

ПРОГНОЗИРОВАНИЕ ГОЛОВ В ФУТБОЛЕ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Сулимов Даниил Андреевич

Стажёр-исследователь, научно-учебная лаборатория
искусственного интеллекта для вычислительной
биологии, Высшая школа экономики, г. Москва
dany.sulimov@yandex.ru

AI-DRIVEN PREDICTION FOR GOAL SCORING IN FOOTBALL

D. Sulimov

Summary. This paper is dedicated to the development of a machine learning model for predicting the probability that a football attack will result in a goal. The goal is to provide clubs with a tool to strengthen their strategies and increase their chances of better performance, thus improving the financial performance of the team. To achieve this goal, tasks were set: the collection and processing of data, training a machine learning model to predict the probability of a goal. During the study, a hypothesis was put forward that artificial intelligence can accurately predict the outcome of an attack. To test this hypothesis, various algorithms were used, such as logistic regression, random forest, and gradient boosting. The results of the study showed that the machine learning model has a high predictive ability. The results of the study can be used by clubs to improve their strategies. This can lead to an improvement in the financial performance of the team, as winning games and achieving good results in competitions is crucial for financial success, especially for small clubs with limited transfer income.

Keywords: machine learning, predictive modelling, forecasting, data science in sports.

Аннотация. В данной статье представлено исследование, посвященное разработке модели машинного обучения для прогнозирования вероятности того, что атака в футболе приведет к голу. Цель — предоставление клубам инструмента для укрепления своих стратегий и повышения шансов на лучшую результативность, улучшение финансовых показателей команды. Для достижения цели были поставлены задачи: сбор и обработка данных, обучение модели машинного обучения для прогнозирования вероятности гола. В ходе исследования была выдвинута гипотеза о том, что искусственный интеллект может точно предсказывать исход атаки. Для проверки этой гипотезы были использованы различные алгоритмы, такие как логистическая регрессия, случайный лес и градиентный бустинг. Результаты исследования показали, что модель машинного обучения обладает высокой предиктивной способностью. Результаты исследования могут быть использованы клубами для улучшения их стратегий. Это может привести к улучшению финансовых показателей команды, поскольку победа в играх и достижение хороших результатов в соревнованиях имеет решающее значение для финансового успеха, особенно для небольших клубов с ограниченными доходами от трансферов.

Ключевые слова: машинное обучение, предиктивное моделирование, прогнозирование, анализ данных в спорте.

Введение

В ходе развития анализа данных в спорте, в контексте моделирования сути игры, появились различные подходы. Например, некоторые из них предназначены для оценки трансферной стоимости игрока с использованием анализа действий игроков. Эти подходы охватывают ряд методологий, включая методы регрессионного анализа [12], где рассматриваются основные действия, такие как результативные передачи и голы. Были использованы дополнительные методы, такие как машинное обучение [7], при этом авторы исследовали такие факторы, как рейтинг «плюс-минус». Также были применены подходы глубинного обучения [11, 8, 14], в которых подчеркиваются относительные и абсолютные параметры производительности игроков, такие как успешные подкаты, дриблинг и многое другое. Таким образом, в центре внимания исследований находится вклад игрока в команду. Кроме того, исследование результативности игрока может использоваться в качестве инструмента для получения обратной связи [3, 1] для тренеров, скаутов, менеджмента. Для оценки показате-

лей производительности также были представлены различные исследования. Однако в основном они рассматривают действия, которые напрямую влияют на игру (например, ключевые передачи, голы и т.д.), такие подходы рассматриваются в [2, 6, 10]. Но некоторые действия сами по себе не приводят к забитому голу, однако они вносят большой вклад в успешную атаку команды (например, первый пас, который открывает пространство для атаки, успешный дриблинг и т.д.). Методы оценки подобных действий описаны в [4, 9], открывая Байесовский подход к оценке показателей эффективности футболиста и находя его применение в обнаружении молодых талантов и замене игроков, покинувших клуб.

Данное исследование основано на использовании данных, собранных после игры, и включает оценку каждого действия в рамках атаки, которая завершается ударом по воротам. В исследовании представлен алгоритм, который предсказывает исход атаки команды, находясь в определённый момент времени на поле, который описывается несколькими факторами, перечисленными ниже. Это служит ценным ресурсом для тренеров и для

более широкого понимания тактики и процессов принятия решений в мире футбола.

Разработка предиктивной модели

Данные, используемые для данного исследования, доступны в открытом доступе [15], предоставленных спортивной аналитической компанией StatsBomb. Предоставленный набор данных состоит из более чем тысячи игр, охватывающих национальные и международные соревнования по мужскому и женскому футболу. Данные о каждом матче состоят из 105 характеристик, описывающих каждое событие игры в контексте временной метки, типа действия, продолжительности действия и т.д. Целью прогноза забитых мячей является оценка следующего: $P(score | \bar{s}_k)$ — вероятность забить гол, учитывая, что мяч находится в положении на поле $\bar{s}_k = (x_k, y_k)^T$, где x_k, y_k — координаты нахождения мяча. Каждое действие на футбольном поле рассматривается с точки зрения следующих факторов:

Категориальные признаки:

- Тип действия;
- Выполняется ли действие под прессингом соперника;
- Тип предыдущего действия;
- Положение игры (например, «с углового»: действие является частью игры после подачи углового, «удар по воротам»: действие является частью игры после удара по воротам и т.д.)

