DOI 10.34185/1562-9945-5-161-2025-09
УДК 004.89

І.Д. Віжуткін, Т.А. Ліхоузова

ПРОГНОЗУВАННЯ АКТИВНОСТІ КОРИСТУВАЧІВ У ВІДЕОІГРАХ

Анотація. Нині прогнозування активності користувачів у відеоіграх є надзвичайно актуальним завданням, яке має важливе значення для розробників, видавців та аналітиків ігрової індустрії. Точне прогнозування кількості онлайн-гравців дозволяє приймати стратегічні рішення щодо оновлень, маркетингових кампаній та розвитку проєктів, а також оптимізувати серверні потужності й ресурси. У рамках даної роботи проведено аналіз історичних та поточних даних про онлайн активність у іграх платформи Steam. Для розв'язання задачі прогнозування були застосовані методи машинного навчання Gradient Boosting, LinearSVR, а також моделі часових рядів (S)ARIMA(X). Результат роботи може бути використаний для прогнозування кількості онлайн-гравців у грі через місяць на основі історичних та поточних параметрів.
Ключові слова: аналіз даних, прогнозування, Gradient Boosting, LinearSVR, ARIMA, SARIMA, ARIMAX, SARIMAX.

Активність гравців у відеоіграх є одним із ключових індикаторів успішності проєкту на ринку розваг. Кількість активних користувачів напряму впливає на прибуток розробників та видавців, розвиток гри, залучення нових гравців і формування спільноти. Steam, найбільша цифрова платформа для розповсюдження ігор, надає доступ до великих обсягів відкритих даних про активність гравців у режимі реального часу. За допомогою цих даних можна відстежувати зміну популярності окремих проєктів, аналізувати загальні тренди індустрії та виявляти залежності між подіями (наприклад, оновленнями гри чи знижками) та поведінкою гравців.

Активність користувачів у грі залежить від багатьох факторів: як внутрішніх (зміни в ігровому процесі, релізи доповнень, маркетингові кампанії), так і зовнішніх (сезонні коливання, економічна ситуація, поява нових конкурентних продуктів). Це створює складне динамічне середовище, де прості моделі часто не можуть забезпечити достатньо точних прогнозів.

Правильне прогнозування кількості гравців має велике практичне значення. Завдяки йому можна планувати навантаження на сервери, оцінювати доцільність випуску оновлень, визначати найкращий час для маркетингових активностей або запуску нових функцій. Також передбачення майбутньої популярності може бути важливим для фінансового планування компанії та аналізу ризиків.

З огляду на складність та варіативність даних, для аналізу і прогнозування використовуватимуться сучасні методи машинного навчання та моделі часових рядів. Вони

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

дозволять виявляти приховані закономірності в даних, обробляти великі обсяги інформації та підвищувати точність прогнозів навіть у нестабільних умовах.

Метою дослідження є побудова та аналіз моделей методами машинного навчання для прогнозування кількості онлайн-гравців.

Матеріали та методи. Для вирішення поставленої задачі були обрані три основні набори даних, які надають основну інформацію для аналізу ігор з різних сторін.

− загальна інформація про ігри на платформі Steam [1];

− окремі відгуки для ігор на платформі Steam [2];

− історія онлайну та цін для ігор на платформі Steam [3].

Для зручної роботи з даними розроблено сховище даних. Даний процес включав такі основні етапи, як завантаження файлів із даними, первинна обробка за допомогою мови Python та бібліотеки Pandas, завантаження даних до Stage-зони бази даних PostgreSQL, обробка та їх перенесення до основного сховища (рисунок 1).

Для розв'язання задачі прогнозування кількості онлайн-гравців у відеогрі через місяць було обрано три різнотипних підходи: Gradient Boosting, LinearSVR та моделі часових рядів (S)ARIMA(X). Кожен із методів має власні особливості та переваги, що дозволяє отримати більш об'єктивну оцінку якості прогнозу в різних умовах.

