

ПРОГНОЗУВАННЯ ПРОДАЖІВ ПРОДУКТОВОЇ КОМПАНІЇ З ВИКОРИСТАННЯМ GRADIENT BOOSTING

Вінницький національний технічний університет

Анотація

Дану роботу присвячено прогнозуванню кількості продажів з використанням алгоритму Gradient Boosting. В рамках дослідження проведено передобробку даних, виконано розвідувальний аналіз даних, розроблено модель для прогнозування кількості продажів та проведено її тестування. Модель було перевірено та підтверджено її ефективність на основі раніше невідомих даних.

Ключові слова: Gradient Boosting, машинне навчання, передобробка даних, нормалізація, аналіз, прогнозування кількості продажів.

Abstract

This work is devoted to forecasting the number of sales using the Gradient Boosting algorithm. As part of the study, data reprocessing was carried out, exploratory data analysis was carried out, a model for predicting the number of sales was developed and tested. The model was tested and its effectiveness was confirmed based on previously unknown data.

Keywords: Gradient Boosting, machine learning, data reprocessing, normalization, analysis, sales forecasting.

Вступ

Ансамблеві методи машинного навчання, такі як Gradient Boosting, використовуються в задачах прогнозування складних систем, оскільки вони мають високу продуктивність у розпізнаванні складних залежностей в даних, що є важливим для створення моделі прогнозування кількості продажів. Алгоритм Gradient Boosting дозволяє аналізувати тенденції та оцінювати вплив факторів на продажі, забезпечуючи вищу точність прогнозу порівняно зі простішими моделями машинного навчання.

Метою даного дослідження є створення моделі для прогнозування кількості продажів на основі історичних, та Gradient Boosting є чудовим інструментом для розв'язання даної проблеми.

Формування датасету

У цьому дослідженні використано відкритий набір даних Company Financials Dataset, автором якого є Atharva Arya, набір є загальнодоступний на платформі Kaggle[1]. Даний набір даних має велику кількість необроблених даних, які не відповідають стандартним форматам, через що значна кількість даних є недоступними для моделі.

Для отримання корисної інформації було проведено попередню обробку даних, що включає — видалення зайвих текстових позначень у даних про фінанси (позначку валюти, коми що розділяють розряди ціни), та нормалізовано числові дані для кращої продуктивності моделі. Завдяки всім цих перетворень, було отримано оптимальний для аналізу набір даних, що наведено нижче (рис.1).

	segment	country	product	discount_band	units_sold	manufacturing_price	sale_price	gross_sales	discounts	sales	g
0	2	0	1	3	1618	3	20	32370	0	32370	1
1	2	2	1	3	1321	3	20	26420	0	26420	1
2	3	1	1	3	2178	3	15	32670	0	32670	2
3	3	2	1	3	888	3	15	13320	0	13320	8
4	3	3	1	3	2470	3	15	37050	0	37050	2

ис. 1. Приклад отриманих показників компанії

Цільовою змінною моделі є кількість продажів, здійснених компанією. Щоб спрогнозувати кількість продажів, було використано ряд показників таких як — розмір знижки, ціна продукту, період року, ціна виробництва, тип продукту та країна в якій здійснюються продажі. Для аналізу зв'язків між всіма цими

показниками було використано ансамблеву модель gradient boosting, що дозволяє створити модель, що здатна створювати більш точні прогнозування базуючись на великій кількості параметрів [2, 5].

Під час розвідувального аналізу даних було сформовано матрицю кореляцій між ознаками, наявними у вибірці. Загальна картина свідчить про переважно низький рівень взаємної кореляції, що вказує на відсутність виражених лінійних зв'язків між більшістю змінних.

Водночас було виявлено кластер ознак, які демонструють високу, а в деяких випадках і пряму кореляцію з цільовою змінною, що може свідчити про їхній вагомий вплив на результати прогнозування (рис. 2).

