів FD, DR або NC, було додано новий стовпець Item_category з категоріями: Foods, Drinks та Non-consumables (Непродовольчі товари).

Корекція Item_Fat_Content: Для непродовольчих товарів (Item_category=’NC’) значення Item_Fat_Content було змінено на відповідне значення, оскільки вони не можуть містити жиру.

б) Outlet_Years: Числова ознака, що відображає кількість років роботи магазину, розрахована на основі Outlet_Establishment_Year та поточного року.

в) Item_Type_new: Широка класифікація, створена на основі змінної Item_Type, шляхом групування категорій на perishable (швидкопсувні) та non_perishable (не швидкопсувні).

г) price_per_unit_wt: Нова числова ознака, що визначає ціну за одиницю ваги товару, розрахована як відношення Item_MRP до Item_Weight.

Категорії нових ознак, розроблених у рамках другої методології:

Другий підхід передбачає видалення рядків з нульовими значеннями (як альтернатива імпутації) та додавання додаткових пояснювальних атрибутів. На основі двовимірного аналізу (Item_MRP vs. Item_Outlet_Sales), де було виявлено 4 цінові сегменти, змінна Item_MRP була використана для створення 4 груп за допомогою методу кластеризації k-середніх (k-Means).

а) Item_MRP_clusters: Змінна, що представляє кластер (групу), до якого належить Item_MRP. Оскільки метод k-Means вимагає попереднього знання k (кількості кластерів), процес було продовжено з k=4, відповідно до емпіричних спостережень. До набору даних було додано стовпець, що містить ідентифікатори цих кластерів.

3.5. Кореляційний аналіз даних та побудова прогностичної моделі

Для виявлення прихованих закономірностей між змінними в обробленому наборі даних було проведено кореляційний аналіз з візуалізацією за допомогою пакета corrplot у середовищі R. Цей інструмент дозволяє візуалізувати кореляційні матриці та автоматично перевпорядковувати змінні.

У візуалізації, представленій на рисунку 3.11, кореляція між будь-якою парою змінних відображається кольоровим квадратом. Блакитні квадрати позначають позитивну кореляцію, тоді як червонуваті — негативну. Інтенсивність кольору квадрата вказує на силу (величину) кореляційного зв'язку. Аналіз цієї діаграми є критичним для визначення подальших дій.

Основні висновки кореляційного аналізу:

а) Item_Visibility та Item_Outlet_Sales. Виявлено негативну кореляцію. Це означає, що збільшення видимості товару (Item_Visibility) може призводити до зниження продажів (Item_Outlet_Sales). Цей висновок суперечить загальноприйнятим бізнес-припущенням про пряму залежність продажів від видимості товару.

