УДК 004.85:796.332

О. І. Зеленцов[1]
С. О. Жуков[1]

ІНФОРМАЦІЙНО-АНАЛІТИЧНА СИСТЕМА ПРОГНОЗУВАННЯ РЕЗУЛЬТАТІВ ФУТБОЛЬНИХ МАТЧІВ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

[1]Вінницький національний технічний університет

Анотація

У роботі розглянуто розробку інформаційно-аналітичної системи для прогнозування результатів футбольних матчів на основі методів машинного навчання. Запропоновано підхід до інженерії ознак, що включає розрахунок динамічного рейтингу ELO, показників поточної форми команд за останні 3 та 5 матчів, атакувальної й оборонної ефективності, а також різницевих метрик між господарями та гостями. Реалізовано порівняння моделей Logistic Regression, Random Forest, Gradient Boosting, XGBoost та Neural Network у мультикласовій і бінарній постановках задачі.

Для загального оцінювання моделей використано хронологічне розбиття 80/20, а для практичної апробації — прогнозування окремих турів сезону. У мультикласовій задачі найвищу Accuracy у загальному тестуванні показала модель XGBoost — 56.6%, що є співставним із букмекерським baseline на рівні 55.7%. У прогнозуванні окремих турів найвищі результати були отримані в бінарній постановці задачі: для 38-го туру модель XGBoost досягла 100% Accuracy на контрольному наборі з 10 матчів. Розроблено інтерактивний вебдодаток на базі Streamlit із модулем сценарного моделювання What-If.

Ключові слова: *машинне навчання, прогнозування футбольних матчів, ELO, XGBoost, Gradient Boosting, Streamlit, What-If аналіз.*

Abstract

This paper considers the development of an information-analytical system for predicting football match outcomes using machine learning methods. The proposed approach includes feature engineering based on dynamic ELO ratings, recent team form over the last 3 and 5 matches, attacking and defensive efficiency, and difference-based features between home and away teams. Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and Neural Network models were compared in both multiclass and binary prediction settings.

Chronological 80/20 splitting was used for general model evaluation, while separate match rounds were used for practical forecasting experiments. In the multiclass task, XGBoost achieved the highest accuracy of 56.6%, which is comparable to the betting market baseline of 55.7%. In round-based forecasting experiments, the best results were obtained in the binary setting: for the 38th round, XGBoost achieved 100% accuracy on a control set of 10 matches. An interactive Streamlit web application with a What-If scenario modeling module was developed.

Keywords: *machine learning, football match prediction, ELO rating, XGBoost, Gradient Boosting, Streamlit, What-If analysis.*

Вступ

Сучасний розвиток спортивної аналітики пов’язаний зі зростанням обсягів статистичних даних та потребою в інтелектуальних системах для їх обробки. Однією з найскладніших задач є прогнозування результатів футбольних матчів, оскільки футбол характеризується високою стохастичністю, низькою результативністю та значним впливом випадкових подій на підсумковий рахунок.

Метою роботи є розробка інформаційно-аналітичної системи для прогнозування результатів футбольних матчів із використанням алгоритмів машинного навчання.

Для досягнення мети було поставлено такі завдання:

- виконати попередню обробку історичних футбольних даних;
- сформувати набір інформативних ознак на основі форми команд, результативності, різницевих метрик та рейтингу ELO;
- порівняти кілька моделей машинного навчання;
- реалізувати мультикласову постановку задачі H/D/A та бінарну постановку H/Not_H;
- виконати прогнозування окремих турів сезону;
- розробити інтерактивний вебдодаток із модулем What-If аналізу.

Результати дослідження

У межах роботи було створено програмний комплекс для прогнозування результатів футбольних матчів Англійської Прем'єр-Ліги. Система реалізована мовою Python із використанням бібліотек Pandas, Scikit-learn, XGBoost, Matplotlib, Seaborn та Streamlit.