Дискретные признаки:

- Координаты действия;
- Количество действий от текущей позиции до конца удара по воротам;
- Временная метка действия;
- Номер действия в рамках одной атаки;

Непрерывные признаки:

- Продолжительность действия;
- Совокупная продолжительность цепочки до действия

Результат каждого рассматриваемого действия попадает в одну из двух категорий: успех (=1) или неуспех (=0). Таким образом, мы сталкиваемся с задачей бинарной классификации. Однако мы имеем дело с проблемой сильно несбалансированных классов. Среди всех действий, включенных в набор данных, только 12 % из них в конечном итоге приводят к успеху. Следовательно, классу 0 присваивается вес 0,12 для решения проблемы несбалансированности классов. Результаты обученных моделей представлены в Таблице 1. В качестве метрик используется стандартный набор метрик бинарной классификации.

Таблица 1.

Метрики разработанного алгоритма

Model	F1-score (weighted)	Precision (weighted)	Recall (weighted)	AUC
Logistic Regression	0.44	0.83	0.37	0.51
Random Forest Classifier	0.66	0.88	0.59	0.79
CatBoost Classifier	0.95	0.95	0.95	0.97

Классификатор CatBoost является наиболее эффективным среди всех рассмотренных алгоритмов, неизменно достигая максимальной производительности по всем показателям оценки. Кривая точности-полноты, представленная на Рисунке 1, показывает, что классификатор CatBoost демонстрирует наиболее значительную площадь под кривой по сравнению с двумя другими алгоритмами. Это указывает на способность алгоритма обнаруживать экземпляры, принадлежащие к классу меньшинства. Визуализация кривой служит убедительным доказательством высокой точности модели в идентификации соответствующих элементов. Используя алгоритм классификатора CatBoost, обученная модель достигает высокой производительности, превосходя свои аналоги с точки зрения точности, полноты и общих возможностей модели. Визуализация кривой точности запоминания убедительно подтверждает эти выводы, подчеркивая полезные качества алгоритма в точном выявлении соответствующих случаев.

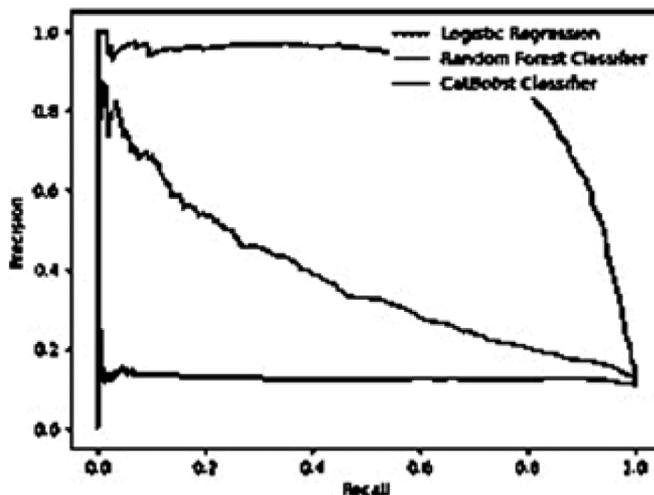


Рис. 1. Кривые точности-полноты классификаторов

Будущая работа

Представленная модель демонстрирует выдающиеся результаты, достигая почти верхней границы каждого предложенного показателя. Однако текущие исследования можно было бы улучшить с нескольких точек зрения. Во-первых, включение данных, происходящий в настоя-

щий момент времени в игре, которые позволяют автоматически отслеживать местоположение всех игроков, расширило бы множество признаков, которые можно было бы учитывать при построении модели. Дополнительные переменные позволили бы более полно моделировать действия футболистов, учитывая действия как атакующей, так и обороняющейся команд. Защитные действия

можно было бы оценить, смоделировав игру как систему «действие-реакция», что позволило бы более целостно оценивать игровые состояния. Эта расширенная модель охватывала бы различные сценарии и обеспечивала более глубокое понимание динамики игры, повышая эффективность прогнозирования.