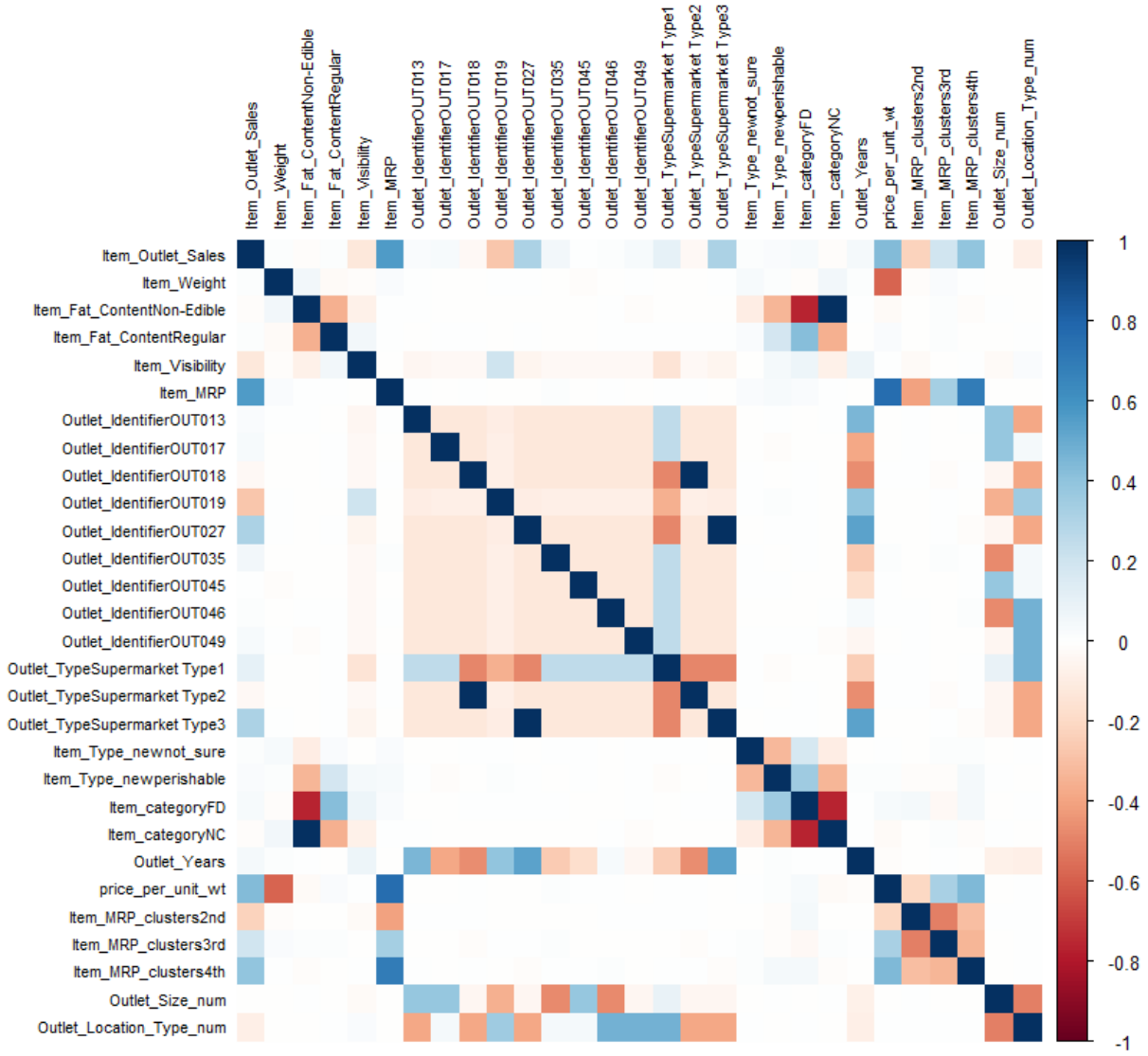


Рис. 3.11. Діаграма кореляції між різними факторами

б) Item_MRP та Item_Outlet_Sales. Спостерігається позитивна кореляція між рекомендованою роздрібною ціною товару (Item_MRP) та його продажами. Це підтверджує, що Item_MRP є важливою ознакою для оцінки Item_Outlet_Sales.

в) Supermarket_Type3: Тип магазину Supermarket_Type3 демонструє позитивну кореляцію з продажами, що узгоджується з висновком про його високу прибутковість.

г) price_per_unit_wt, Item_Weight та Item_MRP: Ознака price_per_unit_wt (ціна за одиницю ваги), будучи похідною від Item_Weight

та Item_MRP, демонструє дуже сильні кореляційні зв'язки з обома цими вихідними змінними.

Після кількох ітерацій підготовки та обробки даних було сформовано валідований набір даних, придатний для побудови прогностичної моделі. Дані були розподілені на тренувальний та тестовий набори, після чого моделі були навчені для кожного обраного алгоритму та оцінені за допомогою відповідних оціночних метрик.

Оскільки проєкт передбачає прогнозування числового значення (Item_Outlet_Sales), він базується на регресійній моделі.

Для аналізу продуктивності та оцінки ефективності різних методологічних підходів у прогнозуванні продажів були розгорнуті та протестовані наступні моделі машинного навчання:

а) Лінійна Регресія (Linear Regression)

б) Випадковий Ліс (Random Forest)

в) Дерево Рішень (Decision Tree)

г) Регресор XGBoost (XGBoost Regressor)

Ці алгоритми використовують різні механізми навчання та були ретельно відібрані, оскільки вони є поширеними та добре зарекомендованими в галузі прогностичного моделювання продажів, що підтверджується попередніми дослідженнями.

Для оцінки узагальнюваності розроблених моделей на нових даних застосовувалася 5-кратна крос-валідація (5-fold cross-validation).

Для побудови моделей використовувалися відповідні пакети та бібліотеки R: xgboost, rpart та randomForest.

1. Для навчання лінійної регресії використовувався метод ’lm’,

2. Для дерева рішень — метод ’rpart’,

3. Для моделі випадкового лісу — метод ’ranger’,

4. Для XGBoost використовувалася функція xgb.cv(), що входить до пакету XGBoost.

3.6. Представлення архітектура та опис інформаційної панелі прогнозування продажів

Архітектура програмного рішення для прогнозування продажів реалізована у форматі інформаційної панелі (дашборду), яка забезпечує інтерактивну взаємодію користувачів з даними та прогностичними моделями. Структура цієї архітектури представлена на рисунку 3.12.

