

УДК 004.8

DOI: <https://doi.org/10.17721/3041-2323.2024.297-303>

Олег ПУРСЬКИЙ¹, д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0002-1230-0305
e-mail: o.pursky@knute.edu.ua

Тетяна ФІЛІМОНОВА¹, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0001-9467-0141
e-mail: t.filimonova@knute.edu.ua

Анна СЕЛІВАНОВА¹, ст. викл.
ORCID ID: 0000-0001-6559-1508
e-mail: a.selivanova@knute.edu.ua

Андрій НЕЧЕПУРЕНКО¹, канд. екон. наук, доц.
ORCID ID: 0009-0003-5828-0762
e-mail: a.nechepurenko@knute.edu.ua

Тетяна ДУБОВИК¹, д-р екон. наук, проф.
ORCID ID: 0000-0001-9223-4629
e-mail: tatiana_dubovik@i.ua

¹ Державний торговельно-економічний
університет, Київ, Україна

ПРОГНОЗУВАННЯ ВАРТОСТІ ТОВАРІВ ЗА ДОПОМОГОЮ РЕГРЕСІЙНОЇ МОДЕЛІ НА ОСНОВІ ФРЕЙМВОРКУ PYTORCH

Розроблено регресійну модель для прогнозування вартості товарів із використанням фреймворку PyTorch. Додатково побудовано дві класичні регресійні моделі машинного навчання – RandomForestRegressor і GradientBoostingRegressor. Реалізовано порівняння моделей. Визначено найкращу.

Ключові слова: регресійні моделі, PyTorch, гіперпараметри, машинне навчання, MAE, MSE.

Вступ

Метою цього дослідження є побудова та порівняння моделей для прогнозування вартості ноутбуків у контексті онлайн-прода-

© Пурський Олег, Філімонова Тетяна, Селіванова Анна,
Нечепуренко Андрій, Дубовик Тетяна, 2024

жив. Для цього застосовано нейронну мережу на базі фреймворку **PyTorch** (PyTorch, n. d.), як одну з методик для розв'язання задачі регресії. Крім того, розглянуто інші методи машинного навчання, зокрема **випадковий ліс**, Random Forest Regressor (RandomForestRegressor, n. d.) та **градієнтний бустинг** (Gradient Boosting Regressor, n. d.); Randomized Search CV, n. d.), для порівняння результатів за ключовими метриками.

Результати

Задачу прогнозування ціни ноутбуків сформульовано так: є множина вхідних характеристик ноутбука (незалежні змінні), необхідно передбачити цільову змінну – **ціну ноутбука**. Таким способом проблема зводиться до багатовимірної регресії, де кожна характеристика виступає як предиктор, а ціна ноутбука – як прогнозоване значення.

Для аналізу та побудови моделі регресії використано датасет ноутбуків, отриманий із платформи Kaggle (Laptop Prices, n. d.). Датасет містить основні технічні характеристики пристроїв.

Після підготовки наборів даних, для їхньої організації у процесі навчання та тестування використано клас DataLoader. Цей клас надає інструменти для ітерації по даних із можливістю їх розбивання на пакети і підвищує ефективність оброблення за рахунок багатопоточності.

Далі розроблено модель NNModel на базі PyTorch. Архітектуру моделі наведено на рис. 1.

Layer (type)	Output Shape	Param #
Linear-1	[-1, 256]	5,888
ReLU-2	[-1, 256]	0
Dropout-3	[-1, 256]	0
Linear-4	[-1, 128]	32,896
ReLU-5	[-1, 128]	0
Dropout-6	[-1, 128]	0
Linear-7	[-1, 64]	8,256
ReLU-8	[-1, 64]	0
Dropout-9	[-1, 64]	0
Linear-10	[-1, 1]	65

Total params: 47,105
 Trainable params: 47,105
 Non-trainable params: 0

Input size (MB): 0.00
 Forward/backward pass size (MB): 0.01
 Params size (MB): 0.18
 Estimated Total Size (MB): 0.19

Рис. 1. Архітектура моделі

Загальна кількість епох навчання становить 500. Для уникнення перенавчання застосовують механізм дострокової зупинки (early stopping) із параметром 20 епох. Для оцінювання результатів побудовано графіки кривих навчання (рис. 2 і 3).

