

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ ЦІНИ ТЕЛЕФОНІВ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Робота присвячена підготовці та розвідувальному аналізу даних для подальшого використання для інформаційної технології передбачення ціни телефонів методами машинного навчання. Було здійснено детальний аналіз датасету та його ознак з метою виявлення ключових факторів, що впливають на формування вартості мобільних пристроїв.

Ключові слова: ціна телефонів, інформаційні технології, машинне навчання, аналіз даних, передбачення, ознаки, мобільні пристрої.

Abstract

The work is dedicated to the preparation and exploratory data analysis for further use in an information technology system for predicting phone prices using machine learning methods. A detailed analysis of the dataset and its features was conducted to identify key factors that influence the pricing of mobile devices.

Keywords: Phone Price, Information Technologies, Machine Learning, Data Analysis, Prediction, Features, Mobile Devices.

Вступ

Сучасні технології розвиваються надзвичайно швидкими темпами — те, що ще вчора здавалося фантастикою, вже завтра стає реальністю. Особливо це помітно у сфері інформаційних технологій, які є основою більшості інновацій сьогодення. Завдяки їм відкриваються нові можливості для розв'язання широкого спектра задач, які раніше здавалися складними або неможливими.

Однією з таких задач є аналіз ринку мобільних пристроїв та прогнозування вартості техніки. Ринок смартфонів постійно змінюється: з'являються нові моделі, технології, функції, а разом із ними змінюється й ціна. Передбачення вартості телефонів дозволяє глибше зрозуміти ринкові тенденції, визначити чинники, що впливають на формування ціни, та приймати більш обґрунтовані рішення — як для споживачів, так і для виробників чи продавців.

Отже, застосування інформаційних технологій для обробки та аналізу даних відкриває нові перспективи у моделюванні, прогнозуванні та дослідженні ринку мобільних пристроїв, роблячи процеси оцінки вартості більш точними та ефективними.

Розвідувальний аналіз

Для проведення розвідувального аналізу було обрано набір даних «Mobile Price Classification», опублікований користувачем Abhishek Sharma та доступний для загального використання на платформі Kaggle [1]. Даний датасет містить структуровану інформацію про технічні характеристики мобільних телефонів, що може бути використана для аналізу ринку та передбачення ціни пристроїв.

Набір даних включає широкий спектр ознак, які характеризують смартфони з різних технічних аспектів. Серед найважливіших стовпців можна виділити параметри, що безпосередньо впливають на формування вартості пристрою, зокрема: ємність акумулятора, обсяг оперативної пам'яті, роздільна здатність екрана, вага телефону, внутрішня пам'ять, актова частота процесора, кількість ядер процесора, характеристики фронтальної та основної камер, наявність підтримки мереж четвертого та третього покоління, підтримка двох SIM-карт, наявність сенсорного екрана, підтримка відповідних бездротових модулів.

Датасет також містить цільову змінну `price_range`, яка класифікує телефони за рівнем вартості (від низького до високого). Це створює умови для побудови моделей машинного навчання, здатних передбачати цінову категорію смартфона на основі його технічних характеристик.

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	...	px_height	px_width
0	842	0	2.2	0	1	0	7	0.6	188	2	...	20	756
1	1021	1	0.5	1	0	1	53	0.7	136	3	...	905	1988
2	563	1	0.5	1	2	1	41	0.9	145	5	...	1263	1716
3	615	1	2.5	0	0	0	10	0.8	131	6	...	1216	1786
4	1821	1	1.2	0	13	1	44	0.6	141	2	...	1208	1212

Рис. 1. Приклад ознак мобільних телефонів, що містить набір даних

Даний датасет включає у себе такі колонки:

- battery_power: Ємність акумулятора телефону (mAh). *(Numeric)*
- blue: Наявність Bluetooth у пристрої (1 – так, 0 – ні). *(Categorical)*
- clock_speed: Тактова частота процесора (GHz). *(Numeric)*
- dual_sim: Підтримка двох SIM-карт (1 – так, 0 – ні). *(Categorical)*
- fc: Роздільна здатність фронтальної камери (в мегапікселях). *(Numeric)*
- four_g: Підтримка мережі 4G (1 – так, 0 – ні). *(Categorical)*
- int_memory: Обсяг внутрішньої пам'яті (GB). *(Numeric)*
- m_dep: Товщина корпусу телефону (см). *(Numeric)*
- mobile_wt: Вага смартфона (грами). *(Numeric)*
- n_cores: Кількість ядер процесора. *(Numeric)*
- px: Роздільна здатність основної камери (мегапікселі). *(Numeric)*
- px_height: Висота екрану в пікселях. *(Numeric)*
- px_width: Ширина екрану в пікселях. *(Numeric)*
- ram: Обсяг оперативної пам'яті (MB). *(Numeric)*
- sc_h: Висота екрану у см. *(Numeric)*
- sc_w: Ширина екрану у см. *(Numeric)*
- talk_time: Максимальний час розмови (годин). *(Numeric)*
- three_g: Підтримка мережі 3G (1 – так, 0 – ні). *(Categorical)*
- touch_screen: Наявність сенсорного екрана (1 – так, 0 – ні). *(Categorical)*
- wifi: Підтримка Wi-Fi (1 – так, 0 – ні). *(Categorical)*
- price_range: Цільова змінна, що показує рівень вартості телефону (0 – низька, 1 – середня, 2 – висока, 3 – преміум). *(Categorical)*