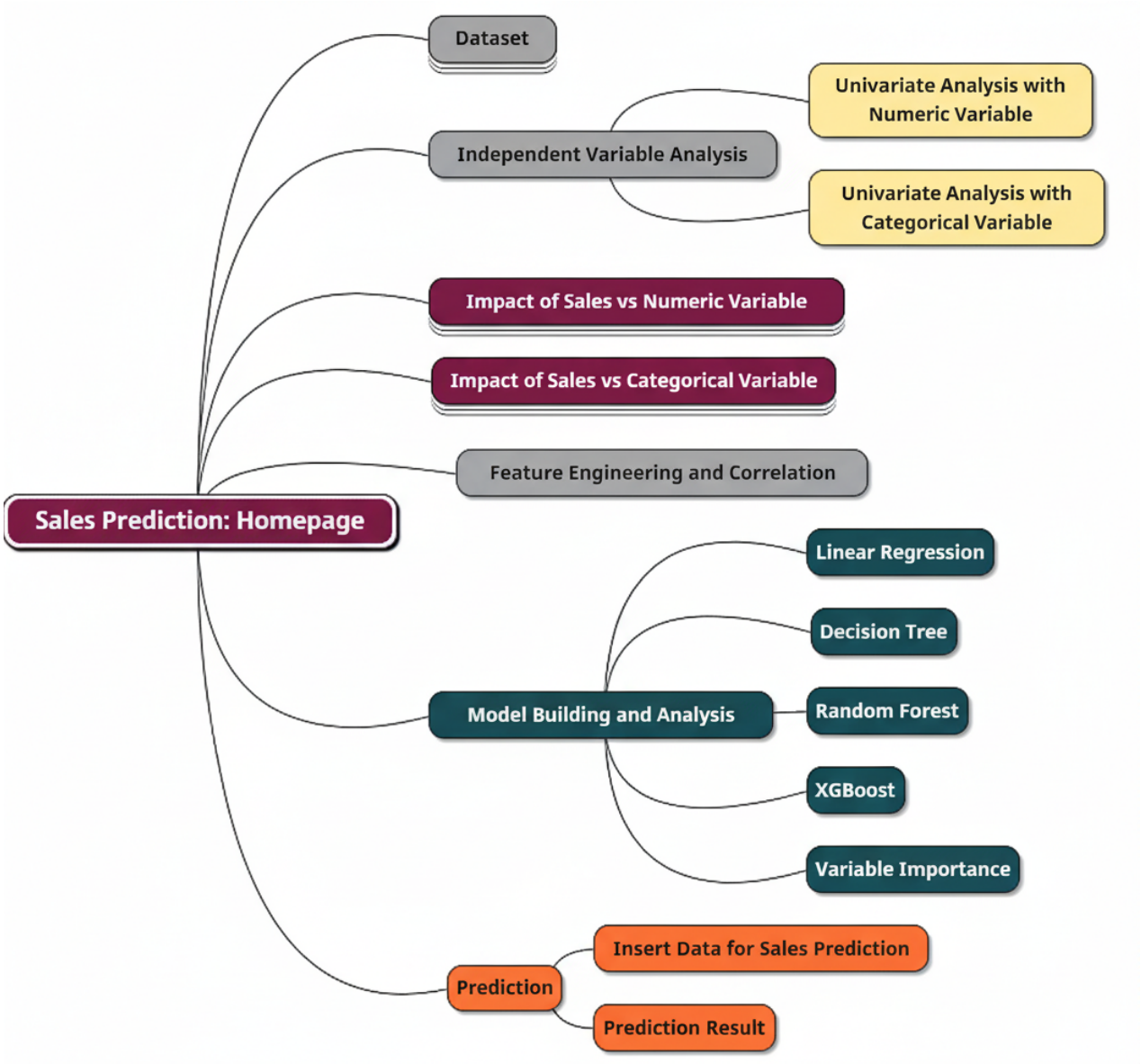
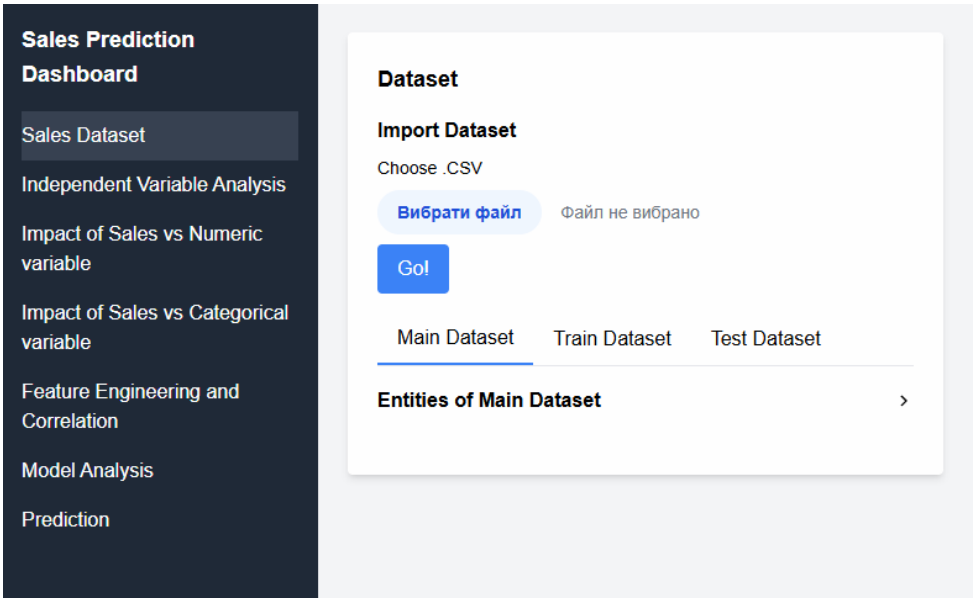