Gradient Boosting (LightGBM, XGBoost) [4-6] є одним із найефективніших ансамблевих методів машинного навчання для задач регресії. Дана технологія будує сильну модель шляхом послідовного поєднання великої кількості слабких моделей (дерев рішень), зосереджуючись на корекції помилок попередніх моделей. Її основною перевагою є висока точність, здатність обробляти як числові, так і категоріальні дані, а також хороша стійкість до великої кількості фіч, пропусків або шумів у даних.

LinearSVR [7] – це модифікація методу опорних векторів для задач регресії. Вона добре працює на задачах, де залежність між ознаками і цільовою змінною можна апроксимувати лінійною функцією. LinearSVR вибрано через його простоту, швидкість навчання та хорошу здатність узагальнювати дані при великій кількості фіч. Вона є корисної базовою моделлю для порівняння.

(S)ARIMA(X) [8] – це класична модель аналізу часових рядів, що поєднує авторегресію (AR), інтегрування (I) і ковзне середнє (MA), а також дозволяє враховувати сезонність (S) та зовнішні регресори (X). Застосування цих моделей є повністю обґрунтованим, оскільки в роботі аналізується типовий часовий ряд – зміна онлайн-активності гравців по днях. У випадку (S)ARIMA використовується лише історія самого онлайну, тоді як (S)ARIMAX дає змогу включити додаткові чинники, такі як знижки, ціни, кількість відгуків.

Таким чином, використання цих трьох підходів дозволяє порівняти результати моделей, що орієнтуються на різні аспекти структури даних: ансамблеві моделі машинного навчання, лінійні моделі та моделі часових рядів.

