

[2] Hurley, N. and Zhang, M. (2011). Novelty and diversity in top-n recommendation analysis and evaluation. ACM Transactions on Internet Technology.

Дарій Бензак, Юрій Процик*

магістрант кафедри КН, НЛТУ України, Львів, Україна.

* к.ф.-м.н., доцент кафедри КН, НЛТУ України, Львів, Україна.
protsyk@nltu.edu.ua.

Розроблення системи прогнозування результатів матчів зі снукеру з використанням методів машинного навчання

Анотація – В роботі розглядаються питання, пов’язані з розробленням системи прогнозування результатів матчів зі снукеру на мові програмування Python з використанням методів машинного навчання. Описуються особливості реалізації засобів збору інформації про результати матчів, попередня обробка даних, налаштування декількох моделей машинного навчання та оцінка якості роботи відповідних алгоритмів.

Keywords – снукер, система прогнозування, машинне навчання, класифікація, Python.

Вступ. Снукер, як один із найпопулярніших видів більярду, завоював серця мільйонів шанувальників по всьому світу. Ця стратегічна гра, заснована на точності та вмінні прогнозувати ходи, викликає захоплення як у професійних спортсменів, так і у любителів. Розвиток технологій та поширення методів машинного навчання приводять до новаторських підходів у різних сферах, включаючи прогнозування результатів спортивних подій.

Задача прогнозування результатів матчів зі снукеру може бути сформульована як задача бінарної класифікації, мета якої полягає в тому, щоб на основі вхідних ознак про матч (наприклад, статистика гравців, рейтинг, історія зустрічей) визначити ймовірність перемоги одного з гравців або класифікувати матч як перемога чи поразка для конкретного гравця.

Процес бінарної класифікації включає наступні кроки:

Збір даних: зібрати вхідні дані про матчі зі снукеру, включаючи ознаки гравців, їхню статистику, рейтинги, історію зустрічей та інші відповідні дані.

Підготовка даних: провести попередню обробку даних, таку як очищення, масштабування та кодування категоріальних ознак.

Вибір моделі: вибрати алгоритм класифікації, який найкраще підходить для даної задачі.

Тренування моделі: навчити модель на навчальній вибірці, використовуючи входні ознаки та відповідні класи (перемога/поразка).

Оцінка моделі: оцінити продуктивність моделі на тестових даних, використовуючи різного роду метрики.

Прогнозування: використати навчену модель для прогнозування результату матчу зі снукеру на основі нових входних даних.

Виклад основного матеріалу. З метою отримання результатів за останні роки та подальшого аналізу даних було вирішено скористатися [snooker.org API](https://api.snooker.org/) [1], що є безкоштовним для некомерційного використання. API [snooker.org](https://api.snooker.org/) надає можливість отримання наступної інформації щодо снукеру: турніри, матчі, гравці, турніри в сезоні, всі матчі заданого турніру, всі матчі для певного гравця в сезоні, рейтинги тощо. Щоб скористатися цим API потрібно зв'язатися з розробниками та отримати значення X-Requested-By, що задається в заголовку запитів.

Таким чином, було зібрано інформацію про результати матчів зі снукеру у професійних турнірах, починаючи з 2004 року, та збережено у csv-файл. Загальна кількість матчів – 14452.

На етапі підготовки даних створено наступні ознаки:

- *DiffSeeding* – відношення різниці посіву гравців в турнірі до посіву першого;
- *AtHome* – гра вдома (по відношенню до першого гравця), що приймає значення -1, 0 та 1;
- *OddsLastRounds* – логарифм відношення кількостей ігор гравців на пізніх стадіях турнірів;
- *OddsRatio* – логарифм відношення шансів на перемогу в матчі на основі історичних даних гравців для заданого турніру;

- *WinDiffPlayer1Player2* – різниця перемог в очних протистояннях;
- *Player1YearsAsPro* та *Player2YearsAsPro* – кількості років в професійному снукері для першого та другого гравців відповідно.