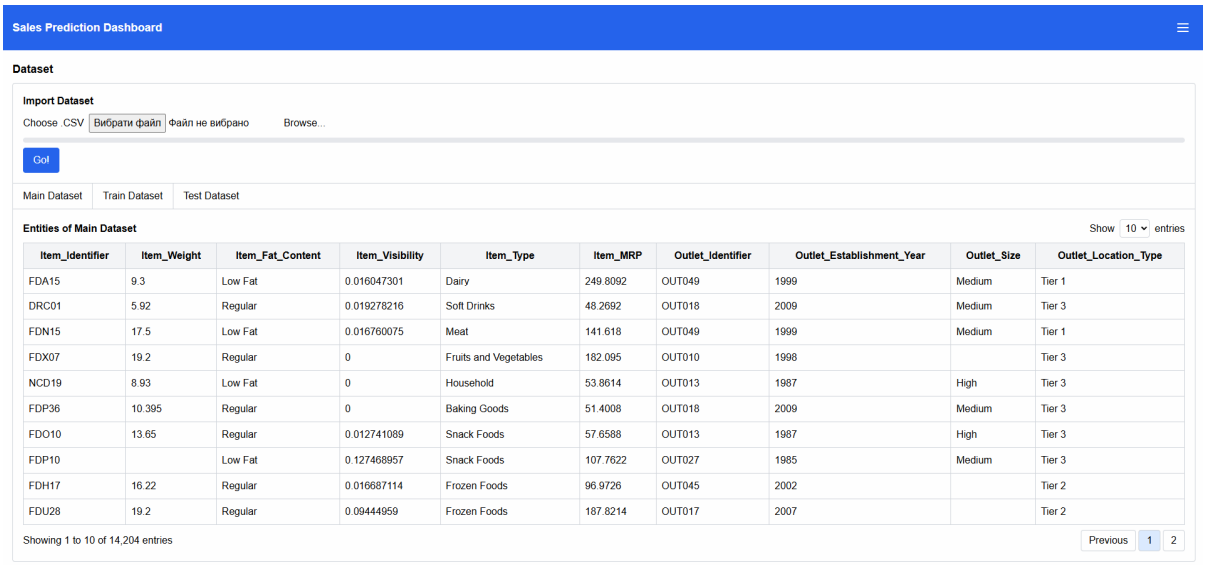
Рис. 3.12. Архітектура дашборду для прогнозування продажів на основі машинного навчання

Дашборд складається з семи ключових пунктів меню (модулів), які організовують робочий процес від завантаження даних до отримання фінальних прогнозів.

Модуль «Набір даних продажів» є початковою точкою взаємодії, що дозволяє користувачам завантажувати та переглядати вихідні набори даних про продажі. Рисунок 3.13 ілюструє початковий вигляд інтерфейсу та його стан після успішного імпорту набору даних.



а) Початковий перегляд



б) Вигляд після імпорту набору даних

Рис. 3.13. Дашборд для прогнозування продажів

Модуль «Аналіз Незалежних Змінних» надає інструменти для одновимірного аналізу незалежних змінних, таких як Item_Weight, Item_Fat_Content та Item_Visibility, відображаючи їхні розподіли та частоти. Його вигляд представлений на рисунку 3.14.

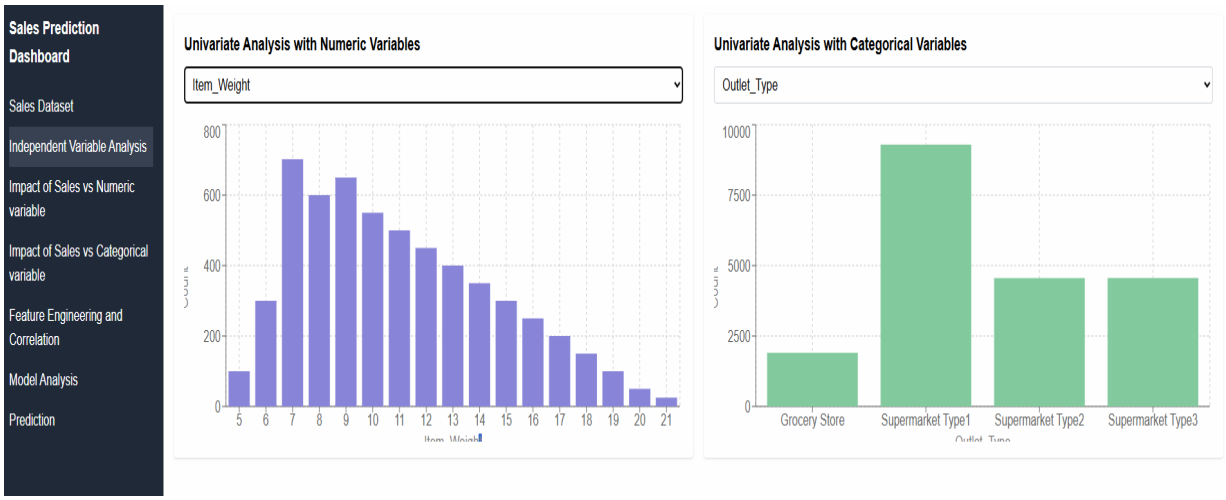


Рис. 3.14. Вигляд дашборду для аналізу незалежних змінних

Модуль «Вплив продажів на числові змінні» візуалізує двовимірні взаємозв'язки між цільовою змінною (Item_Outlet_Sales) та числовими предикторами (Item_Weight, Item_MRP, Item_Visibility), сприяючи ідентифікації кореляцій (рисунок 3.15).

