УДК 519.237.5 : 004.852 DOI: https://doi.org/10.30837/IYF.CVSAMM.2024.226

РОЗВ'ЯЗАННЯ ЗАДАЧІ ПРОГНОЗУВАННЯ ВІДТОКУ ГРАВЦІВ В ІГРОВІЙ ІНДУСТРІЇ

Ломія С.Г.
Науковий керівник – канд. техн. наук, доц. Гибкіна Н.В.
Харківський національний університет радіоелектроніки, каф. ПМ,
м. Харків, Україна
e-mail: serhii.lomiia@nure.ua

The problem of predicting the outflow of players in the gaming industry as a classification problem is considered. To solve the problem using machine learning methods, it is proposed to use the gradient boosting method. To improve the quality of classification, it is proposed to identify the features that are most influential for the churn forecasting.

Ігрова індустрія займається створенням, розробкою та розповсюдженням відеоігор. Останніми роками вона швидко розвивається і відіграє важливу роль не лише у культурі завдяки можливостям для відпочинку, соціалізації та творчості, а й у економіці та технологічному прогресі за рахунок сприяння розвитку графічних та звукових технологій, віртуальної реальності, штучного інтелекту. Комп'ютерні та мобільні ігри – це серйозний бізнес, який привертає увагу інвесторів та великих корпорацій. Величезна конкуренція у ігровій галузі призводить до того, компанії вимушені утримувати свою аудиторію. Відтік гравців – це явище, коли гравці припиняють активну участь у грі. Виходячи з цього необхідно оптимізувати утримання гравців і підвищити їхню участь у грі, запобігаючи відтоку. Отже, задача прогнозування відтоку гравців є актуальною.

Як задача машинного навчання прогнозування відтоку гравців є задачею класифікації, яка полягає у розбитті гравців на два класи: тих, які залишаються у грі, та тих, які йдуть з неї; а також у передбаченні для кожного окремого гравця класу, до якого він належить.

Для розв'язання задачі було використано набір даних гравців мобільної гри в жанрі Fantasy Collection RPG за один з кварталів 2023 року. Набір даних складається з 19882 рядків та 165 стовпців (показників гри гравця за останні 7, 14, 30 та 60 днів). Особливостями набору даних, по-перше, є велика кількість пропусків (коли гравець не має дій в тій чи іншій активності), і, по-друге, його незбалансованість (близько 10 % гравців мають статус відвалу).

Після попередньої обробки даних, яка передбачає обробку пропущених значень та викидів, кодування категоріальних ознак, масштабування числових ознак, розбиття вибірки на тренувальний, валідаційний та тестовий набори, переходимо до розробки базової моделі та її подальшого вдосконалення. Для класифікації у дослідженні використовувався метод Gradient Boosting. Він є потужним методом ансамблевого навчання, що передбачає комбінування слабких учнів (зазвичай дерев рішень) для отримання більш точної і сильної моделі.

Для налаштування гіперпараметрів моделі запропоновано використовувати бібліотеку Optuna, що на досліджуваному наборі даних дозволило визначити оптимальні значення кількості дерев у градієнтному алгоритмі, швидкості навчання, параметру обрізання дерев, коефіцієнту регуляризації, кількості раундів без покращень для зупинки навчання. Використання цих значень дозволить побудувати базову модель класифікації з найвищою очікуваною ефективністю за визначеною цільовою метрикою.

При побудові моделі класифікації ми спирались на мінімізацію False Positive (FP – помилкове віднесення до класу тих, хто покине гру) та не високе значення False Negative (FN – помилкове віднесення до класу тих, хто лишиться) і шукали найкращий баланс між цими показниками [1]. За значеннями матриці невідповідності розраховано оптимальне значення $F1score = 0,66$ для оптимального співвідношення показників Precision (0,667) та Recall (0,634). З 4188 гравців валідаційної вибірки модель розмічає правильно 291 гравця як таких, що відвалюються (True Positive, TP – churn-гравець), але ще 168 гравців, які також є churn-гравцями, пропускає (FN). Помилково до класу churn модель відносить 145 гравців (FP). Це середня якість моделі, отже, в майбутньому намагатимемось її підвищити.

Для цього дослідимо вплив окремих ознак. Shap (SHapley Additive exPlanations) – метод, який використовує числа Шеплі (Shapley values) для встановлення важливості кожної ознаки у моделі прогнозу. На досліджуваному наборі даних на рівні 0,1 він як найважливіші виділив ознаки з груп: Summon (показує, як багато гравець відкриває нових героїв зі спеціального айтему); ознаки активності за деякий період (зокрема, DaysInAROW – кількість днів, які підряд грає гравець); BattleDoomTower (показує кількість битв у розділі на мапі, який є доволі складним для гравців). Так при високому значенні DaysInARow ймовірність відтоку гравця знижується. При низькому значенні SummonsSacredLegendaryD7 ймовірність відтоку підвищується, але як тільки цей показник підвищується, тобто гравці отримують найкращих героїв, то ймовірність відтоку знижується. Аналогічно визначаються нечіткі ознаки, котрі не дають чіткого розуміння впливу на ймовірність відтоку гравця.

Врахування оптимальних значень параметрів та тільки значущих ознак дозволило отримати модель, яка краще прогнозує ймовірність відтоку гравців у незбалансованих вибірках з використанням таких ознак як, наприклад, активність гравців у батлах на всіх існуючих локаціях у грі, у призові героїв за останні 7,14,30 та 60 днів ($F1score = 0,70$, TP=257, FN=113, FP=111).

Список використаних джерел:

1. Raschka S., Mirjalili V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow. Packt Publishing, 2019. 622 p.