

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ КЛАСИЧНИХ ТА НЕЙРОМЕРЕЖЕВИХ МЕТОДІВ РЕГРЕСІЙНОГО ПРОГНОЗУВАННЯ

Іван Хамар, Ігор Оленич

STUDY OF THE EFFECTIVENESS OF CLASSICAL AND NEURAL NETWORK METHODS FOR REGRESSION FORECASTING

Ivan Khamar, Ihor Olenych

Львівський національний університет імені Івана Франка,
вул. Драгоманова, 50, Львів, 79005, Україна

Abstract. *This paper presents a comparative analysis of classical and neural regression models for processing large datasets. Three approaches were implemented: Random Forest Regressor and two Dense Neural Networks — a basic PyTorch version and an improved TensorFlow model with Dropout and BatchNormalization layers. While the classical model achieved moderate accuracy ($R^2 = 0.7404$), the enhanced DNN yielded the best results ($RMSE = 0.1570$, $R^2 = 0.9346$). The study confirms the potential of neural networks for high-precision regression, provided that regularization, normalization, and hyperparameter tuning are applied.*

У сучасних умовах стрімкого зростання обсягів даних, які генеруються з різноманітних джерел — від фінансових операцій і медичних систем до інтернет-сервісів та сенсорів Інтернету речей — зростає потреба в ефективних методах обробки, аналізу та прогнозування. Зокрема, у разі регресійного прогнозування на основі великомасштабних і часто насичених шумовими проявами наборів даних, виникає необхідність вибору між класичними алгоритмами прогнозування, використанням оптимізованих методів машинного навчання та сучасними нейромережевими підходами [1]. Класичні методи, такі як лінійна регресія, регресійні дерева та ансамблеві підходи на кшталт Random Forest Regressor (RFR), відзначаються відносною простотою реалізації, інтерпретованістю та стійкістю до перенавчання [2]. Водночас нейромережеві моделі — особливо глибокі архітектури — демонструють високий потенціал при роботі з високорозмірними, гетерогенними та нелінійними даними, хоча й часто вимагають значно більших обчислювальних ресурсів, а також складнішого налаштування і оптимізації [3].

У роботі досліджено ефективність моделей машинного навчання для регресійного прогнозування рейтингу фільму за структурованими числовими та категоріальними ознаками на основі табличного набору даних IMDb Top 5000 Movies. Кожен запис у наборі даних містить характеристики різного типу, зокрема назву фільму, рік випуску, жанр, інформацію про режисера та головних акторів, тривалість фільму, дані про країну походження, мову, компанію-виробника, рейтинг IMDb, кількість голосів і валовий дохід тощо. Цільовою змінною для завдання регресії є рейтинг IMDb, представлений у вигляді дійсного постійного значення в діапазоні від 1,0 до 10,0. Після попереднього аналізу даних (EDA [4]) виконано one-hot кодування жанрів і цільове кодування імені режисера, що дозволило сформувати оптимізовану матрицю ознак. Це створило підґрунтя для навчання моделей з покращеною здатністю до узагальнення та стійкістю до перенавчання [5].

У рамках дослідження було порівняно дві групи моделей: RFR та глибока щільна нейронна мережа (DNN) [3]. Метод RFR реалізовано в scikit-learn з 300 деревами та вилучено ознаку «rank» для запобігання витоку. DNN реалізовано у двох варіантах: базовому (на PyTorch) і покращеному (на TensorFlow), де використано Dropout, Batch Normalization, early

stopping, adaptive learning rate та архітектурне масштабування. Для оцінювання застосовано метрики MAE, RMSE та R^2 [6]. Оптимізація моделі DNN дозволила досягти $R^2 > 0,93$. Навчання здійснювалося на 80 % даних, при цьому для DNN виділено 20 % валідаційної вибірки [7]. Єдина тестова вибірка з `random_state=42` забезпечила стабільність і відтворюваність порівняння.

Для побудови нейромережевої моделі DNN було застосовано кількешарову архітектуру типу Dense(256) → Dense(128) → Dense(64), яка забезпечує поступове зниження розмірності та кращу апроксимацію складних нелінійних залежностей. Щоб запобігти перенавчанню, використано техніки Dropout (0,2–0,1) та Batch Normalization після кожного шару, що також прискорює збіжність і стабілізує градієнти. Навчання виконувалося із застосуванням Adam-оптимізатора (learning rate = 0,001) у зв'язі з learning rate scheduling для адаптивного зменшення кроку при стагнації. Додатково реалізовано Early Stopping з patience = 15, що дозволило зберегти найкращі ваги та уникнути перенавчання. Було обрано розмір мініпакету 32, що забезпечило часті оновлення ваг і сприяло стохастичній стабільності. Усі параметри архітектури та гіперпараметри моделі були підібрані вручну на основі серії валідаційних експериментів.

У результаті моделювання було проведено порівняння трьох конфігурацій регресійних моделей: RFR, базової DNN та покращеної DNN. Найвищу точність за всіма метриками (RMSE = 0,1570, MAE = 0,1008, $R^2 = 0,9346$) продемонструвала налаштована щільна нейромережа DNN, що свідчить про ефективність глибоких архітектур за умови оптимізації. Базова DNN без додаткового налаштування показала гірші результати, поступившись навіть класичній моделі RFR ($R^2 = 0,7404$). Аналіз підтвердив ключову роль архітектурних рішень і оптимізації гіперпараметрів у точності DNN, а також актуальність класичних моделей у задачах з обмеженими ресурсами.

Бібліографія

1. Bjerre L.M., Peixoto C., Alkurd R., Talarico R., Abielmona R. Comparing AI/ML approaches and classical regression for predictive modeling using large population health databases: Applications to COVID-19 case prediction. *Global Epidemiology*. 2024. Vol. 8. 100168. <https://doi.org/10.1016/j.gloepi.2024.100168>
2. Breiman L. Random Forests. *Machine Learning*. 2001. Vol. 45(1). P. 5–32. <https://doi.org/10.1023/A:1010933404324>
3. Chen Y., Zhang Y. A Study of Optimization in Deep Neural Networks for Regression. *Electronics*. 2023. Vol. 12(14). Article 3071. <https://doi.org/10.3390/electronics12143071>
4. Mehta V., Batra N., Poonam, Goyal S., Kaur A., Dudekula K. V., Victor G. J. Machine Learning based Exploratory Data Analysis (EDA) and Diagnosis of Chronic Kidney Disease (CKD). *EAI Endorsed Transactions on Pervasive Health and Technology*. 2024. Vol. 10. <https://doi.org/10.4108/eetpht.10.5512>
5. Pargent F., Pfisterer F., Thomas J., Bischl B. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Statistical Papers*. 2022. Vol. 63. P. 1353–1375. <https://doi.org/10.1007/s00180-022-01207-6>
6. Chicco D., Warrens M. J., Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*. 2021. Vol. 7. Article e623. <https://doi.org/10.7717/peerj-cs.623>
7. Joseph A., Singh A. Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 2022. Vol. 15(1). P. 1–10. <https://doi.org/10.1002/sam.11583>