МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ

Кафедра функционального анализа и аналитической экономики

Анализ и визуализация футбольных данных. Математические методы в футболе

Курсовая работа

Шульжика Дмитрия Николаевича
студента 3-го курса
специальности 1-31 03 01-01
«Математика
(научно-производственная деятельность)»
Научный руководитель:
доцент, кандидат физ.-мат. наук Е. М. Радыно

Минск, 2022

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ **3**

1 Элементы визуализации **4**

1.1 Визуализация передач 4

1.1.1 Pass map 4

1.1.2 Heat map 5

1.1.3 Passing networks 6

1.1.4 Диаграмма Вороного 7

1.2 Удары, графики xG 8

1.3 Скаутинг 10

2 Математические методы **14**

2.1 Модель ожидаемых голов 14

2.2 Цепи Маркова. Оценка эффективности 17

ЗАКЛЮЧЕНИЕ **21**

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ **22**

ВВЕДЕНИЕ

Анализ данных с каждым годом становится всё более неотъемлемой частью футбола. Так, довольно весомый вклад в победы «Ливерпуля» в 2018-2020 годах сыграли не только тренерское мастерство и новаторство Юргена Клоппа, но и отличная работа Data Science отдела под руководством Иана Грэма.

Классический подход, когда матч воспринимается через призму видения тренера или скаута, зачастую упускает многие вещи, происходящие на поле и является всё же субъективной оценкой. Анализ данных же призван дополнить эту оценку и дать объективные выводы и по каждому игроку, и по командным действиям в целом.

Многие команды не выносят на широкую публику информацию об устройстве своих отделов по Data Science. Например, в бундеслиге Германии играет 18 команд. И мне известны всего 6 из них, в которых есть хотя бы один Data Scientist. В Английской Премьер лиге (АПЛ) и Чемпионшипе (2 главные лиги Англии) такие отделы есть практически у всех.

Что касается ситуации в Беларуси, то говорить о неком институте Data Science или отраслевых стандартах аналитики в белорусском футболе пока не приходится. Компании, которые собирают данные, пока не думают о чемпионатах уровня белорусского, а методы компьютерного зрения дают большую погрешность в данных, что не позволяет достаточно точно анализировать информацию. Тем не менее, Data Science и анализ данных – это одна из самых горячих и бурно развивающихся индустрий 21 века, инструменты которой в большей степени применяются в бизнесе. Футбол – это тоже своего рода бизнес, и не маленький. Поэтому клубы безусловно будут смотреть в сторону современных подходов к анализу данных, накапливать экспертизу, выстраивать бизнес-процессы в этом направлении.

В данной работе я хотел бы показать как визуализируются футбольные данные, и что может говорить о футбольном матче данная визуализация. Также хотелось бы показать математическую сторону вопроса и привести несколько устоявшихся математических моделей.

ГЛАВА 1

Элементы визуализации

1.1 Визуализация передач

Одним из самых частых действий на футбольном поле является передача мяча от одного игрока другому. В последнее время все больше команд стараются больше владеть мячом и как можно быстрее передвигать его по футбольному полю, чтобы заставить соперника бегать за ним, тем самым контролируя ситуацию. "Пока мяч у нас — нам не забьют"— цитата Валерия Лобановского, одного из основоположников данного стиля игры.

В связи с этими возникают вопросы. Как команда двиагает мяч? Какой футболист отдал больше всего передач? В каких зонах команда имеет наибольшее преимущество?

1.1.1 Pass map

Первым инструментом, с помощью которого можно ответить на поставленные вопросы, является карта передач. Выглядит она следующим образом:

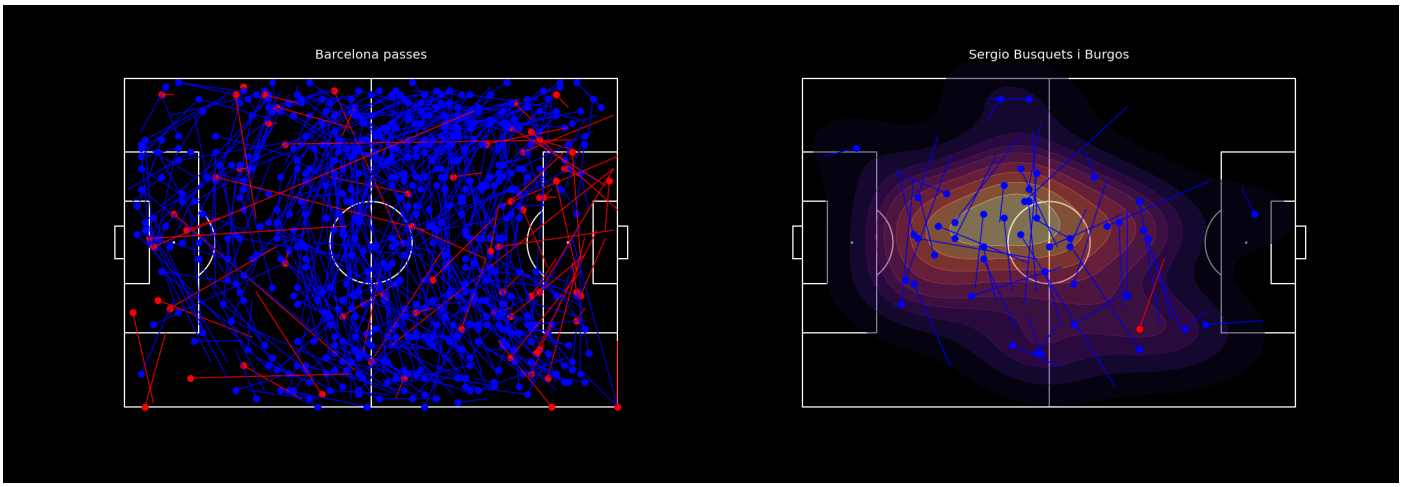

Рисунок 1.1 Карта передач Барселоны Рисунок 1.2 Карта передач Бускетса

На рисунке 1.1 отображены передачи футбольного клуба Барселона(Барселона-Реал Сосьедад, 2018 год). Точные передачи отображены синим цветом, неточ-

ные — красным. По данной карте можно сказать, что большая часть передач сосредоточена в центре поля или же у ворот соперника, что свидетельствует о высоком уровне продвижения мяча к чужой штрафной площади. Также видно, что большая часть передач сосредоточена на левом фланге. Это может говорить о том, что акцент в атаке сделан на него и в дальнейших матчах соперникам стоит акцентировать на этом внимание.

На рисунке 1.2 представлена та же карта передач, но данные берутся по отдельному игроку(в качестве примера взят Серхио Бускетс — один из лучших распасовщиков в мировом футболе). Здесь можно оценить влияние игрока на игру команды. Посмотреть какой сложности передачи он совершает, и определить основную зону действий.

1.1.2 Heat map

Похожим инструментом анализа передач и характера владения в целом является тепловая карта: Она подтвержает гипотезу о том, что атакующие дей-