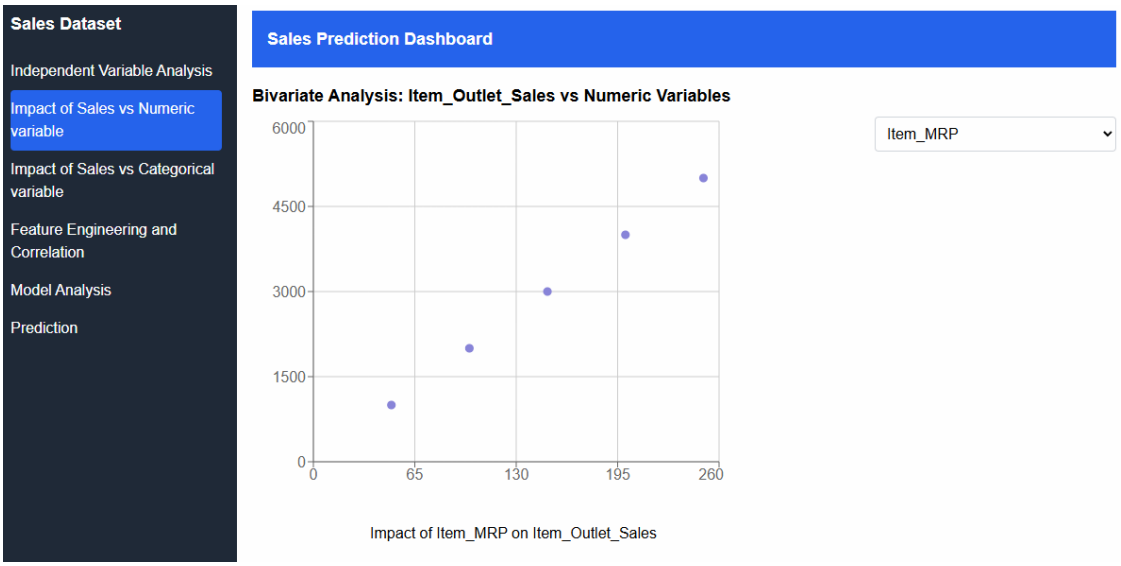


Рис. 3.15. Дашборд для аналізу впливу продажів на числові змінні

Модуль «Вплив продажів на категорійні змінні» призначений для аналізу впливу категоріальних змінних (Item_Fat_Content, Item_Type, Outlet_Size тощо) на Item_Outlet_Sales, що відображає відмінності у продажах залежно від характеристик товару та точки продажу. Модуль «Інженерія ознак та кореляція» надає інтерфейс для перегляду результатів інженерії ознак та аналізу кореляційної матриці між усіма змінними в обробленому наборі даних, що є ключовим для розуміння мультиколінеарності та значущості ознак.

Модуль «Аналіз моделі» дозволяє користувачам оцінювати продуктивність розгорнутих прогностичних моделей. Він відображає результати для чотирьох ключових алгоритмів: лінійна регресія, дерево рішень, випадковий ліс та XGBoost.

Фінальний модуль забезпечує функціональність прогнозування. Він дозволяє користувачам вводити специфічні параметри товару та точки продажу та отримувати прогностичні значення продажів на основі обраної найкращої моделі. Рисунки 3.16 – 3.18 демонструють інтерфейс для введення даних та відображення отриманих результатів прогнозування.