Проведено попереднє очищення даних та побудовано інтерактивні кругові діаграми, а також гістограми для того щоб дослідити розподіл категоріальних та числових ознак.

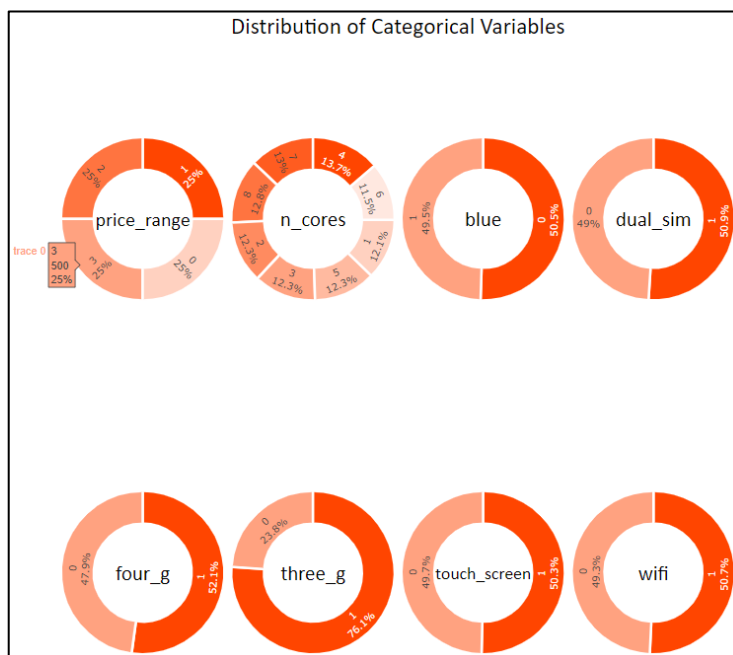


Рис. 2. Інтерактивні кругові діаграми категоріальних ознак

З наведених інтерактивних кругових діаграм категоріальних ознак на рисунку 2, можна зробити висновок, що мобільні телефони розподілені з однаковою частотою між чотирма класами price_range. Це свідчить про те, що набір даних є повністю збалансованим. Також мобільні телефони в наборі даних мають майже однакову частоту з точки зору наявності або відсутності Bluetooth, 4G, сенсорного екрану, Wifi, а також підтримки двох SIM-карт і кількості використовуваних процесорних ядер. Близько 76% мобільних телефонів підтримують 3G.