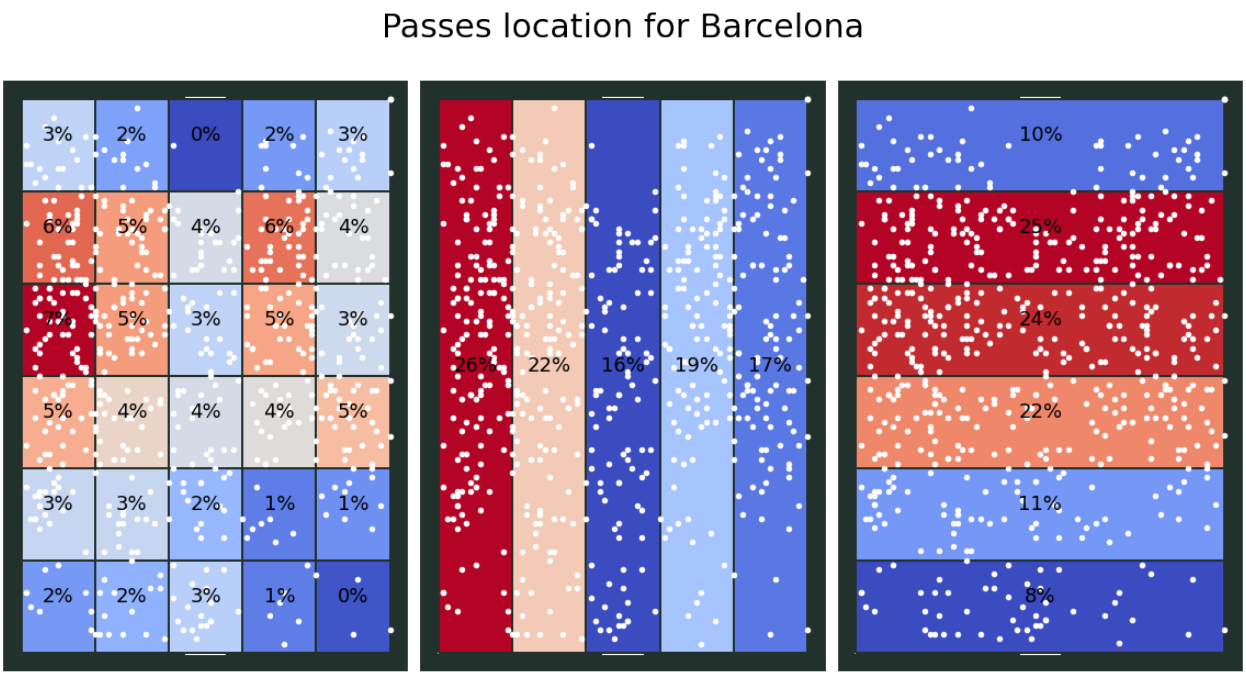

Рисунок 1.3 Тепловая карта Барселоны

ствия команды сконцентрированы на левом фланге, и, действительно, большая

часть передач приходится на часть поля у штрафной площади соперника, что свидетельствует о высоком уровне доминирования.

Тепловая карта, в отличие от карты передач, дает больше количественных сведений о передачах. На ней четко можно увидеть самые активные зоны владения и определить характер матча, однако если нужно узнать примерную сложность и расстояние передачи, лучше воспользоваться обычной.

1.1.3 Passing networks

Самым настоящим трендом в последние годы стали, так называемые, "сети" передач. На рисунке 1.4 представлена "сеть" все той же Барселоны(матч тоже остался неизменным).