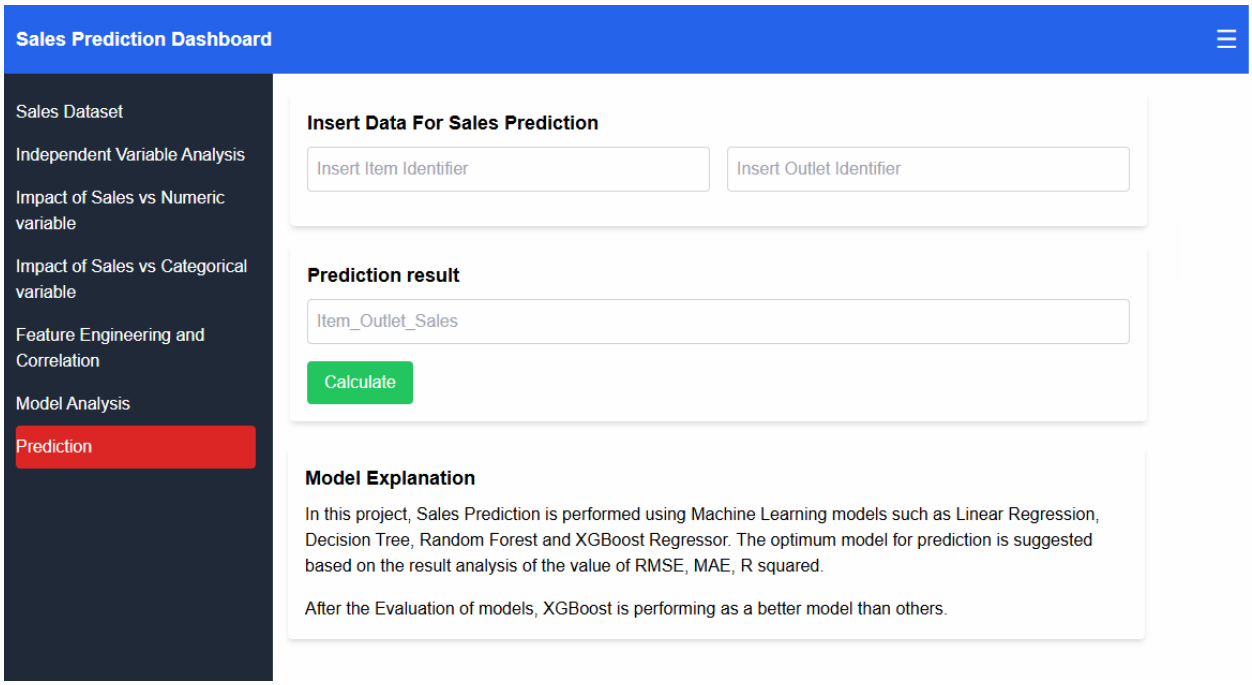


Рис. 3.16. Вигляд дашборду для прогнозування

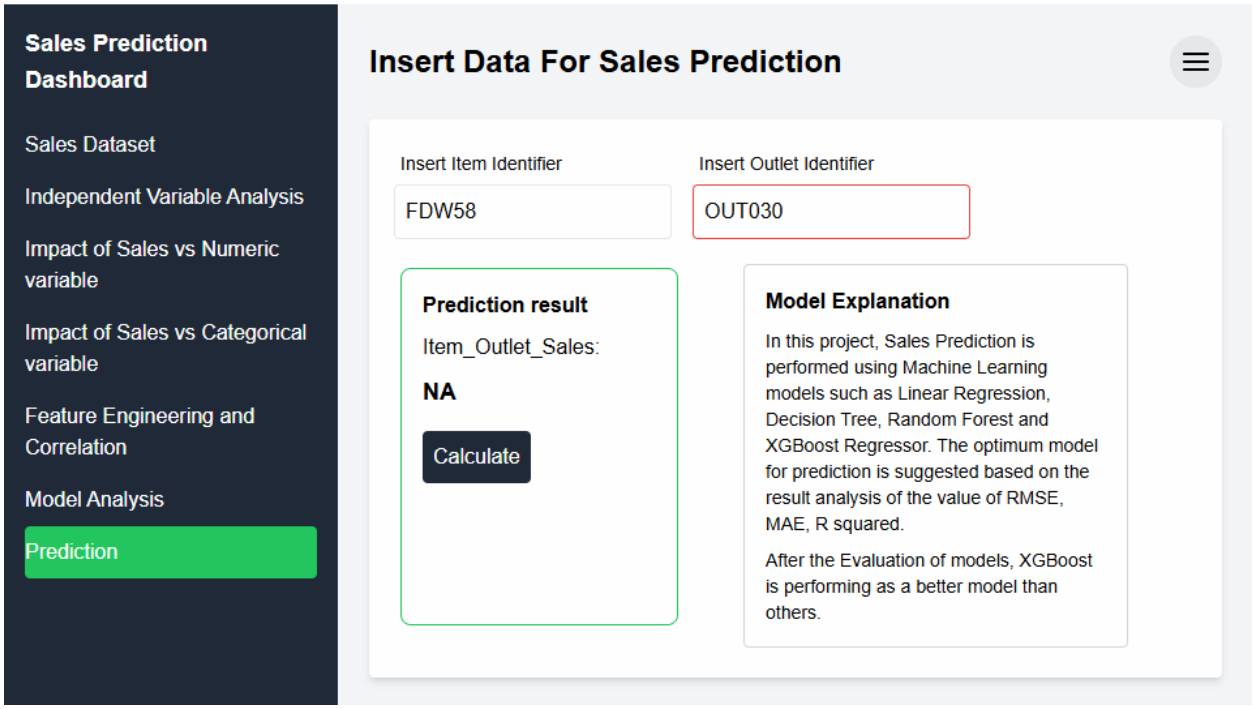


Рис. 3.17. Вигляд дашборду для результатів прогнозування (при неправильному вводі)