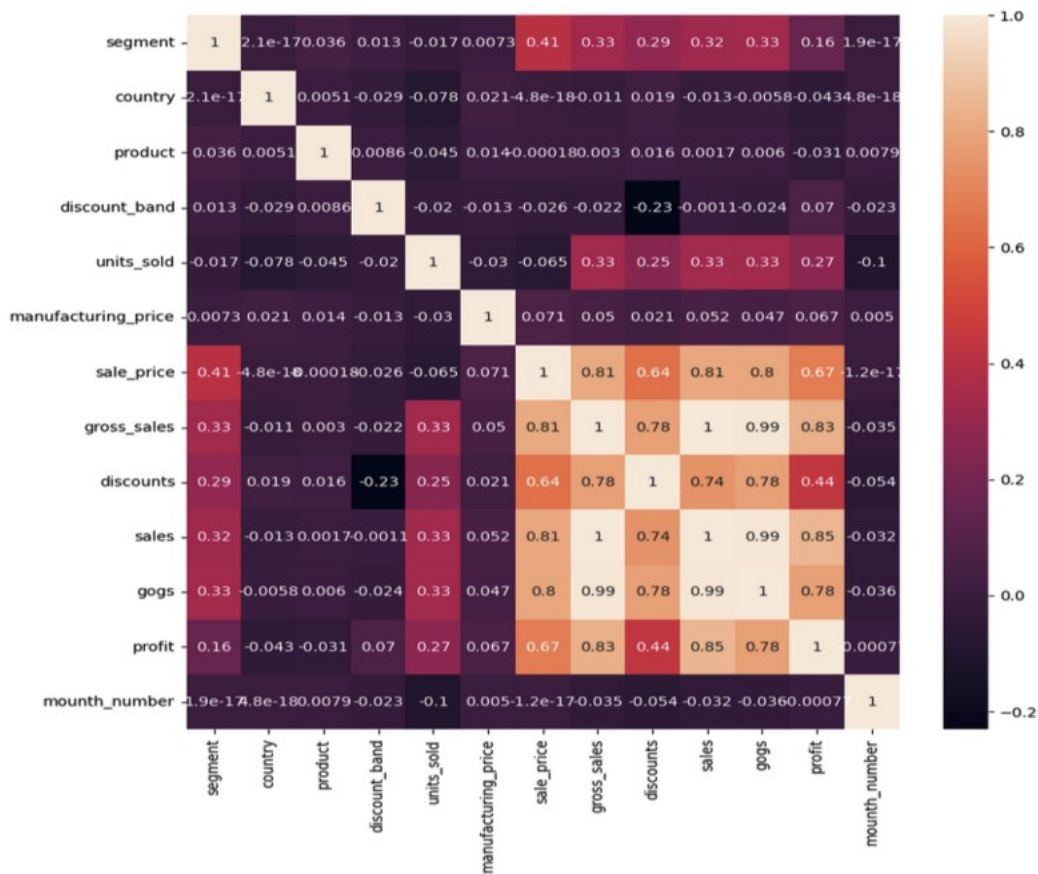


Рис. 2. Матриця кореляції

Під час підготовки ознак до навчання моделі було проаналізовано ступінь кореляції кожної змінної з цільовою ознакою sales. В результаті аналізу виявлено ряд ознак, що мають надзвичайно високий рівень кореляції ($r > 0.85$) з цільовим значенням. Такі змінні не слід використовувати як вхідні до моделі, оскільки вони або напряду обчислюються на основі цільової змінної, або є похідними від неї, що може призвести до витoku інформації і формально завищити якість прогнозування. Опираючись на це було видалено наступні ознаки: gross_sales, gogs та profit. Після видалення надлишкових ознак було створено оновлену кореляційну матрицю (рис. 3)