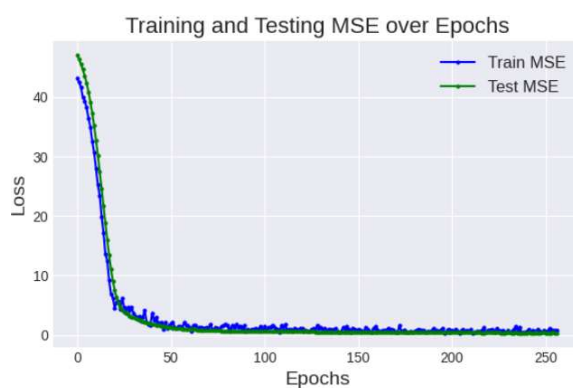


Рис. 2. Mean Squared Error (MSE)

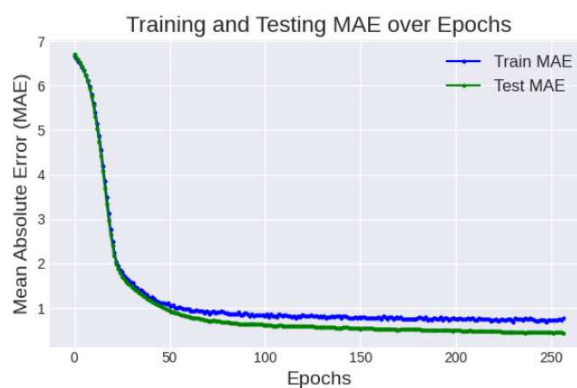


Рис. 3. Mean Absolute Error (MAE), n. d.

У процесі навчання моделі спостерігалось значне зниження тренувальних і тестових значень для MSE (Randomized Search CV, n. d.), середня абсолютна помилка (MAE) (R^2 Score, n. d.) зменшилася з 5.8011 до 0.7247 на тренувальних даних і з 5.7509 до 0.4555 на тестових. Загалом, модель демонструє високий рівень

адаптації до даних, що підтверджує її потенціал для практичного застосування.

Далі для порівняння результатів побудовано дві регресійні моделі `RandomForestRegressor`, `GradientBoostingRegressor`. Проведено оптимізацію параметрів двох алгоритмів машинного навчання – `RandomForestRegressor` (`RandomForestRegressor`, n. d.) і `GradientBoostingRegressor` (`GradientBoostingRegressor`, n. d.) – з використанням методу пошуку `RandomizedSearchCV` (`RandomizedSearchCV`, n. d.).

З табл. 2 видно, що `RandomForestRegressor` демонструє кращі результати на навчальних і тестових наборах даних порівняно з `GradientBoostingRegressor`. Класичні моделі машинного навчання показують значно кращі результати порівняно з нейронною мережею на даних метриках.

Реалізовано крос-валідацію для моделей `RandomForestRegressor` і `GradientBoostingRegressor` із використанням 5-кратної валідації для покращення узагальнення моделей.

Результати крос-валідації свідчать, що обидві моделі демонструють подібну ефективність за метриками MSE, яка для обох моделей становить 0.019. Однак `RandomForestRegressor` дещо перевершує `GradientBoostingRegressor` за показником MAE, що свідчить про її точніше передбачення середніх значень. Обидві моделі мають однакове значення $R^2 = 0.951$.

Побудуємо графіки для справжніх і прогнозованих значень для зазначених моделей (рис. 4).

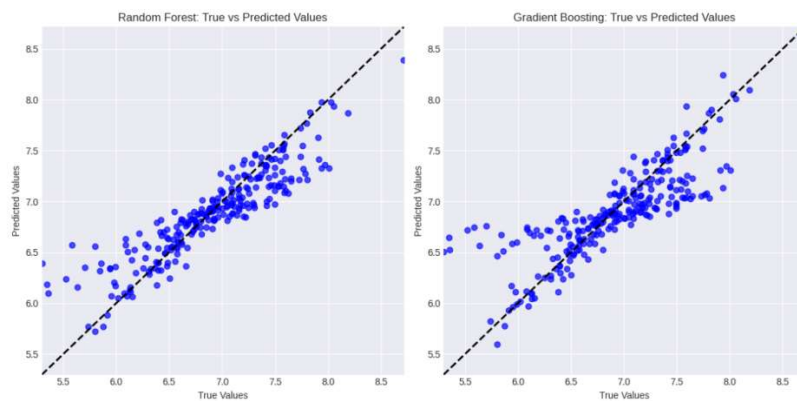


Рис. 4. Графіки справжніх і прогнозованих значень для моделей

Можна зробити висновок, що обидві моделі показали відносно високі результати, але найкраща модель RandomForestRegressor.

