УДК 004.85

А. М. Макаров
Студент 2 курса магистратуры
«Санкт-Петербургский государственный университет»

В. И. Кияев
Профессор кафедры информатики
«Санкт-Петербургский государственный экономический университет»

ПРИМЕНЕНИЕ НОВЫХ МЕТОДОВ СПОРТИВНОЙ АНАЛИТИКИ В ПОСТРОЕНИИ ПРОГНОЗНЫХ МОДЕЛЕЙ В ИГРОВЫХ ВИДАХ СПОРТА

Аннотация. В статье рассматриваются возможности применения новых методов машинного обучения для вычисления параметров более точных xG-моделей, определяются наиболее важные игровые атрибуты, определяющие рейтинги игроков после и по ходу матча, в соответствии с их более детальными позициями на поле.

Ключевые слова: спортивная аналитика, прогнозные модели, анализ данных, алгоритмы машинного обучения.

Введение

В игровых видах спорта часто случается так, что результат матча не отражает суть происходившего на поле. При нулевой ничьей борьба может быть очень острой, а при большой разнице забитых мячей игра может протекать вяло и неинтересно. Такая ситуация может складываться в любом командном игровом виде спорта – баскетболе, волейболе, гандболе, бейсболе и т. д. На игровом поле постоянно возникают ситуации, которые отражают суммарные усилия игроков – ввод мяча в игру при стандартных положениях, командный прессинг, создание опасного момента, применение искусственного офсайда. Таким образом, формирование ситуации, приводящей к полезным действиям на поле, складывается из согласованных действий нескольких игроков или всей команды.

Из этого следует, что для анализа игры в целом, оценки действий каждого игрока по отдельности и построению прогноза действий команды в будущих играх необходим набор оценочных элементов (показателей, индикаторов) и соответствующих им метрик, с помощью которых можно характеризовать действия каждого игрока в отдельности (вычислить «коэффициент полезности» игрока) и эффективность действий всей команды. Такой оценочный механизм чрезвычайно полезен при разборе предыдущих игр, составлении планов подготовки команды на будущие игры текущего сезона, а также построении прогнозов успешности своей команды и анализа успешности команд соперников.

Помимо того, что существуют сайты, на которых футбольные эксперты выставляют каждому игроку рейтинговую оценку после игры, за базовую основу можно взять часто применяемую метрику Expected Goals Model (xG). По этой метрике, в частности, каждому удару, нанесенному в сторону ворот, присваивается вероятность того, что он закончится голом. Сейчас существует несколько xG-моделей, выложенных в открытый доступ, однако, мы не знаем насколько адекватны результаты, полученные с помощью этих моделей, и не можем провести верификацию моделей на практике, так как самого кода, по которому можно верифицировать конкретную модель xG, в открытом доступе нет.

Исторически методы спортивной аналитики начали применяться в бейсболе. Причиной этому послужили общедоступность статистических данных и обширное сообщество, заинтересованное в анализе этой игры. Основоположником бейсбольной статистики считается Уильям Джеймс, который вёл термин «сабметрика» (submetrics).

Изложенные им методы получили широкое применение благодаря книге Михаэля Льюиса «Moneyball», которая считается мировым феноменом в области спортивной журналистики в XXI веке [5].

Публикация C. Reep и B. Benjamin на эту тему является одной из основных работ в области аналитического анализа эффективности игроков в футболе и считается ориентиром в анализе эффективности матчей в футболе [7]. Исследователи задавались вопросом: игра в пас или игра забросами («possession play» or «direct play») наиболее эффективна при достижении целей в футболе. Основываясь на этой работе, смешанные результаты получил Collet который утверждал, что корректное владение мячом влияет на результат матча лишь в случае команд, входящих в элиту мирового футбола [3].

Отметим, что в большинстве исследований использовался одномерный подход, состоявший в выявлении одного тактического фактора, характеризующего разницу между успешными и безуспешными группами выступлений. Однако, с течением времени стало понятно, что, из-за сложности анализа эффективности футбольного матча и возможного взаимодействия нескольких тактических факторов, необходимо использовать многомерный статистический анализ [4].

Исследования также указывают на то, что выводы, сделанные из нескольких взаимосвязанных исследований, могут подвергаться сомнению на основании недостаточного размера выборки. Также ключевым вопросом в анализе эффективности является определение того, какие атрибуты (факторы) следует рассматривать как ключевые показатели эффективности. В большинстве случаев эти атрибуты построены на основе экспертного понимания игры. В работе [9] несколько аналитиков провели совместный анализ, чтобы определить ключевые показатели эффективности для каждой позиции игрока в футбольном матче.