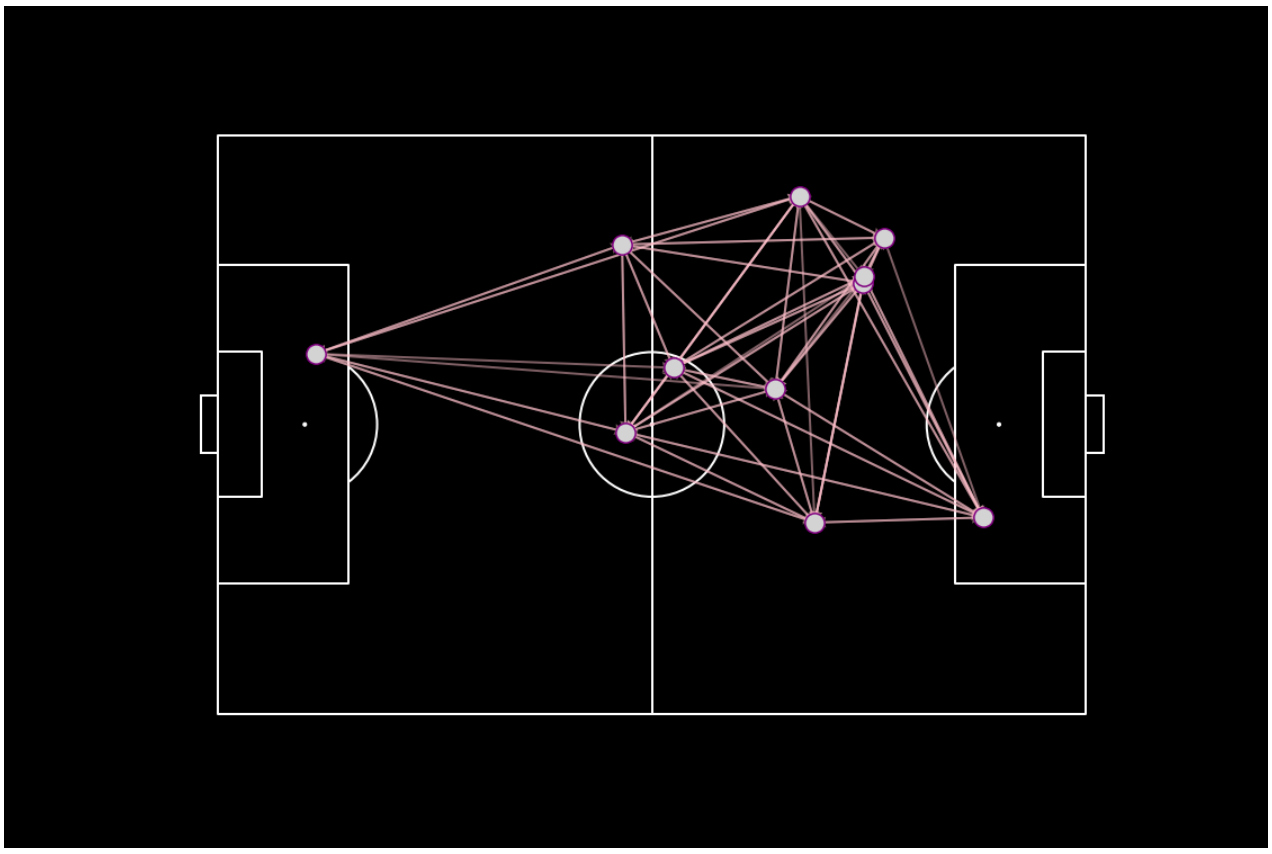

Рисунок 1.4 "Сеть" передач Барселоны

Точками обозначены усредненная позиция футболиста на поле на протяжении всего матча, линии же отвечают за передачи. Чем линия жирнее, тем больше передач в данном направлении было сделано.

По данной сети можно много рассказать о структуре владения и основных принципах игры команды с мячом. Сразу стоит обратить внимание на высокие позиции центральных защитников, что говорит о заточенности команды на атаку, однако соперник может пользоваться пространством за спинами, и совершать рывки в данную зону, что может приводить к опасным моментам. Также можно сразу заметить и схему игры, в которой вышла команда на данный матч. Можно сказать о высокой позиции крайних защитников, и о низкой позиции опорного полузащитника(на данной "сети" между двумя центральными защитниками). Видно, что на левом фланге зачастую создается перегруз в виде четырех футболистов, взаимодействующих между собой(еще одно подтверждение гипотезы о заточенности на левый фланг).

Обращая внимания не на расположение точек, а на толщину линий, обычно делают вывод о взаимодействии между игроками. Чем толще линия, тем больше между ними передач, следственно, больше взаимодействия. Так можно выявлять различные связки, которые в будущем соперники будут стараться нейтрализовать.

1.1.4 Диаграмма Вороного

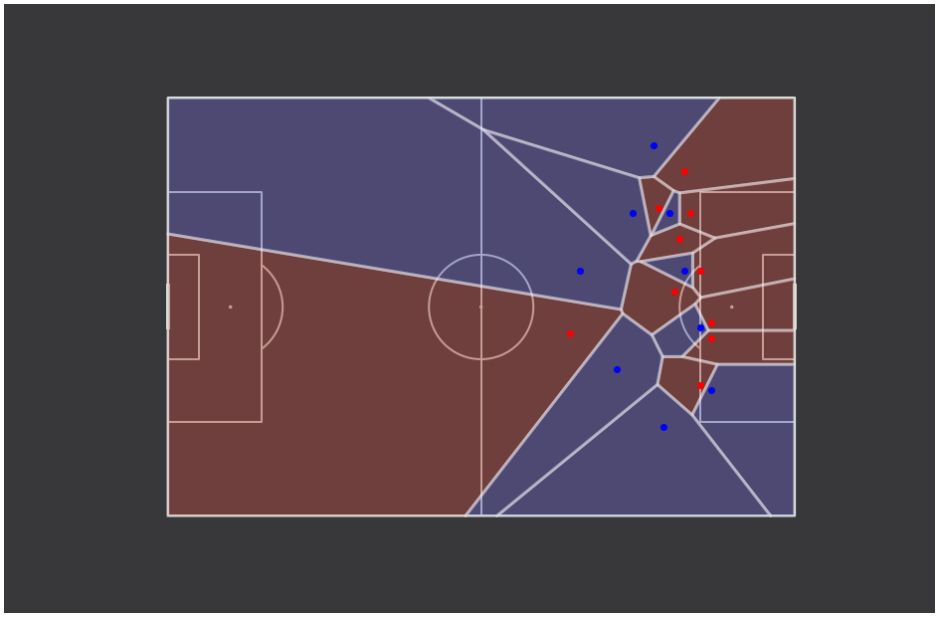

Рисунок 1.5 Диаграмма Вороного

Диаграмма Вороного конечного множества точек S на плоскости представляет такое разбиение плоскости, при котором каждая область этого разбиения образует множество точек, более близких к одному из элементов множества S, чем к любому другому элементу множества. В футболе это применяется редко, однако небольшой интерес все-таки вызывает.