Дискусія і висновки

У цій роботі побудовано моделі для прогнозування вартості товарів у контексті онлайн-продажів. Для цього розроблено нейронну мережу на базі фреймворку PyTorch. Також побудовано регресійні моделі Random Forest і Gradient Boosting, для порівняння результатів за ключовими метриками.

Результати вказують на те, що моделі класичного машинного навчання є більш придатними для розв'язання цієї задачі регресії, демонструючи кращі узагальнювальні властивості та точність прогнозів. Нейронні мережі потребують ретельнішого налаштування й оптимізації для досягнення подібної ефективності.

Список використаних джерел

GradientBoostingRegressor. (б. д.). *Документація scikit-learn*. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html> (Дата звернення: 16.09.2024).

Laptop Prices. (б. д.). *Kaggle*. <https://www.kaggle.com/datasets/owm4096/laptop-prices> (Дата звернення: 16.09.2024).

Mean Absolute Error. (б. д.). *Документація scikit-learn*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html (Дата звернення: 16.09.2024).

Mean Squared Error. (б. д.). *Документація scikit-learn*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html (Дата звернення: 16.09.2024).

PyTorch. (б. д.). *Офіційний сайт*. <https://pytorch.org/> (Дата звернення: 16.09.2024).

R² Score. (б. д.). *Документація scikit-learn*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html (Дата звернення: 26.09.2024).

RandomForestRegressor. (б. д.). *Документація scikit-learn*. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html> (Дата звернення: 16.09.2024).

Randomized Search CV. (б. д.). *Документація scikit-learn*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html (Дата звернення: 17.09.2024).

References

GradientBoostingRegressor. (n. d.). *Scikit-learn documentation*. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html> (Accessed: 16.09.2024).

Laptop Prices. (n. d.). *Kaggle*. <https://www.kaggle.com/datasets/owm4096/laptop-prices> (Accessed: 16.09.2024).

Mean Absolute Error. (n. d.). *Scikit-learn documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html (Accessed: 16.09.2024).

Mean Squared Error. (n. d.). *Scikit-learn documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html (Accessed: 16.09.2024).

PyTorch. (n. d.). *Official website*. <https://pytorch.org/> (Accessed: 16.09.2024).

R² Score. (n. d.). *Scikit-learn documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html (Accessed: 16.09.2024).

RandomForestRegressor. (n. d.). *Scikit-learn documentation*. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html> (Accessed: 16.09.2024).

Randomized Search CV. (n. d.). *Scikit-learn documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html (Accessed: 17.09.2024).

Отримано редакцією журналу / Received: 23.09.24

Прорецензовано / Revised: 27.09.24

Схвалено до друку / Accepted: 01.10.24

Oleh PURSKYI¹, DSc (Phys. & Math.), Prof.
ORCID ID: 0000-0002-1230-0305
e-mail: o.pursky@knute.edu.ua

Tetiana FILIMONOVA¹, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0001-9467-0141
e-mail: t.filimonova@knute.edu.ua

Anna SELIVANOVA¹, Senior Lecturer
ORCID ID: 0000-0001-6559-1508
e-mail: a.selivanova@knute.edu.ua

Andrii NECHEPURENKO¹, PhD (Econ.), Assoc. Prof.
ORCID ID: 0009-0003-5828-0762
e-mail: a.nechepurenko@knute.edu.ua

Tetiana DUBOVYK¹, DSc (Econ.), Prof.
ORCID ID: 0000-0001-9223-4629
e-mail: tatiana_dubovik@i.ua

¹State University of Trade and Economics, Kyiv, Ukraine

FORECASTING THE COST OF GOODS USING A REGRESSION MODEL BASED ON THE PYTORCH FRAMEWORK

In this work, a regression model for predicting the cost of goods using the PyTorch framework has been developed. Additionally, two classical machine learning regression models – RandomForestRegressor and GradientBoostingRegressor – were constructed. A comparison of the models was carried out, and the best one was identified.

Keywords: *regression models, PyTorch, hyperparameters, machine learning, MAE, MSE.*

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.