ЛИТЕРАТУРА

1. Л.В. Андерсен, Дж.В. Фрэнсис и М. Бейтман. Восприятие футбольными тренерами Датской ассоциации анализа результативности. *Международный журнал анализа результативности в спорте*, 22(1):149–173, 2022. URL <https://doi.org/10.1080/24748668.2021.2012040>.
2. Г. Анзер и П. Баур. Модель вероятности забитых голов для бросков, основанная на синхронизированных данных о положении и событиях в футболе. *Границы спорта и активного образа жизни*, 3, 2021 г. URL <https://doi.org/10.3389/fspor.2021.624475>.
3. С. Кронин, Т. Бампурас и П. Миллер. Процессы анализа производительности в спортивной практике высших достижений: исследовательское исследование перспектив спортивного ученого, тренера и спортсмена. *Международный журнал анализа результативности в спорте*, 12: 468–483, 2012. URL <https://doi.org/10.1080/24748668.2012.11868611>.
4. Т. Декроос, Л. Брансен, Дж. Ван Хаарен и Дж. Дэвис. Действия говорят громче, чем цели: оценка действий игроков в футболе. В материалах 25-й Международной конференции ACM SIGKDD по обнаружению знаний и интеллектуальному анализу данных, стр. 1851–1861, Нью-Йорк, штат Нью-Йорк, США, 2019. Ассоциация вычислительной техники. ISBN 9781450362016. URL <https://doi.org/10.1145/3292500.3330758>.
5. Дж. Фернандес, Л. Борн и Д. Сервон. Разложение неизмеримого спорта: структура ожидаемой ценности владения в футболе с глубоким обучением. В исследовательской статье, MIT Sloan, конференция по спортивной аналитике, Бостон, 2019 г.
6. Ф. Гое-Смит, М. Кемпе, Р. Меерхофф и А.П.М. Коэн Лемминк. Не каждый пас может быть результативным: модель на основе данных для измерения эффективности пасов в профессиональных футбольных матчах. *Большие данные*, 7, 2018. URL <https://doi.org/10.1089/big.2018.0067>.
7. Р. Ковач и Л. Тока. Прогнозирование трансферов игроков в маленьком мире футбола, 2022 год. URL: https://doi.org/10.1007/978-3-031-02044-5_4.
8. Г. Лю и О. Шульте. Глубокое обучение с подкреплением в хоккее для контекстно-зависимой оценки игроков. В материалах двадцать седьмой международной совместной конференции по искусственному интеллекту, июль, IJCAI-18., страницы 3442–3448, 2018 г. URL <https://doi.org/10.24963/ijcai.2018/478>.
9. Г. Лю, Ю. Луо, О. Шульте и Т. Харрат. Глубокая футбольная аналитика: изучение функции значения действия для оценки футболистов. *Данные Мин. Знать. Discov.*, 34 (5):1531–1559, 2020. ISSN 1384–5810. URL <https://doi.org/10.1007/s10618-020-00705-9>.
10. П. Люси, А. Бялковски, М. Монфорт, П. Карр и И. Мэтьюз. «Качество против количества»: улучшенное прогнозирование бросков в футболе с использованием стратегических характеристик на основе пространственно-временных данных. В материалах 8-й ежегодной конференции MIT Sloan Sports Analytics (стр. 1–9), 2015 г. I. Г. Макхейл и Б. Холмс. Оценка трансферных сборов профессиональных футболистов с использованием передовых показателей производительности и машинного обучения. *Европейский журнал операционных исследований*, 306(1):389–399, 2023. ISSN 0377-2217. URL <https://doi.org/10.1016/j.ejor.2022.06.033>.
11. Д. Патнаик, Х. Прахарадж, К. Пракаш и К. Самдани. Исследование моделей прогнозирования оценок футболистов путем количественной оценки статистических и экономических характеристик мирового трансферного рынка. На Международной конференции IEEE по системам, вычислениям, автоматизации и сетям (ICSCAN) 2019 г., страницы 1–7, 2019 г. URL <https://doi.org/10.1109/ICSCAN.2019.8878843>.
12. Р. Поли, Р. Бессон и Л. Равенель. Эконометрический подход к оценке трансферных сборов и стоимости профессиональных футболистов. *Экономика*, 10 (1), 2022. ISSN 2227-7099. URL <https://doi.org/10.3390/economies10010004>.
13. П. Пауэр, Х. Руис, К. Вэй и П. Люси. Не все передачи одинаковы: объективно оценивайте риск и выгоду от передач в футболе на основе данных отслеживания. В материалах 23-й Международной конференции ACM SIGKDD по обнаружению знаний и интеллектуальному анализу данных, страницы 1605–1613, 2017 г. ISBN 978-1-4503-4887-4. URL <https://doi.org/10.1145/3097983.3098051>.
14. П. Рахимян, Дж. Ван Хаарен, Т. Абжанова и Л. Тока. Помимо оценки действий: система глубокого обучения с подкреплением для оптимизации решений игроков в футболе. На 16-й ежегодной конференции по спортивной аналитике MIT Sloan. Бостон, Массачусетс, США: Массачусетский технологический институт, стр. 25, 2022 г.
15. Статсбомб. открытые данные, 2023 г. URL <https://github.com/statsbomb/open-data>.

© Сулимов Даниил Андреевич (dany.sulimov@yandex.ru)

Журнал «Современная наука: актуальные проблемы теории и практики»