В последнее время также предпринимались попытки привлечения методов искусственного интеллекта к проблеме прогнозирования. Например, A. C. Constantinou с соавторами использовал байесовскую сетевую модель, чтобы предсказать результаты матчей [2]. H. Rue и Øyvind Slavensen использовали байесовский подход в сочетании с Марковскими цепями и методом Монте-Карло [8]. Эти модели достаточно сложны, используют многие предположения, требуют больших статистических выборок и не всегда могут быть легко и однозначно интерпретированы. Нейронные сети использовались для прогнозирования в нескольких видах спорта, включая американский футбол [6].

На данный момент существует несколько общедоступных ресурсов в Интернете, которые предоставляют xG-статистику для футбольных Топ-чемпионатов (АПЛ, Ла Лига, Бундеслига), Michael Caley [12] и сайт understat.com [14], однако саму модель в открытый доступ никто из них не выкладывал.

Стоит также отметить, что в футболе существуют и другие метрики для оценки действий игроков, например, метрики результативности продвижения мяча Packing\Impect [11]. Packing – это сумма соперников, оказывающихся за линией мяча либо в результате передачи вперёд, либо после удачной обводки. Impect – разновидность показателя Packing, учитывающая исключительно «отрезанных» защитников – эта статистика хорошо коррелирует с результатами матчей. До недавнего времени считалось, что в 66% случаев команда победит если она переиграла соперника по показателю Impect. Главным минусом подобных статистик является то, что они пока ещё «молоды» и не проверены на достаточно больших выборках матчей.

Гипотеза

В нашей предыдущей работе мы разработали программный сервис для построения xG-модели, а также проанализировали наиболее важные игровые атрибуты, определяющие рейтинги игроков, выставленные футбольными экспертами [1]. Мы можем

улучшить данную xG-модель, используя новые данные, новые методы машинного обучения, а также расширить структурное представление действий групп игроков, сделав их более детализированными.

Методы

В нашей работе мы использовали методы машинного обучения (Machine Learning – ML) для анализа данных футбольного матча для достижения двух целей. Мы построили новую модель формирования xG-метрик, улучшили результаты, показанные в работе [1]. Также, мы определили наиболее важные атрибуты производительности групп игроков, которые определяют их рейтинги, при этом расширив нашу задачу классификации.

Результаты и обсуждение

Для построения обновленной xG-модели мы выбрали язык программирования Python (последней версии 3.8.2), библиотеки scikit-learn, NumPy, SciPy и Pandas. В качестве платформы написания кода мы выбрали среду Jupyter Notebook. Для автоматизации процесса получения информации с сайта whoscored.com [16] об ударах мяча из всех футбольных матчей, сыгранных в рамках турнира, мы использовали инструмент Selenium WebDriver для автоматизации действий с Web-браузером [13].

Базовый алгоритм нашей программы выглядит так же, как и в предыдущей работе [1], однако мы пользовались обновленными алгоритмами ML, добавили в наш dataset информацию за сезоны 2017/2018 и 2018/2019 – в итоге в совокупной выборке оказалась информация обо всех ударах по мячу из топ-5 лиг (АПЛ, Ла Лига и Бундеслига, Лига 1, Серия А) за сезоны 2010-2019, всего 427543 удара. Добавили также несколько признаков (помимо уже добавленных в работе [1]) – расстояние до ворот и угол, под которым они видны из точки удара, степень блокировки удара (какую площадь обзора ворот закрывают защитники, блокирующие удар, и какая площадь ворот в зоне досягаемости вратаря) и число защитников команды оказавшихся вообще за линией удара мяча.

Пошаговый алгоритм разработанного программного сервиса выглядит следующим образом:

1. Формируем функцию, которая извлекает информацию о таких событиях матча, как: касание мяча, пас, удар по мячу (и т. д.) в конкретном матче, как это сделано в работе [3].
2. Из этого массива данных получаем всё, что касается ударов по мячу.
3. Применяем функцию, которая по ссылке на турнир получает ссылки на все матчи, сыгранные в нём – для этого используем данные с ресурса Selenium [13].
4. Далее с помощью созданного парсера загружаем извлечённую информацию об ударах по мячу и обучаем логистическую регрессию, обученную модель сохраняем в отдельный файл.
5. Переходим к формированию оценок в соответствии с выбранной xG-метрикой – для этого формируем функцию, которая с помощью обученной модели вычисляет xG-метрику для каждого момента в матче.