На етапі підготовки даних було виконано сортування матчів за датою та сформовано динамічні ознаки, які розраховуються лише на основі попередніх матчів. Це дозволило уникнути витоку інформації з майбутнього. Для кожної команди були розраховані показники форми, атаки та захисту за останні 3 і 5 матчів, а також рейтинг ELO.

Для загального порівняння моделей було використано хронологічне розбиття 80/20. У мультикласовій задачі прогнозування результату матчу H/D/A найкращу Accuracy показала модель XGBoost — 56.6%. Neural Network продемонструвала найнижчий Log-Loss — 0.956, що свідчить про кращу якість імовірнісних оцінок. Logistic Regression мала найкращий F1-score — 0.48, що вказує на більш збалансовану поведінку між класами.

Важливим орієнтиром для інтерпретації результатів став букмекерський baseline. Простий прогноз за найменшим коефіцієнтом Bet365 показав 55.7% Accuracy. Отже, результат XGBoost на рівні 56.6% є співставним із ринковим орієнтиром, що підтверджує складність задачі футбольного прогнозування.

Окремо було проведено практичну апробацію системи на конкретних турах сезону. На 30-му турі в бінарній постановці задачі H/Not_H найкращі моделі досягли 80% Accuracy. У режимі навчання лише на поточному сезоні найкращою стала Neural Network завдяки найнижчому Log-Loss — 0.456. У режимі з додаванням попереднього сезону найкращою стала Random Forest із 80% Accuracy та Log-Loss 0.579.

Для 38-го туру в мультикласовій задачі найкращий результат показала Gradient Boosting. При використанні попереднього та поточного сезонів модель досягла 90% Accuracy, Precision 0.917, Recall 0.900 та F1-score 0.899. У бінарній постановці задачі для 38-го туру найвищий результат показала XGBoost — 100% Accuracy, Precision 1.000, Recall 1.000 та F1-score 1.000 на вибірці з 10 матчів.

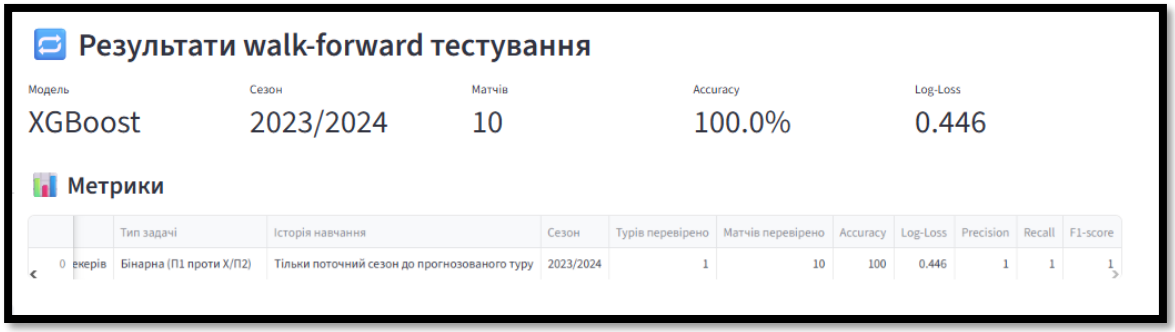
Варто зазначити, що результати на окремих турах не є середньою точністю системи на всьому сезоні. Один тур містить приблизно 10 матчів, тому кожен правильний або неправильний прогноз змінює Accuracy приблизно на 10%. Тому результати 80–100% характеризують успішність моделі на конкретному контрольному турі, а загальна якість моделей оцінювалася за хронологічною схемою 80/20.

Також було реалізовано модуль сценарного моделювання What-If. Він дозволяє користувачеві вручну змінювати показники форми команд і спостерігати за зміною прогнозних імовірностей. Це робить систему не лише інструментом автоматичного прогнозування, а й засобом сценарного аналізу.