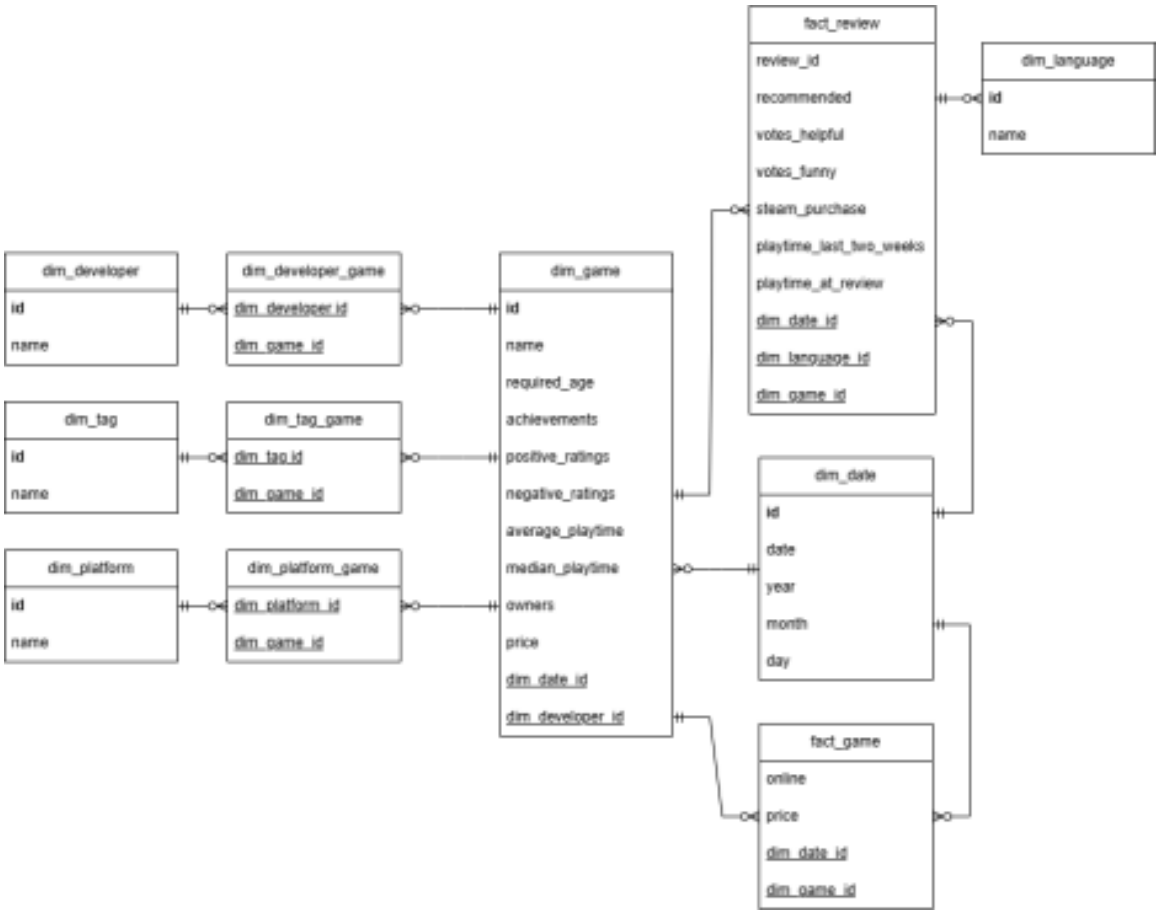

Рисунок 1 – Модель основного сховища

Результати та основний матеріал дослідження. Після побудови моделей Gradient Boosting, LinearSVR та SARIMA(X) було визначено, що останні показують найкращі результати (таблиця 1), які, в цілому, вже можна використовувати для прогнозування онлайну відеоігор через місяць.

Таблиця 1

Порівняння використаних моделей

Модель	Швидкість навчання, включаючи підбір параметрів	Точність результатів
XGBoost (Gradient Boosting)	20 хв	Близько 90% відхилень від реальних значень.
LGBMBoost (Gradient Boosting)	20 хв	Близько 40% відхилень від реальних значень.
LinearSVR	3 хв	Близько 40% відхилень від реальних значень.
SARIMA	1 хв	Близько 20% відхилень від реальних значень.
SARIMAX	1.5 хв	Близько 20% відхилень від реальних значень, але з меншою дисперсією.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Для обрання найкращої моделі для вирішення поставленої задачі прогнозування онлайну відеоігор через місяць, проведемо кінцеве тестування (рисунки 2-4) для трьох різних по популярності ігор та винесемо кінцеве рішення.