Как уже было сказано ранее, в предыдущей работе мы использовали информацию об ударах из топ-3 лиг (АПЛ, Ла Лига и Бундеслига) за сезоны 2010-2017. В более ранней работе[1] величина средней квадратичной ошибки, вычисленной по разности значений реальных голов и xG была равна RMSE = ± 0.29, для новой модели на большей выборке было получено RMSE = ±0.25. Это говорит о том, что новая модель xG метрики дает более точную оценку чем используемая нами ранее.

Также, используя разработанный сервис, можем посмотреть, например, на соотношение голов и созданных моментов (xG/goal), в этот раз для таблицы РФПЛ сезона 2018/2019 (табл. 1). В таблице использованы слелующие параметры: team_id- название команды, is_goal- число реальных голов, которое забила команда, xG- ожидаемое число голов, вычисленное нами.

Таблица 1. Оценка игровых моментов, созданных командами РФПЛ сезона 2018/2019

team_id	is_goal	xG	xG/goal
Зенит	57	49.21	0.86
Краснодар	55	51.19	0.93
Локомотив	45	42.31	0.94
ЦСКА	46	44.52	0.96
Спартак	36	39.07	1.09
Арсенал	40	32.45	0.81
Оренбург	39	32.04	0.82
Ахмат	28	25.87	0.92
Ростов	25	27.59	1.10
Урал	33	31.25	0.95
Рубин	24	27.50	1.14
Динамо	28	34.32	1.23
Крылья Советов	25	29.57	1.18
Уфа	24	25.21	1.05
Анжи	13	17.65	1.35
Енисей	24	29.61	1.23

Метрику xG/goal можно использовать, чтобы сделать выводы о навыке реализации игровых моментов игроками команды — чем меньше этот показатель, тем меньше игровых моментов, описываемых величиной xG-метрики, требуется команде на реализацию одного гола, то есть она имеет лучшую реализацию с меньшими игровыми усилиями. Как видно из результатов, приведенных в таблице, наилучший навык реализации моментов по метрике xG/goal имеют команды «Зенит», «Арсенал» и «Оренбург», наихудший – «Анжи» и «Енисей», замыкающие турнирную таблицу

Можем посмотреть также на различные метрики, оценивающие атакующие действия нападающих команды «Зенит», в том же сезоне 2018/2019 в рамках чемпионата (табл. 2).

Таблица 2. Оценка игровых моментов, созданных игроками команды «Зенит» за сезон 2018/2019

player_name	games	min	is_goal	xG	xG/goal	xG/90
Сердар Азмун	12	1025	9	10.35	1.15	0.91
Артем Дзюба	27	2366	8	8.56	1.07	0.33
Себастьян Дриусси	27	1853	11	9.52	0.86	0.46
Эмилиано Ригони	11	726	3	2.10	0.7	0.24

Тут довольно логично помимо метрики xG/goal, ввести метрику xG/90- количество созданных моментов (xG) за 90 мин (ровно столько длится полный футбольный матч без дополнительного времени), так как игроки проводят на поле не равное число минут (они могут выходить с замены, быть травмированными, пропускать матчи из-за

восстановления и т. д.). Значения метрики xG/90 мы будем вычислять по формуле xG/(min/90), где min- количество минут, проведенных игроком на поле.

В данной таблице: player_name – имя игрока, games – число игр сезона, в которых игроки выходили на поле, min – количество минут, проведенных игроком на поле, is_goal – реальное число голов, забитое игроками по итогу сезона, xG – ожидаемое число голов, вычисленное нами, xG/goal, xG/90 – метрики, описанные нами выше.

Из результатов, приведённых в таблице 2, видно, что игрок Сердар Азмун провел всего 12 игр в сезоне, при этом создал больше всех опасных моментов (xG)/90 (за 90 мин), а игрок Эмилиано Ригони имеет лучшую результативность, так как ему на гол требуется меньше всего моментов (xG/goal).