Крім зазначених ознак, використано повні імена гравців (*Player1FullName*, *Player2FullName*), національність (*Player1Nationality*, *Player2Nationality*) та стадію турніру (*RoundName*). Цільова ознака *Winner* – бінарна ознака, що набуває значення 1 у разі перемоги першого гравця.

Вибірку було розбито на навчальну (11000 записів) та тестову (3452 записи), а також проведено масштабування кількісних ознак та кодування категоріальних. Останнє полягає у створенні для категоріальної ознаки набору фіктивних бінарних ознак, кожна з яких відповідає за певне значення категоріальної (одна з ознак дорівнює одиниці, а решта – нулі).

Для розв’язання задачі класифікації використано такі Python-бібліотеки як Scikit-learn та XGBoost, які дозволяють спростити процес створення класифікатора та більш чітко виділити концепції машинного навчання. Було обрано три моделі: логістичну регресію, випадковий ліс та XGBoost, що є однією з ефективних реалізацій градієнтного бустінгу. Підбір гіперпараметрів та навчання кожної з них здійснювалися з використанням перехресної перевірки (крос-валідації) [2].

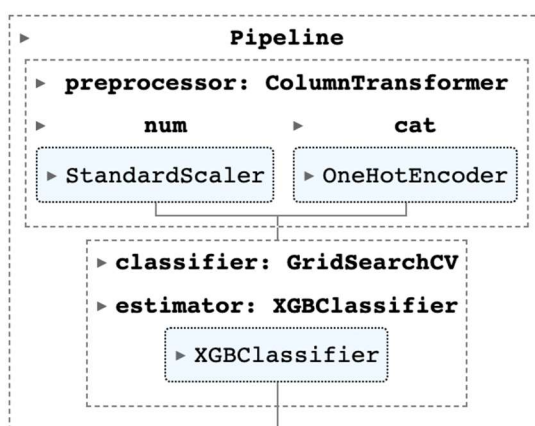


Рис. 1. Етапи налаштування XGBoost класифікатора

Порівнюючи якість роботи алгоритмів на тестовій вибірці з використанням такої метрики як збалансована частка правильних відповідей

(*balanced_accuracy_score*) [3], було отримано наступні результати: випадковий ліс – 0.63, логістична регресія та XGBoost – 0.64.

Оскільки всі моделі показали приблизно однакові результати, а логістична регресія є найбільш простою та добре інтерпретованою моделлю, то було обрано саме її. Розгорнуту оцінку якості моделі логістичної регресії з використанням метрик точності, повноти та F-міри подано на рисунку 2.

	precision	recall	f1-score	support
0	0.60	0.54	0.57	1445
1	0.69	0.75	0.72	2007
accuracy			0.66	3452
macro avg	0.65	0.64	0.64	3452
weighted avg	0.65	0.66	0.65	3452

Рис. 2. Оцінка якості моделі логістичної регресії

Слід зауважити, що тут частка правильних відповідей (*accuracy*) відрізняється від збалансованої частки (*balanced_accuracy_score*), оскільки тестова вибірка не є збалансованою (1 – 58%, 0 – 42%).

Висновок. В роботі розглянуто питання, пов’язані з розробленням системи прогнозування результатів матчів зі снукеру з використанням методів машинного навчання. Реалізовано засоби збору даних для системи, на основі яких побудовано модель прогнозування. Впровадження такої системи може сприяти покращенню якості спортивних прогнозів, розвитку аналітики у галузі снукеру та забезпеченню нових можливостей для професіоналів та любителів цієї гри.

Список використаних літературних джерел

- [1] snooker.org API [Електронний ресурс]. – Режим доступу: <https://api.snooker.org/>.
- [2] Scikit-learn User Guide. Tuning the hyper-parameters of an estimator [Електронний ресурс]. – Режим доступу: https://scikit-learn.org/stable/modules/grid_search.html#grid-search.
- [3] Scikit-learn User Guide. Balanced accuracy score [Електронний ресурс]. – Режим доступу: https://scikit-learn.org/stable/modules/model_evaluation.html#balanced-accuracy-score.