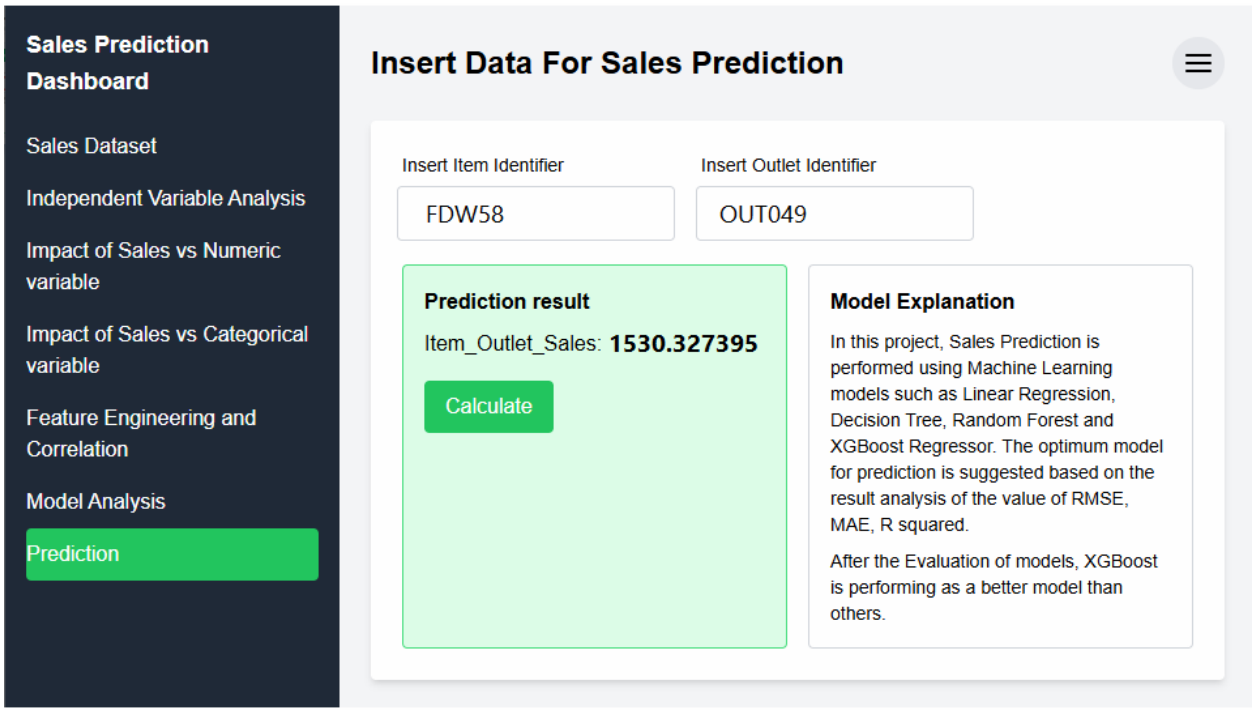


Рис. 3.18. Вигляд дашборду для результатів прогнозування

Отже, у сучасному ринковому середовищі кожен торговий центр прагне заздалегідь знати попит споживачів для запобігання дефіциту або надлишку товарів у будь-який сезон. Здатність компаній та торгових центрів

передбачати попит на товари щоденно зростає. На рівні організацій проводяться інтенсивні дослідження для забезпечення прецизійного прогнозування продажів, оскільки точні прогнози напряму корелюють із фінансовими показниками та прибутком компанії.

Дане дослідження зосереджене на застосуванні алгоритмів машинного навчання до вирішення класичної проблеми ланцюга постачання, а саме — прогнозування продажів. Для досягнення найвищого рівня точності прогнозування продажів у рамках цього проєкту було оцінено низку моделей та алгоритмів.

Висновки до розділу

У третьому розділі реалізовано практичну частину дослідження — створено програмний прототип системи прогнозування продажів із використанням технологій машинного навчання. Проведено підготовку та очищення набору даних, здійснено візуалізацію змінних і кореляційний аналіз для виявлення ключових факторів впливу. Навчання моделей показало, що алгоритм XGBoost забезпечує найвищу точність прогнозування серед розглянутих підходів. Розроблено інтерактивну інформаційну панель, що відображає результати прогнозів і підтримує прийняття управлінських рішень у реальному часі. Практична реалізація підтвердила ефективність запропонованої методики та її придатність до застосування у сфері бізнес-аналітики.

ВИСНОВКИ

У магістерській роботі здійснено комплексне дослідження моделей, методів та технологій машинного навчання, що використовуються для побудови ефективних систем прогнозування продажів у сучасному бізнес-середовищі. На основі проведеного теоретичного аналізу, моделювання та програмної реалізації запропоновано підхід, який підвищує точність прогнозів та оптимізує процес прийняття управлінських рішень у сфері роздрібної торгівлі та електронної комерції.

У першому розділі було розглянуто предметну область прогнозування продажів і проаналізовано сучасний стан використання технологій машинного навчання у цій галузі. Визначено, що традиційні статистичні методи (ARIMA, ETS, прості регресійні моделі) демонструють обмежену ефективність у контексті великих обсягів даних та високої варіативності ринкових факторів. Натомість алгоритми машинного навчання, зокрема дерева рішень, ансамблеві методи (Random Forest, XGBoost, LightGBM) та нейронні мережі, забезпечують адаптивність, здатність враховувати нелінійні залежності та підвищену точність прогнозів. Проаналізовано наукові джерела, в яких підтверджено ефективність таких моделей у реальних бізнес-застосуваннях, і сформовано теоретичну базу для подальшого розроблення системи прогнозування продажів.

У другому розділі досліджено моделі, алгоритми та архітектуру системи прогнозування продажів на основі технологій машинного навчання. Проведено порівняльний аналіз алгоритмів лінійної регресії, дерева рішень, випадкового лісу та градієнтного бустингу. Встановлено, що ансамблеві методи демонструють вищу стабільність результатів та меншу схильність до перенавчання порівняно з окремими моделями. Запропоновано методологічний підхід до створення моделей прогнозування, який включає етапи розуміння бізнес-проблеми, аналізу та підготовки даних, інженерії

ознак, навчання моделі, оцінювання ефективності та інтеграції результатів у бізнес-середовище.