Рассмотрим пример все той же Барселоны, только в матче с другим соперником(диаграмма представлена на рисунке 1.5). Диаграмма делается в какой-то конкретный момент времени. По ней можно судить какие зоны контролирует команда, и как обороняется соперник. Видно, что обороняющаяся команда контролирует зону своих ворот, однако обороняется достаточно глубоко(почти все футболисты защищают штрафную площадь) Большой красный четырехугольник слева предупреждает, что в случае потери мяча, обороняющейся команде стоит просто доставить туда мяч, чтобы создать опасный момент.

1.2 Удары, графики xG

Главная цель футбольной команды на матч—забить больше голов, чем соперник. Чтобы забивать голы, нужно создавать опасные моменты и бить по воротам. Однако удары бывают разные, поэтому оценивать шансы забить гол опираясь на количество ударов будет неправильно. Для этого был придуман более объективный показатель, который называется xG(expected goals), его описание будет дано в следующей главе, сейчас же нужно знать, что xG — это вероятность забить гол отдельным ударом(в одной атаке не может быть больше 1).

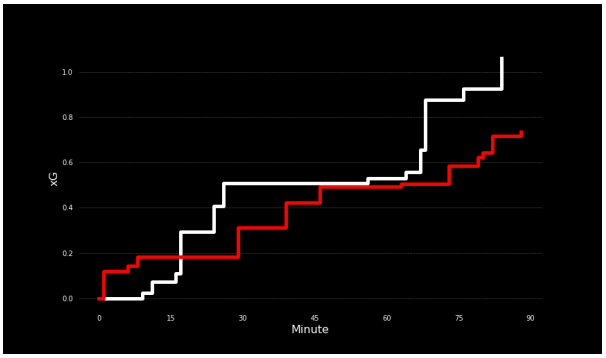

Рисунок 1.6 График набора xG

Обычно, чтобы определить какая команда была ближе к победе, в конце матча просто сравнивают суммарный xG каждой из команд. На рисунке 1.6 показан график набора xG с того же матча(красная линия — Барселона, белая — Реал Сосьедад). С помощью таких графиков можно наблюдать, на каких отрезках матча одна команда дейстовала эфективнее другой, а также сказать о справедливости исхода матча.

Иногда возникает интерес узнать не только на сколько ожидаемых голов наиграла команда, но и посмотреть, откуда наносились удары. Тогда можно посмотреть следующую карту:

