НТУУ «КПІ ім. ІГОРЯ СІКОРСЬКОГО»
НН ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА ММСА

Комп'ютерний практикум

№2

Варіант 2

З дисципліни

«Інтелектуальний аналіз даних»

Виконали:

Студентка групи КА-31

Сачек В.

Студентки групи КА-32

Богданова О. і Ревенко І.

Перевірив:

Андросов Д. В.

Київ 2025

Тема: Побудова та оцінювання якості моделей класифікації та регресії засобами бібліотеки Scikit-Learn Python.

Завдання:

Побудувати моделі регресії:

• Лінійної регресії з різними значеннями гіперпараметру fit_intercept, використовуючи клас sklearn.linear_model.LinearRegression.

• Гребневої регресії з різними значеннями гіперпараметру alpha, використовуючи sklearn.linear_model.Ridge.

• Поліноміальної регресії, використовуючи pipeline, PolynomialFeatures в поєднанні з LinearRegression.

Початкові дані:

(а) sklearn.datasets.load_boston

(б) www.kaggle.com/rahulsah06/gooogle-stock-price

Виконання:

Завдання, виконані згідно з ходом виконання роботи, з текстом програми, результатами аналізу моделей, підбором гіперпараметрів, значеннями метрик якості моделей, оцінкою впливу розміру навчальної вибірки на якість моделі, розміщено у відповідних файлах:

репозиторій GitHub:

https://github.com/oleksandrabohdanova/IAD_Team2/tree/main/Lab2

- для датасету (а) Boston:

linear_boston.ipynb

ridge_boston.ipynb

polynomial_boston.ipynb

- для датасету (б) Google Stock Price:

linear_google.ipynb

ridge_google.ipynb

polynomial_google.ipynb

Висновки:

(а) Для датасету **Boston Housing** розглянуа задача регресії полягала у прогнозі медіанної вартості заселених будинків за набором ознак.

Лінійна регресія забезпечує базовий рівень точності:

- значення R^2 ≈ 0.67 показує, що модель пояснює ~67% варіації цільової змінної.
- помилки RMSE і MAE свідчать, що реальні значення в середньому відхиляються від прогнозу на ≈3-5 одиниць (в $1000s).

Гребнева регресія дала майже ідентичні результати, тобто регуляризація незначно вплинула на якість. Значення критеріїв якості дуже близькі до звичайної лінійної регресії, що підтверджує стабільність моделі.

Поліноміальна регресія дала доволі суттєве покращення метрик якості:

- R^2 зріс до 0.84
- Помилки RMSE, MAE, MAPE зменшилися, що вказує на кращу відповідність даним.

Це свідчить, що зв'язок між ознаками та ціною є нелінійним, і поліноміальні перетворення краще його описують.

Отже, з усіх моделей найкращою за якістю прогнозу є поліноміальна регресія, оскільки вона має найвищий коефіцієнт детермінації R^2 та найнижчі помилки RMSE і MAE, проте якщо потрібна простота, може бути використана і звичайна лінійна чи гребнева регресія.

(б) Для датасету **Google Stock Price** розглянута задача регресії полягає у прогнозі ціни відкриття акції Open наступного дня на основі ціни Open поточного дня.

Лінійна регресія показує дуже високий рівень точності:

- Пояснює ≈ 92,2% варіації цільової змінної (R^2 = 0.922).

- Середня похибка MAE ≈ 7.05 і корінь середньоквадратичної помилки RMSE ≈ 9.72 свідчать про стабільну роботу моделі.

Можемо зробити висновок, що зв’язок є переважно лінійним, що добре видно і з графіку початкових даних.

Гребнева регресія має значення R^2 = 0.9196 та RMSE = 9.86. Ця модель є стабільнішою, але дещо менш точною.

Поліноміальна регресія показує найкращі метрики серед трьох: найвищий R^2 = 0.9225 і найменші RMSE та MAE. Водночас різниця незначна, тобто у даному наборі даних немає яскраво вираженої нелінійності.

Отже, усі три моделі мають дуже близькі значення критеріїв якості. Для гребневої регресії найкращий параметр близький нулю, для поліноміальної рівний 1, тобто моделі близькі до лінійної регресії або співпадають з нею. Тому за найкращу модель варто обрати звичайну лінійну регресію.