Також у межах другого розділу розроблено концептуальну архітектуру системи прогнозування, яка реалізує поетапний робочий процес відповідно до стандарту CRISP-DM. Описано ключові етапи: від збору та очищення даних до розгортання готової моделі. Для оцінювання якості моделі використано стандартні метрики — MAE, RMSE, R^2, що дозволило забезпечити об’єктивну перевірку точності прогнозів.

У третьому розділі реалізовано практичну частину роботи — створення програмного прототипу системи прогнозування продажів із використанням алгоритмів машинного навчання. Проведено детальний аналіз набору даних, який включав як числові, так і категоріальні змінні, здійснено візуалізацію цільової змінної та виконано комплексну попередню обробку даних. Реалізовано процес очищення, кодування та нормалізації ознак, проведено двовимірний кореляційний аналіз для виявлення ключових взаємозв’язків між факторами.

У результаті моделювання побудовано кілька варіантів прогностичних моделей, серед яких найкращі результати продемонструвала модель XGBoost, що забезпечила мінімальну похибку та найвищий показник точності за всіма метриками. На основі одержаних результатів створено інтерактивну інформаційну панель для моніторингу та візуалізації прогнозів продажів, яка дозволяє користувачеві в реальному часі аналізувати ключові тенденції та приймати обґрунтовані управлінські рішення.

Підсумовуючи результати дослідження, можна сформулювати такі основні наукові та практичні висновки:

- Проведений аналіз підтвердив доцільність застосування алгоритмів машинного навчання для прогнозування продажів, особливо в умовах високої динамічності ринку та багатофакторності вхідних даних.

- Запропонована методика побудови моделей базується на послідовності етапів, що охоплюють збір, очищення, інженерію ознак, навчання та оцінювання моделі, що забезпечує підвищену точність і стійкість прогнозів.

Реалізована архітектура системи дозволяє масштабувати рішення, адаптувати його до нових наборів даних та інтегрувати у корпоративні інформаційні системи.

Створена інформаційна панель сприяє підвищенню прозорості аналітичного процесу, наочності представлення результатів і підтримує процес прийняття стратегічних рішень у сфері управління продажами.

ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Vaishnavi Nath Dornadula and Gheeta S (2019) Credit Card Fraud Detection using Machine Learning Algorithms. International Conference On Recent Trends In Advanced Computing. Pp 631-641. doi-10.1016/j.procs.2020.01.057
2. Xinjie Li; Jiakai Du; Yang Wang; Yuanm Cao; Automatic Sales forecasting using LSTM Networks, IEEE International Workshop, Shangai,China,20-22 November 2020.21
3. Y. Guo and X. Li, “Application of an improved algorithm of a mobile e-commerce system,” Industrial Management and Data Systems, vol. 118, no. 2, pp. 297– 303, 2017.
4. Narkhede, S., & Shinde, N. (2020). Evaluation Metrics for Regression and Classification Problems in Machine Learning. International Journal of Computer Applications, 178(1).
5. Shinde, N., et al. (2017). Enhancing Predictive Modeling using Random Forest and LDA with R package for Real-Life Business Case. International Conference on Data Mining and Applications.
6. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. SPSS.
7. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann.
8. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
9. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control (5th ed.). Wiley.
10. Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.

11. Varian, H. R. (2014). Predictive Big Data Analytics: The Power of R in Forecasting. Journal of Economic Perspectives, 28(2), 3–20.
12. Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81–106.
13. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
14. Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow (2nd ed.). O'Reilly Media.
15. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
16. Tufte, E. R. (2001). The Visual Display of Quantitative Information (2nd ed.). Graphics Press.
17. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (2nd ed.). Springer.
18. Donoho, D. (2017). 50 Years of Data Science. Journal of Computational and Graphical Statistics, 26(4), 745–766.
19. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate Data Analysis (7th ed.). Prentice Hall.
20. Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence, 14, 1137–1143.
21. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189–1232.
22. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
23. Chang, W., & Cheng, J. (2021). Shiny: Web Application Development in R (R Package Version 1.7.1).
24. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

25. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.
26. McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and Stan (2nd ed.). CRC Press.
27. Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. Proceedings of the 20th International Conference on Very Large Data Bases, 487–499.
28. Efron, B., & Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman and Hall.
29. Heaton, J. (2017). Introduction to Deep Learning (2nd ed.). Heaton Research.
30. Gorunescu, F. (2011). Data Mining Applications Using R. CRC Press.
31. Waller, L. A., & Gotway, C. A. (2004). Applied Spatial Statistics for Public Health Data. Wiley.
32. Griffin, P. M. (2004). Sales Forecasting Management: A Global Perspective. Prentice Hall.
33. Armstrong, J. S. (2001). Principles of Forecasting: A Handbook for Researchers and Practitioners. Springer.
34. Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). Forecasting: Methods and Applications (3rd ed.). Wiley.
35. Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence, 1(5), 206–215.
36. Fei-Fei, L., Deng, J., & Li, K. (2010). ImageNet: A Large-Scale Hierarchical Image Database. IEEE Conference on Computer Vision and Pattern Recognition, 248–255.