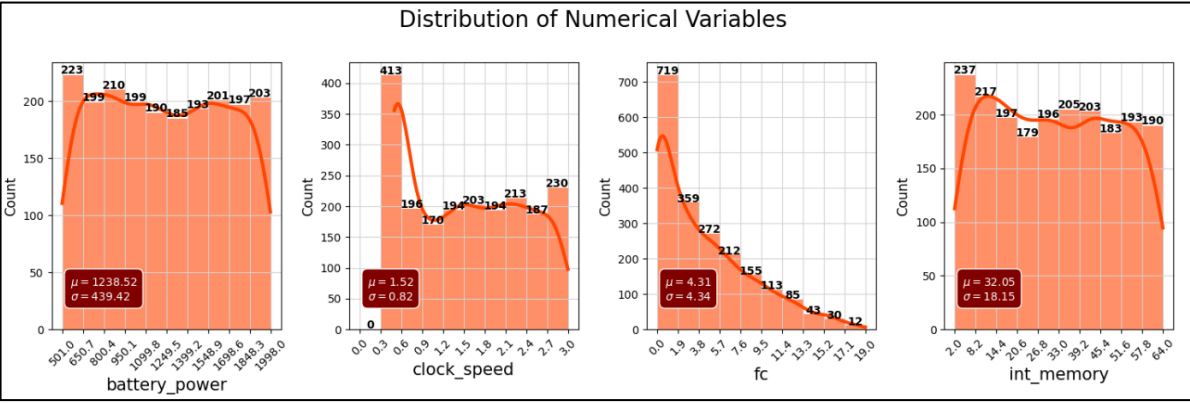


Рис. 3. Гістограми перших чотирьох числових ознак

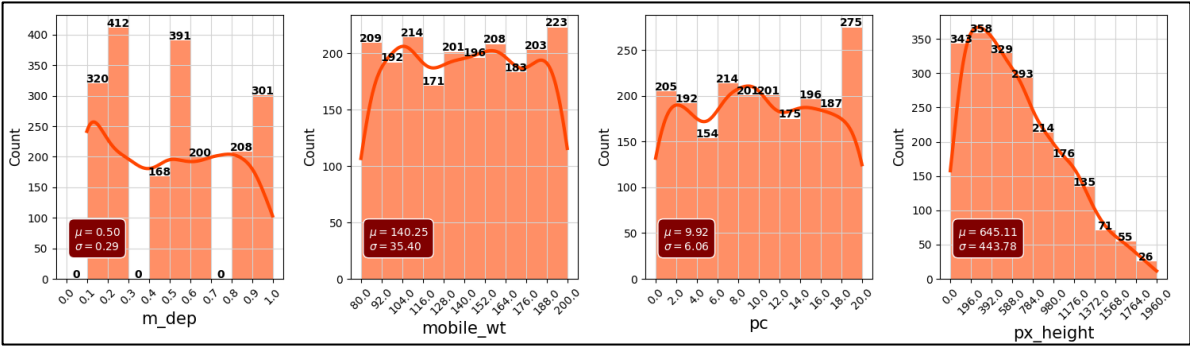


Рис. 4. Гістограми наступних чотирьох числових ознак

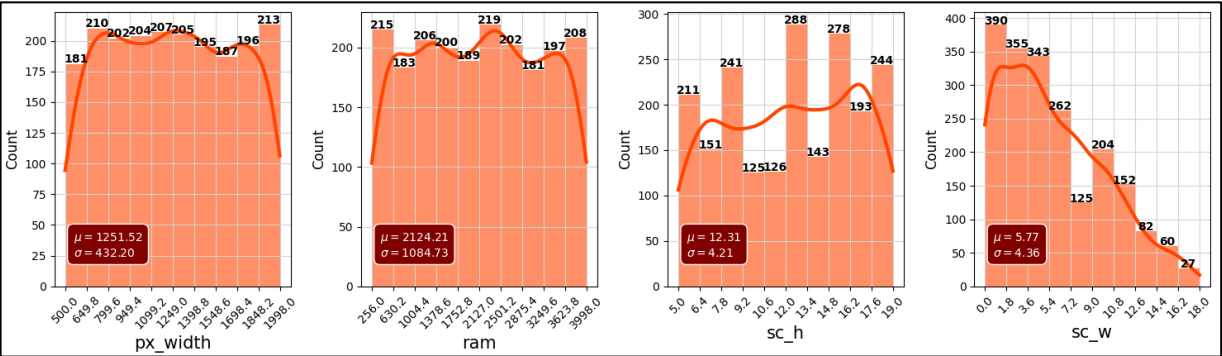


Рис. 5. Гістограми останніх чотирьох числових ознак

Дивлячись на наведені графіки на рисунках 3-5, ми можемо легко зрозуміти статистичні значення кожної ознаки, включаючи мінімальне і максимальне значення, а також середнє і стандартне відхилення. У деяких характеристиках, зокрема px_height (висота роздільної здатності пікселів) та sc_w (ширина екрану мобільного в см), спостерігається багато значень, близьких до нуля, які здаються шумом.

Підготувавши дані створимо для них матрицю кореляції, що дасть змогу виявити явну залежність та зв'язки між різними параметрами та ознаками.