Повна порівняльна таблиця метрик

	Алгоритм	Accuracy	Log-Loss	F1-score	Recall	Precision
0	Логістична Регресія	51.4%	1.008	0.48	0.49	0.49
1	Random Forest	52.3%	0.973	0.4	0.44	0.41
2	Gradient Boosting	54.1%	0.971	0.39	0.44	0.35
3	XGBoost	56.6%	0.97	0.41	0.47	0.37
4	Neural Network	54.4%	0.956	0.42	0.46	0.5

Рисунок 1 — Порівняння метрик моделей у мультикласовій задачі прогнозування



Результати walk-forward тестування

Модель	Сезон	Матчів	Accuracy	Log-Loss
XGBoost	2023/2024	10	100.0%	0.446

Метрики

	Тип задачі	Історія навчання	Сезон	Турів перевірено	Матчів перевірено	Accuracy	Log-Loss	Precision	Recall	F1-score
0	екерів Бінарна (П1 проти Х/П2)	Тільки поточний сезон до прогнозованого туру	2023/2024	1	10	100	0.446	1	1	1

Рисунок 2 — Результат прогнозування 38-го туру в бінарній постановці задачі моделлю XGBoost

Висновки

У роботі розроблено інформаційно-аналітичну систему прогнозування результатів футбольних матчів на основі алгоритмів машинного навчання. Система виконує обробку історичних даних, розраховує динамічні ознаки форми команд, формує ELO-рейтинг, навчає моделі машинного навчання та надає користувачеві інтерактивний інтерфейс для аналізу результатів.

У загальному хронологічному тестуванні 80/20 найкращу Accuracy у мультикласовій задачі показала модель XGBoost — 56.6%. Цей результат є співставним із букмекерським baseline на рівні 55.7%, що підтверджує складність прогнозування футбольних матчів у форматі H/D/A.

Практичне прогнозування окремих турів показало, що в бінарній постановці задачі система може досягати вищої точності. Для 30-го туру найкращі моделі показали 80% Accuracy, а для 38-го туру модель XGBoost у бінарній задачі досягла 100% Accuracy на контрольному наборі з 10 матчів. У мультикласовій задачі для 38-го туру Gradient Boosting досягла 90% Accuracy при використанні попереднього та поточного сезонів для навчання.

Отримані результати підтверджують працездатність розробленої системи та доцільність використання ELO, показників поточної форми і різницевих метрик для прогнозування футбольних матчів. Бінарна постановка задачі H/Not_H виявилася корисною для оцінювання ризику втрати очок домашньою командою, тоді як мультикласова задача H/D/A залишається складнішою через необхідність окремого прогнозування нічиєї.

Подальший розвиток системи може включати додавання розширених футбольних метрик, зокрема xG, даних про травми, склади команд, календарну втому, а також використання методів локальної інтерпретованості SHAP і автоматизацію збору даних через спортивні API.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. L. M. Hvattum and H. Arntzen, "Using ELO ratings for match result prediction in association football," International Journal of Forecasting, vol. 26, no. 3, pp. 460-470, 2010.
2. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference, 2016, pp. 785-794.
3. Наука про дані: машинне навчання та інтелектуальний аналіз даних : електронний навчальний посібник комбінованого (локального та мережевого) використання [Електронний ресурс] / В. Б. Мокін, М. В. Дратований – Вінниця : ВНТУ, 2024. – 258 с.
4. M. C. Malamatinos, V. Vrochidou, and G. A. Papakostas, "On Predicting Soccer Outcomes in the Greek League Using Machine Learning," Computers, vol. 11, no. 9, article 133, 2022, doi: 10.3390/computers11090133.
5. Streamlit Documentation [Електронний ресурс]. – Режим доступу: https://docs.streamlit.io/

Зеленцов Олександр Ігорович – студент групи СА-22б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: sasha.zelentsov777@gmail.com

Сергій Олександрович Жуков – канд. техн. наук, доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: sazhukov@vntu.edu.ua

Zelentsov Oleksandr Ihorovych – student of group SA-22b, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: sasha.zelentsov777@gmail.com

Serhii Oleksandrovych Zhukov – Candidate of Technical Sciences, Associate Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@vntu.edu.ua