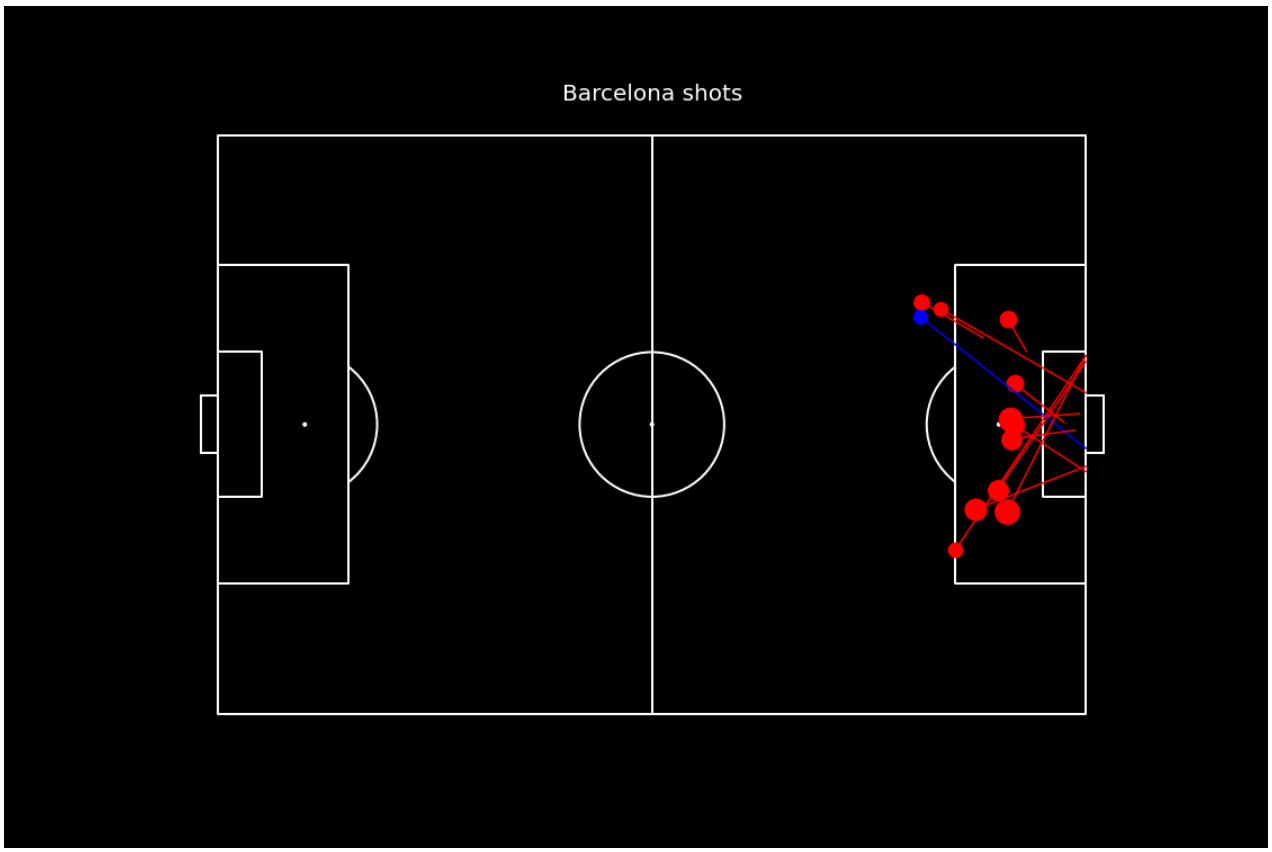

Рисунок 1.7 Карта ударов Барселоны

На карте показаны позиция, с которых наносились удары, направление ударов, а также можно узнать по размеру круга опасность данного удара(чем толще круг, тем больше xG весил удар). Стоит обратить внимание, что синим цветом показан удар, с которого был забит гол. Видно, что круг данного удара не самый большой, отсюда можно сделать вывод, что везение в футболе играет немаловажную роль.

Также показатель ожидаемых голов применяется на более длинные дистанции, к примеру, на сезон. Иногда результаты команды на протяжении таких дистанций могут сильно отличаться от реального качества игры. Команде может сопутствовать удача по ходу сезона или же наоборот(оверперфморминг и андерперформинг). С помощью xG можно правильнее оценить работу тренера и команды в целом, и принять решение о продлении или расторжении контракта с ним. На рисунке ниже представлен график созданного и допущенного xG по сезону.

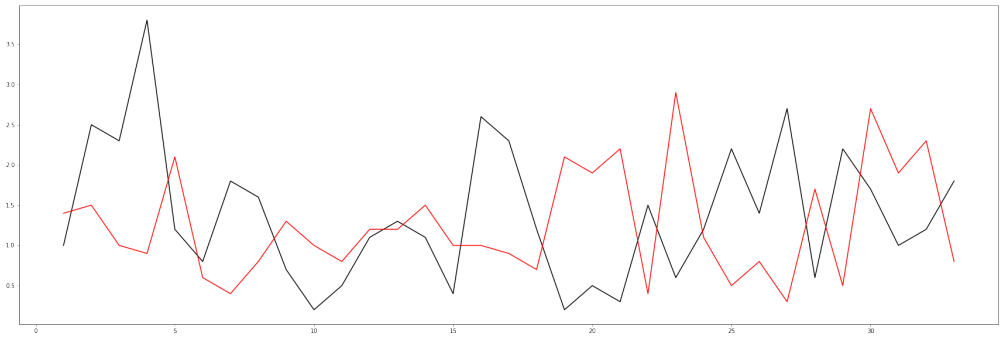

Рисунок 1.8 Тоттенхем, сезон 2020-2021

В качестве примера был взят английский клуб Тоттенхем в сезоне 2020-2021, черная линия — созданный xG, красная — допущенный. Можно заметить, что в первой половине сезона команда выступала неплохо, однако во второй начался спад. Это может быть связано с травмами основных игроков, сложностью матчей, усталостью и так далее. Проанализировав всю ситуацию, руководство клуба может принять наилучшее решения для будущего клуба.

1.3 Скаутинг

Многие футбольные клубы живут за счет грамотной селекции. Обычно они покупают футболистов из лиг более низкого уровня, делают из них звезд и продают за более крупные суммы в другие клубы. Поэтому в селекции также активно используют данные.