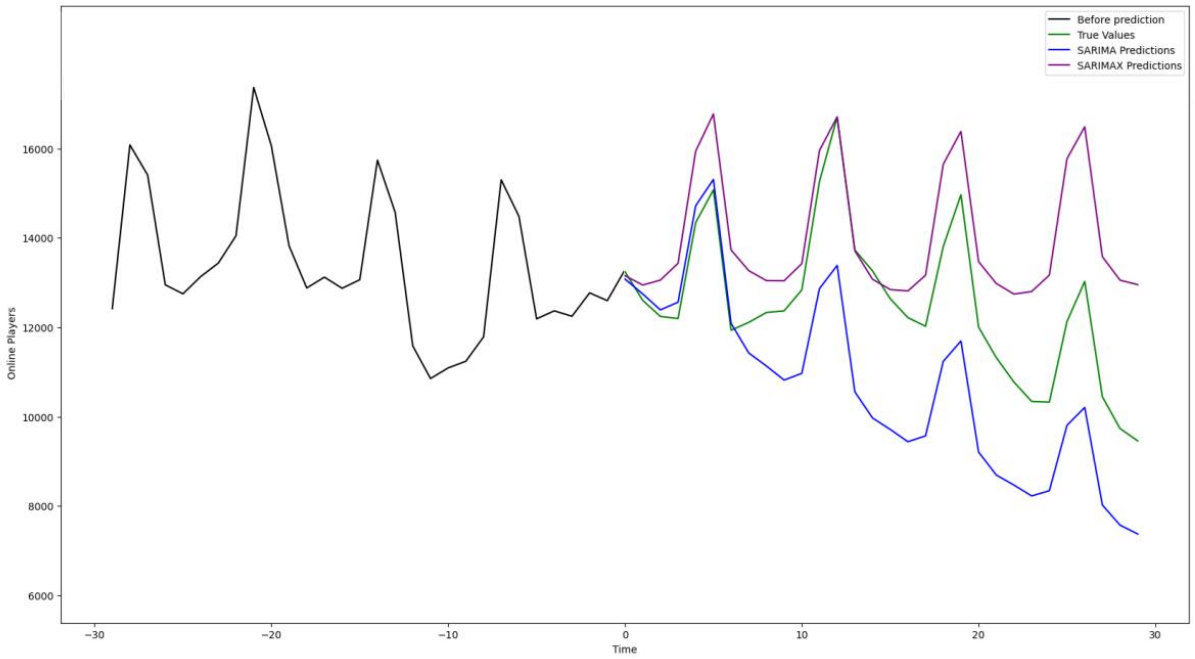

Рисунок 2 – Графіки прогнозованого онлайну популярної гри

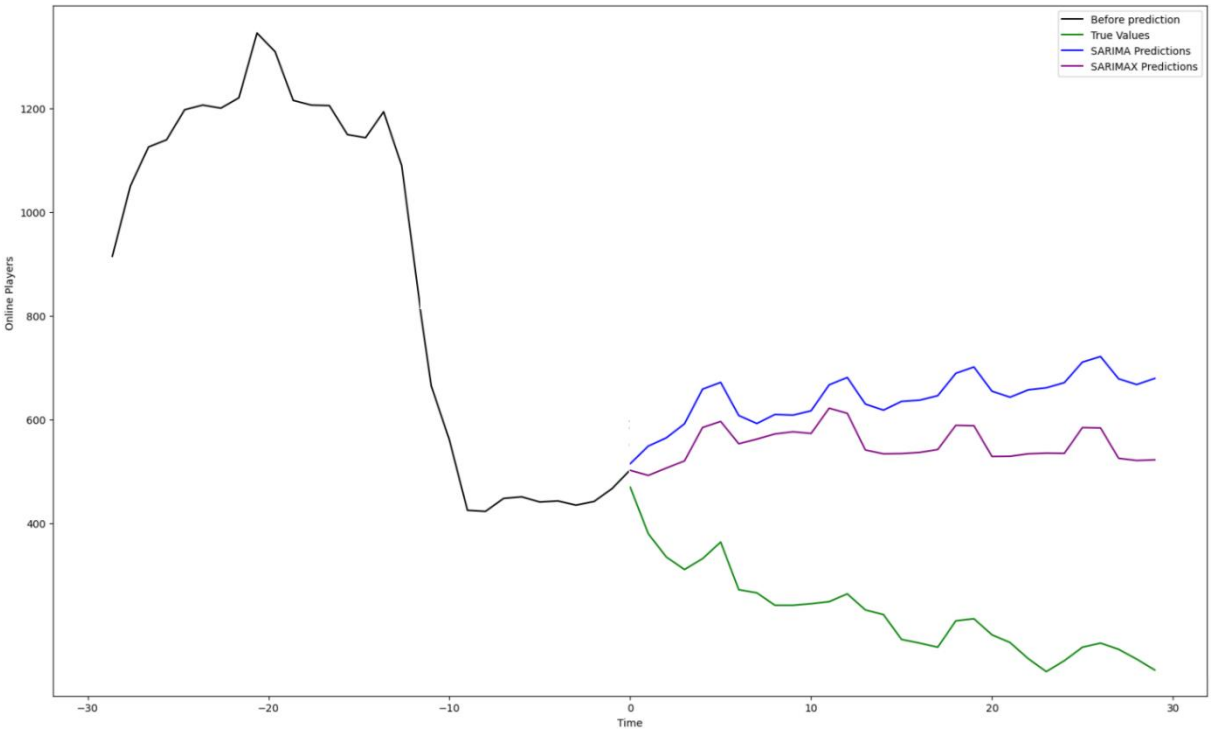

Рисунок 3 – Графіки прогнозованого онлайну середньопопуярної гри

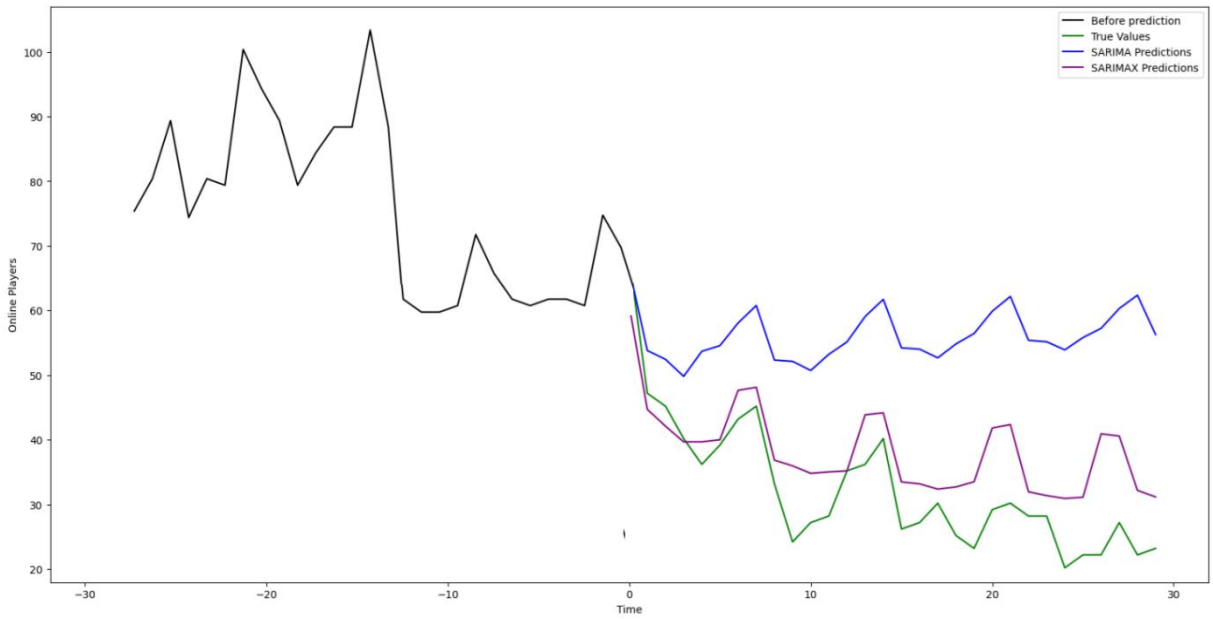

Рисунок 4 – Графіки прогнозованого онлайну непопулярної гри

Бачимо, що незважаючи на те, що середній показник MAPE при побудові та тестуванні моделі SARIMA був трохи кращий, SARIMAX показує більш стабільні результати для будь-яких обраних ігор, найкраще наближуючись до реальних значень.

Висновки. У даній роботі описано результати реалізації повного циклу вирішення задачі прогнозування кількості онлайн-гравців у відеоіграх через місяць на основі історичних та поточних даних.