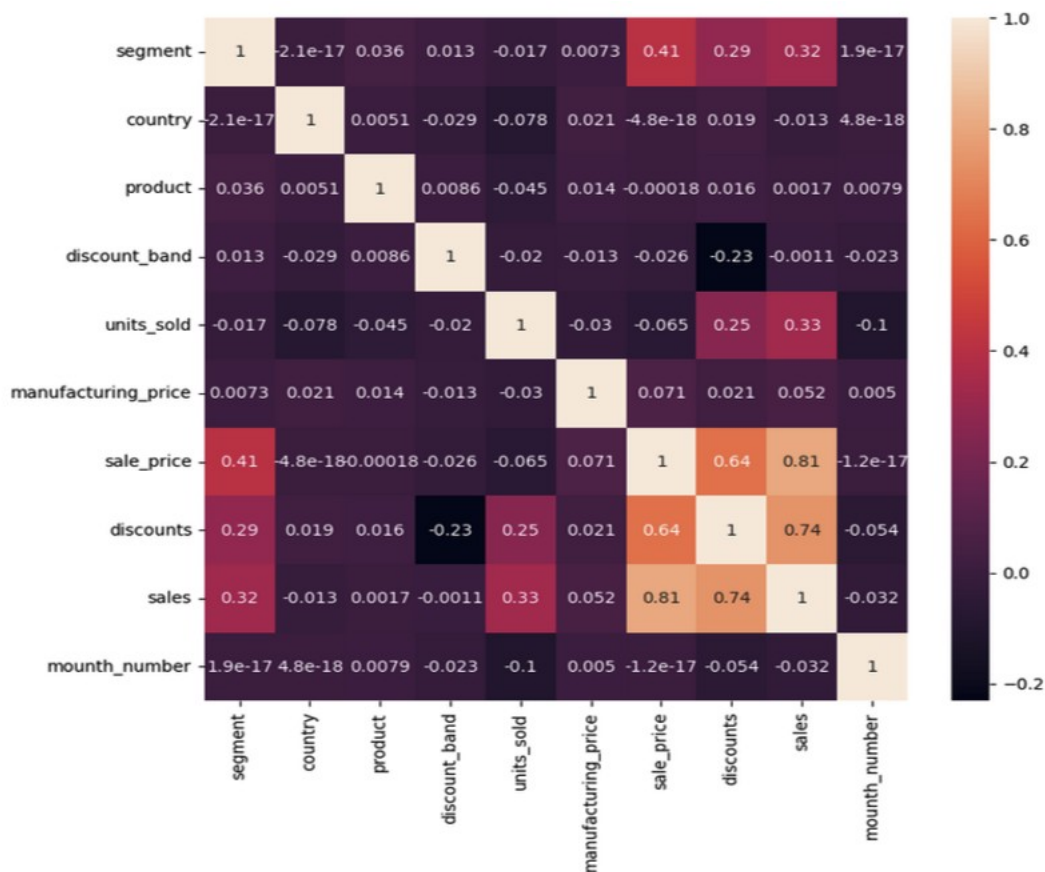


Рис. 3. Очищена матриця кореляції

Аналіз матриці кореляції показав помітний зв'язок між продажами та знижками ($r = 0.74$), що свідчить про вплив знижкової політики на реалізацію товарів. Також простежується певна залежність між продажами та ціною категорією товару ($r = 0.64$), що може вказувати на вплив виробничих витрат на кінцевий обсяг реалізованої продукції.

Цікавим спостереженням є те, що такі ознаки, як країна походження товару та його конкретний тип, мають незначний вплив на кількість проданих одиниць, про що свідчить їх низький рівень кореляції з цільовою змінною. Натомість набагато вагомішим чинником виявився сегмент, у якому реалізується товар, що може свідчити про важливість правильної маркетингової стратегії та позиціонування продукту для підвищення обсягів продажів.

Результати дослідження

У цьому дослідженні модель Gradient Boosting застосована до фінансових показників компанії. Дана форма ансамблевого машинного навчання, яка вивчає модель, вирощуючи кілька слабких елементів (часто дерева рішень), а потім додаючи їх по одному, щоб зменшити помилку. Його головна перевага полягає в тому, що він може ефективно працювати зі складними взаємозв'язками в даних і водночас стійкий до шуму та мінливості. Gradient Boosting також має вбудовану регуляризацию, яка запобігає перенавчанню, що робить модель стабільною та дуже точною в прогнозах [2, 4].

Модель градієнтного бустингу було натреновано на даних про продажі компанії, щоб прогнозувати майбутні тенденції та покращувати стратегічне планування. Для того, щоб зробити процес навчання більш точним і перевірити якість моделі, весь набір даних було розділено на дві частини: тренувальну, яка використовувалася для навчання алгоритму, і тестову, яка допомагала оцінити її здатність до узагальнення на нових, раніше невідомих даних.

Для перевірки роботоздатності моделі було використано кросс-валідацію. Тестування показало оцінку r^2 рівну 0.8753, що є високим показником, що свідчить про гарне навчання моделі, без

перенавчання. Вона здатна якісно спрогнозувати раніше невідомі дані. Також було визначено параметри які мають найбільший вплив на прогнозування. До них належать:

- роздрібна ціна;
- місяць року;
- розмір знижки;
- сегмент в якому відбувається продаж.