Ещё одной целью нашей работы, в продолжение работы [1], является проведение анализа наиболее важных игровых атрибутов**,** определяющие рейтинги игроков, выставленные футбольными экспертами, разделив их на более детальные группы. Мы разделили всех игроков на 8 групп (кластеров) в соответствии с их амплуа – на центральных и фланговых защитников, на центральных опорных и центральных атакующих полузащитников, на центральных полузащитников, на фланговых и центральных нападающих и вратарей. Далее, используя алгоритмы ML, для каждого кластера мы решали задачу классификации для набора данных, содержащего некоторую информацию об игровых моментах матчей, так, чтобы они максимально точно определяли рейтинги игроков по итогу матчей, которые им присваивают эксперты с сайта whoscored.com.

В нашей работе [1] мы выбрали наиболее эффективные алгоритмы ML для нашей задачи с помощью программного приложения Weka Toolkit [15]. Теперь, с выходом версии Weka-3, у нас появилась возможность использовать Auto-Weka – эта функция автоматически подбирает алгоритм и параметры, лучше всего классифицирующие параметры задачи. Как и в прошлый раз мы пользовались DataSet, предоставленным OPTA [10] за сезон английской премьер-лиги 2011-2012 годов, так как новых, настолько обширных данных не появилось, а также, чтобы была возможность сравнить полученные результаты.

Для каждого набора данных мы использовали 10-кратную перекрестную проверку, чтобы избежать «перекоса» в сторону одного из атрибутов в процессе обучения и также из-за отсутствия достаточного количества исходных данных для разделения на тренировочные, валидационные и тестовые наборы.

Таблица 3. Результат работы алгоритмов для всех позиций

Position	Algorithm	CC	MAE	BTime(s)	TTime(s)
st	SMOreg	0.9658	0.1675	75.0720	448.9650
lf/rf	SMOreg	0.9646	0.1659	72.0060	520.9650
cam	SMOreg	0.9504	0.1829	237.5800	2837.2170
cm	GaussianProcesses	0.8530	0.1711	615.9230	4815.6860
cdm	GaussianProcesses	0.8515	0.2095	133.5630	1074.4590
lb/rb	SMOreg	0.8484	0.2766	17.0750	150.9360
cb	LMT	0.8515	0.2728	7.4000	61.5210
gk	LMT	0.8815	0.2695	12.5630	97.4590

В таблице 3 мы можем видеть результаты работы восьми наиболее оптимальных алгоритмов, которые были подобраны с помощью Auto-Weka из 107 доступных алгоритмов. В таблице показаны следующие позиции: cfd- центральный нападающий, lf/rf- левый/правый фланговые нападающие (сюда же мы отнесли и фланговых полузащитников), cam- центральный атакующий полузащитник, cm- центральный полузащитник, cdm- центральный опорный полузащитник, lb/rb- левый/правый крайние защитники, cb-

центральный защитник, gk- вратарь. В этой таблице CC означает коэффициент корреляции (рейтинга игрока и рейтинга, который получит модель на основе данного алгоритма), MAE – среднюю абсолютную ошибку, а BTime и TTime – время в секундах построения и обучения модели, соответственно.

Алгоритм functions.SMOreg использует метод опорных векторов, алгоритм LMT – метод на основе деревьев с функциями логистической регрессии на листьях, алгоритм GaussianProcesses – метод на основе Гауссовских процессов.

Далее, для всех восьми кластеров амплуа игроков, используя результаты из таблицы 3, мы выбрали 20 наиболее важных *оценочных игровых атрибутов*, входящих в модели с наибольшим весом (с помощью этих атрибутов можно достаточной точно оценивать рейтинг игрока как по результатам конкретной игры или выбранной совокупности игр – например, по результатам конкретного сезона).

В качестве примера в таблице 4 приведены эти атрибуты для центральных и для фланговых нападающих, соответственно. Названия атрибутов приведены в английской транскрипции, в соответствии с данными, предоставленными ресурсом OPTA – как это принято в подобных моделях. Каждый из этих атрибутов описывает определенную игровую ситуацию, которая может привести к голевому моменту.