У рамках інтелектуального аналізу даних протестовано декілька моделей прогнозування, серед яких Gradient Boosting (XGBoost, LightGBM), LinearSVR та моделі часових рядів SARIMA та SARIMAX. За результатами попереднього тестування було встановлено, що моделі SARIMA та SARIMAX забезпечують найнижчі середні показники похибки (близько 20–22%), тоді як інші методи мали значно вищу похибку (близько 40%), що не дозволяє ефективно застосовувати їх для вирішення поставленої задачі

Фінальне тестування моделей SARIMA та SARIMAX на прикладі трьох ігор з різними рівнями популярності дозволило зробити висновок, що, незважаючи на незначне відставання SARIMAX за середнім значенням MAPE, ця модель забезпечує більш стабільні результати прогнозування, краще пристосовується до особливостей різних типів ігор та точніше наближається до реальних значень онлайну. Таким чином, найоптимальнішим підходом до прогнозування майбутнього онлайну відеоігор виявилася модель SARIMAX, яка демонструє високу точність, стійкість та гнучкість.

Отримані результати можуть бути корисними для подальшого застосування в ігровій індустрії з метою планування навантажень на сервери, маркетингових кампаній та загального моніторингу динаміки інтересу до продукту.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

ЛІТЕРАТУРА / REFERENCES

1. General information for games on the Steam platform.
URL: https://www.kaggle.com/datasets/nikdavis/steam-store-games (date of access: 28.04.2025).
2. Reviews for games on the Steam platform.
URL: https://www.kaggle.com/datasets/najzeko/steam-reviews-2021/data (date of access: 28.04.2025).
3. Online and pricing history for games on the Steam platform.
URL: https://data.mendeley.com/datasets/ycy3sy3vj2/1 (date of access: 28.04.2025). 12. PostgreSQL. URL: https://www.postgresql.org/ (date of access: 28.04.2025).
4. Gradient Boosting. URL: https://en.wikipedia.org/wiki/Gradient_boosting (date of access: 28.04.2025).
5. XGBoost. URL: https://xgboost.readthedocs.io/en/release_3.0.0/ (date of access: 28.04.2025).
6. LightGBM. URL: https://lightgbm.readthedocs.io/en/stable/ (date of access: 28.04.2025).
7. LinearSVR. URL: https://habr.com/ru/articles/802185/ (date of access: 28.04.2025).
(S)ARIMA(X). URL: https://habr.com/ru/articles/477206/ (date of access: 28.04.2025).

Received 26.08.2025.
Accepted 29.08.2025.

Predicting user activity in video games

This paper describes the results of the implementation of a full cycle of solving the problem of predicting the number of online players in video games in a month based on historical and current data.

Within the framework of data mining, several forecasting models were tested, including Gradient Boosting (XGBoost, LightGBM), LinearSVR, and the SARIMA and SARIMAX time series models. According to the results of preliminary testing, it was found that the SARIMA and SARIMAX models provide the lowest average error rates (about 20–22%), while other methods had a significantly higher error (about 40%), which does not allow them to be effectively used to solve the problem.

The final testing of the SARIMA and SARIMAX models on the example of three games with different levels of popularity allowed us to conclude that, despite the slight lag of SARIMAX in terms of the average MAPE value, this model provides more stable forecasting results, better adapts to the characteristics of different types of games, and more accurately approaches real online values. Thus, the most optimal approach to predicting the future of online video games was the SARIMAX model, which demonstrates high accuracy, stability and flexibility.

The results obtained may be useful for further application in the gaming industry for the purpose of planning server loads, marketing campaigns and general monitoring of the dynamics of interest in the product.

Keywords: data analysis, forecasting, Gradient Boosting, LinearSVR, ARIMA, SARIMA, ARIMAX, SARIMAX.

Ліхоузова Тетяна Анатоліївна – к.т.н., доцент кафедри інформатики та програмної інженерії, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», ORCID: https://orcid.org/ 0000-0002-4132-3979
Віжуткін Ілля Дмитрович – студент кафедри інформатики та програмної інженерії, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», ORCID: https://orcid.org/0009-0007-2824-9392

Likhouzova Tetiana – PhD, associate professor, Department of Informatics and Software Engineering National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», ORCID: https://orcid.org/ 0000-0002-4132-3979
Vizhutkin Illia – student, Department of Informatics and Software Engineering National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», ORCID: https://orcid.org/0009-0007-2824-9392

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)