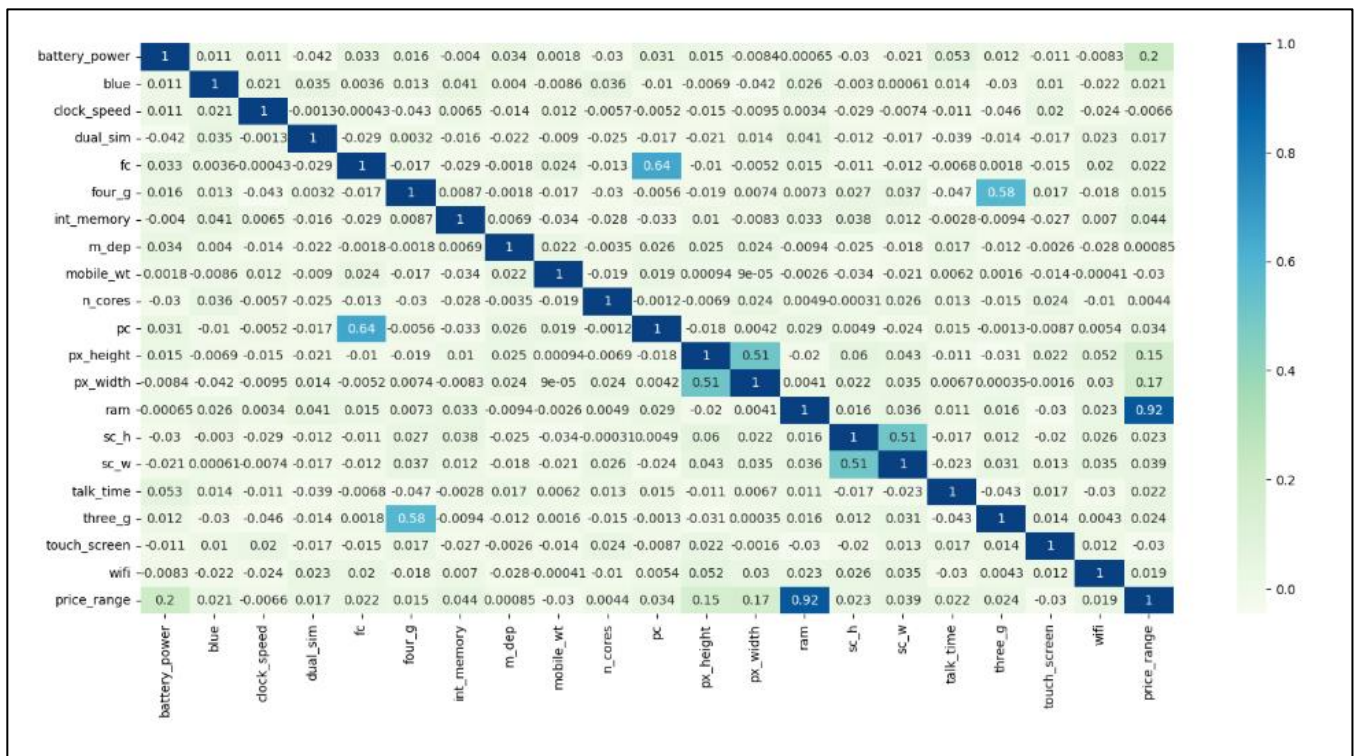


Рис. 6. Матриця кореляції

Як бачимо рисунок 2 показує, що існує багато сильних кореляцій між змінними. «price_range» сильно корелює із «ram». Така висока кореляція свідчить про те, що телефони з більшою кількістю оперативної пам'яті можуть бути доступні за більш високими цінами.

Це пов'язано з тим, що оперативна пам'ять є важливим компонентом телефону, який впливає на його продуктивність. Більше оперативної пам'яті дозволяє телефону запускати більше програм одночасно, обробляти більше даних та працювати швидше. Тому виробники телефонів, як правило, встановлюють більше оперативної пам'яті у свої флагманські моделі, які також мають й вищу ціну.

Крім того, сильні кореляції між деякими іншими змінними, такими як pc та fc, px_height та px_width, sc_w та sc_h, а також three_g та four_g, вказують на те, що ці пари характеристик зазвичай зростають або зменшуються разом. Інші змінні мають слабку кореляцію між собою, що свідчить про незалежність їх впливу на досліджувані показники.

Висновки

При розвідувальному аналізі набору даних «Mobile Price Classification», що містить у собі інформацію про параметри та характеристики цінової політики мобільних телефонів, було досліджено вплив різних ознак на показник ціни. Побудовано інтерактивні кругові діаграми категоріальних ознак, які дали змогу дослідити їх розподіл та виявити з якою частотою вони розподілені між класами цільової ознаки.

Далі було побудовано гістограми числових ознак за допомогою яких було визначено статичні значення кожної ознаки

Побудовано матрицю кореляції та з її допомогою виявлено ті фактори, які найбільше впливають на збільшення ціни мобільних телефонів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Mobile Price Classification Dataset. Kaggle. 2023 [Електронний ресурс]. – Режим доступу: <https://www.kaggle.com/datasets/iabhishekkofficial/mobile-price-classification>
2. Pandas Getting started. 2024 [Електронний ресурс] – Режим доступу: https://pandas.pydata.org/docs/getting_started/index.html
3. Matplotlib Pyplot Documentation. 2024 [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html
4. Seaborn Tutorial. 2023 [Електронний ресурс]. – Режим доступу: <https://seaborn.pydata.org/tutorial.html>

Нереуцький Олег Віталійович – студент групи 2ІСТ-24м, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: onereutsk@gmail.com

Козачко Олексій Миколайович – к.т.н., доцент кафедри системного аналізу та інформаційних технологій,

Вінницький національний технічний університет, e-mail: lekoz80@gmail.com

Nereutsky Oleg Vitaliyovych. - student of Faculty of Intelligent Information Technology and Automation, 2IST-20b, Vinnytsia National Technical University, Vinnytsia, e-mail onereutsk@gmail.com

Kozachko Oleksiy Mykolayovych - Ph.D., Assistant Professor of the Department of Systems Analysis and Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: lekoz80@gmail.com