Таблица 4. Двадцать наиболее важных игровых атрибутов: для кластера центральные нападающие (слева) и кластера фланговые нападающие (справа)

Top 20 st
Successful Crosses in the Air
Leading to Goal Errors
Penalties Scored
Last Man Tackle Assists
Penalties Not Scored
Throws leaded to attempt
Headed Goals
Corners in the air
Foul
Goals from Set Play
Own Goal
Error leading to Attempt
Red Cards
Won Penalty
Foot Goals
Other Goals
Open Play
Goals from Corners
Goals from Outside Box
Foul Won Penalty

Top 20 lf/rf
Direct Free-kick Goals
Last Man Tackle Assists
Penalties Not Scored
Goals from Throws
Fouls Won in Danger Area including penalties
Other Goals
Headed Goals
Corners in the air
Foul
Goals from Set Play
Successful Crosses in the Air
Crosses leaded to attempt
Own Goal
Error leaded to attempt
Red Cards
Won Penalty
Goals from Outside Box
Foul Won Penalty
Open Play
Goals from Corners

Аналогичные результаты были получены и для остальных кластеров амплуа игроков. Если провести сравнение между полученными результатами, например, для центральных форвардов и фланговых нападающих, можно заметить, что в случае фланговых более важными атрибутами являются: кроссы, вбрасывания, навесы с угловых, а также

заработанные пенальти. Это вполне соответствует реальному положению вещей в современном футболе, так что, можно сделать вывод о том, что наше разделение игроков на более детальные кластеры является обоснованным.

Выполненное исследование и предложенная новая модель формирования xG-метрик позволяют сказать, что предложенные характеристики оценки игровых моментов, создаваемые игроками команд, можно использовать для назначения персональных рейтингов игрокам, совокупных рейтингов команд и прогнозов возможных результатов команд на предстоящие матчи. Полученные результаты позволяют также по-новому оценивать эффективность игроков указанных выше амплуа по ходу сезона и строить прогнозы их результативности в предстоящих играх.

Список литературы

1. Макаров А. М., Кияев В. И. Методы спортивной аналитики в прогнозных моделях // Сборник научных статей по материалам Международной научно-практической конференции "Конвергенция цифровых и материальных миров: экономика, технологии, образование". Санкт-Петербург, 21-22 июня 2018 года. СПб. : Изд-во СПбГЭУ, 2018. — с. 171-180.
2. A. C. Constantinou N. E. Fenton, Neil M. A bayesian network model for forecasting association football match outcomes. — Knowledge-Based Systems, 2012. — P. 322–339.
3. Collet C. The possession game: a comparative analysis of ball retention and team success in european and international football, 2007-2010. — Journal of Sports Sciences, 2012. — P. 1–14.
4. Hughes M., Franks I. Analysis of passing sequences, shots and goals in soccer. — Journal of Sports Sciences, 2005. — P. 509–514.
5. Lewis M. Moneyball: The Art of Winning an Unfair Game. — W.W. Norton Company Inc, 2003.
6. Purucker M. C. Neural network quarterbacking.— IEEE Potentials, 1996. — P. 9–17.
7. Reep C., Benjamin B. Skill and chance in association football.— Journal of Royal Statistical Society, vol. January, no. 1, 1968— P. 581–586.
8. Rue H., Øyvind Slavensen. Predictive and retrospective analysis of soccer matches in a league.— Technical report, Department of Mathematical Sciences, NTNU, 1998. — P. 322–339.
9. Tenga A. Reliability and validity of match performance analysis in soccer: a multi- dimensional qualitative evaluation of opponent interaction. — Journal of Sports Sciences, 2010. — P. 22–34
10. OPTA [Электронный ресурс] // OPTA [Сайт]. [2018]. URL: https://www.optasports.com (дата обращения: 24.03.2020).
11. Packing/impact [Электронный ресурс] // pspfrench [Сайт]. [2019]. URL: https://pspfrench.com/blog/packing (дата обращения: 10.04.2020).
12. Premier League Projections and New Expected Goals [Электронный ресурс] // cartilagefreecaptain [Сайт]. [2019]. URL: hhttps://cartilagefreecaptain.sbnation.com/2015/10/19/ 9295905/premier-league-projections-and-new-expected-goals (дата обращения: 13.04.2020).
13. Selenium [Электронный ресурс] // Selenium [Сайт]. [2017]. URL: hhttps://www.seleniumhq.org (дата обращения: 9.04.2020).
14. Understat [Электронный ресурс] // Understat [Сайт]. [2018]. URL: https://understat.com (дата обращения: 10.03.2020).
15. Weka classifiers sourceforge [Электронный ресурс] // Weka [Сайт]. [2020]. URL: https://weka.sourceforge.io/doc.dev/weka/classifiers (дата обращения: 11.03.2020).
16. Whoscored.com [Электронный ресурс] // Whoscored [Сайт]. [2018]. URL: https://www.whoscored.com (дата обращения: 10.03.2020).