Значення розмірів впливу параметрів на результат прогнозування продемонстровано на рис. 4.

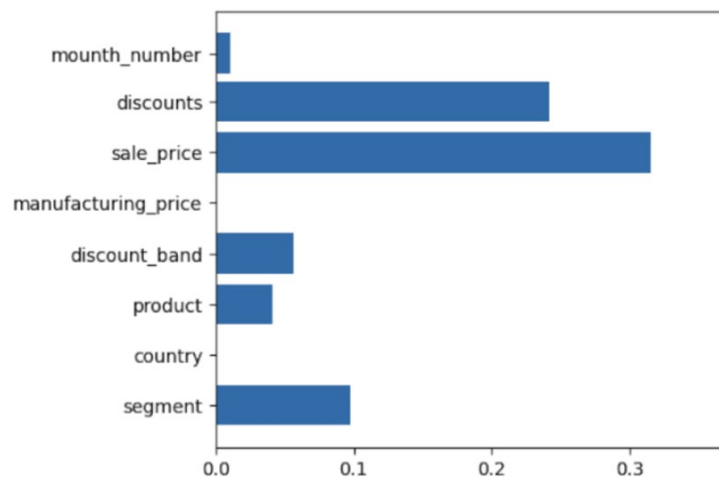


Рис. 4. Ваги властивостей у прогнозі моделі

Висновки

Результати дослідження показують значні можливості моделі градієнтного бустингу для прогнозування та планування продажів. Модель продемонструвала здатність обробляти складні зв'язки даних, для точного визначення майбутніх тенденцій на основі комерційних показників. Через високу стійкість моделі до шуму та мінливості даних, вона продемонструвала значну продуктивність для прогнозування кількості продажів, яке має велике значення для стратегічного планування компанії.

Найважливішими факторами, що впливають на точність прогнозів, є роздрібна ціна, місяць року, розмір знижки та сегмент продажів. Це може бути використано для правильного розподіляти ресурси для підвищення виторгів.

Результати підтверджують ефективність моделі Gradient Boosting для стратегічного планування в бізнесі, оскільки вона не лише забезпечує точні прогнози, але й визначає важливі фактори, які можуть вплинути на результат. Модель продемонструвала значну точність у передбаченні кількості проданих товарів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Company Financials Dataset Kaggle. [Електронний ресурс]. – Режим доступу: <https://www.kaggle.com/datasets/atharvaarya25/financials>
2. Zhang Z., Zhao Y., Canes A., Steinberg D., Lyashevskaya O. Predictive analytics with gradient boosting in clinical medicine. *Annals of Translational Medicine*. 2019. Vol. 7, no. 7, [Електронний ресурс] – Режим доступу: <https://atm.amegroups.org/article/view/24543/23475>
3. Wang J., Chong W. K., Lin J., Hedenstierna C. P. T. Retail Demand Forecasting Using Spatial-Temporal Gradient Boosting Methods. *Journal of Computer Information Systems*. 2023. Vol. 64, no. 5, pp. 652–664. [Електронний ресурс] – Режим доступу: <https://doi.org/10.1080/08874417.2023.2240753>
4. Malik S., Khan M., Abid M. K., Aslam N. Sales Forecasting Using Machine Learning Algorithm in the Retail Sector. *Journal of Computing & Biomedical Informatics*. 2024. Vol. 6, no. 2, pp. 282–294. [Електронний ресурс] – Режим доступу: <https://www.jcibi.org/index.php/Main/article/view/370>
5. Neba C., Gerard S. F. B., Nsuh G., Amouda P., Neba A., et al. Advancing Retail Predictions: Integrating Diverse Machine Learning Models for Accurate Walmart Sales Forecasting. *Asian Journal of Probability and Statistics*. 2024. Vol. 26, no. 7, pp. 1–23. [Електронний ресурс] – Режим доступу: <https://doi.org/10.9734/ajpas/2024/v26i7626>

Дудар Анатолій Михайлович – студент групи СА-22Б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: gigsollt@gmail.com

Жуков Сергій Олександрович – к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@vntu.edu.ua

Dudar Anatolii M. - student of Faculty of Intelligent Information Technologies and Automation, SA-22B, Vinnytsia National Technical University, Vinnytsia, e-mail gigsoll@gmail.com

Zhukov Serhii O. - Ph.D., Assistant Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@vntu.edu.ua