Банальным примером может служить задача описать отдельного футболиста. Обычно этим занимаются футбольные скауты, но так как человек не всегда

может объективно оценить ситуацию и все учесть, на помощь приходят цифры. Первым инструментом могут служить круговые диаграммы.

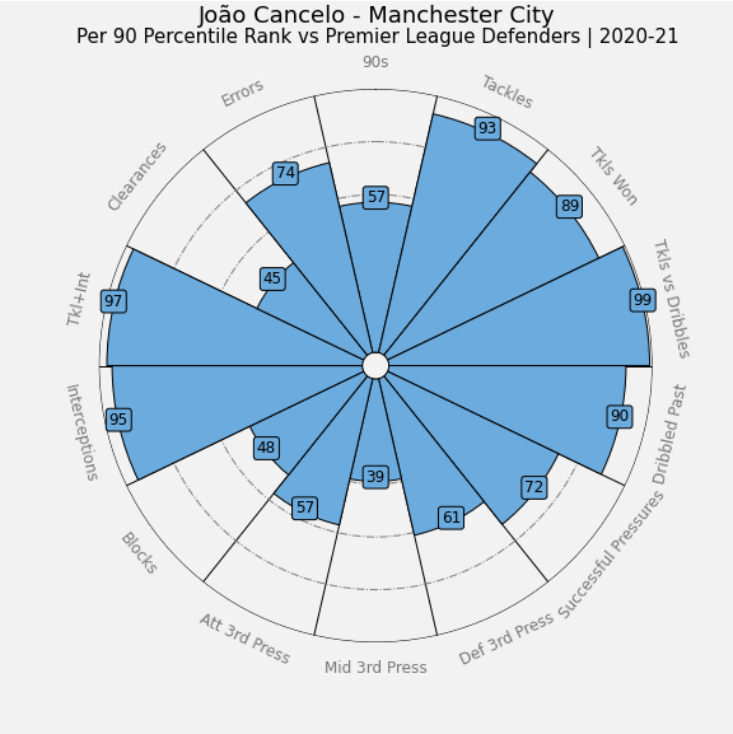

Рисунок 1.9 Оценка Жоао Кансело

На рисунке 1.9 представлен, как говорят, "pizza plot" по Жоао Кансело(крайний защитник). Параметры оценивания обычно выбираются в зависимости от позиции футболиста. Из особенностей, на диаграмме можно отметить высокий процент отборов(tackles), перехватов(interceptions). Последнее свидетельствует о хорошем чтении игры. Также бросается в глаза показатель прессинг действий, процент успешных в атакующей трети и в оборонительной почти одинаковы, что может говорить об установке высоко встречать соперника.

Зачастую бывает необходимо заменить проданного футболиста другим, при этом не сильно изменяя стиль игры команды. Одинаковых игроков не существует, однако можно найти максимально близких по стилю игры. Тут также на помощь приходят круговые диаграммы, но они имеют другой вид. Сравнивать исключительно цифры довольно долго и нудно, поэтому обычно одну диаграмму накладывают на другую и ищут игроков, многоугольники которых максимально близки друг к другу. Выглядит это следующим образом:

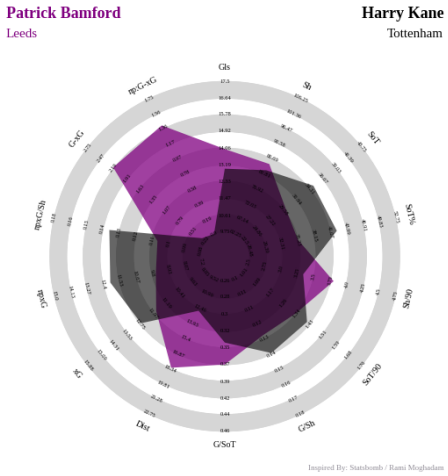

Рисунок 1.10 Сравнение двух форвардов

На рисунке 1.10 представлены нападающий футбольного клуба Лидс Патрик Бэмфорд и нападающий футбольного клуба Тоттенхем Гарри Кейн. Игроки, как говорит диаграмма, довольно разноплановые, но некоторые совпадения все-таки есть. Изучив всех игроков на рынке, можно выбрать наиболее подходящего.

Иногда нужно оценить футболиста по лиге, чтобы найти лучших. Здесь может помочь следующий график:

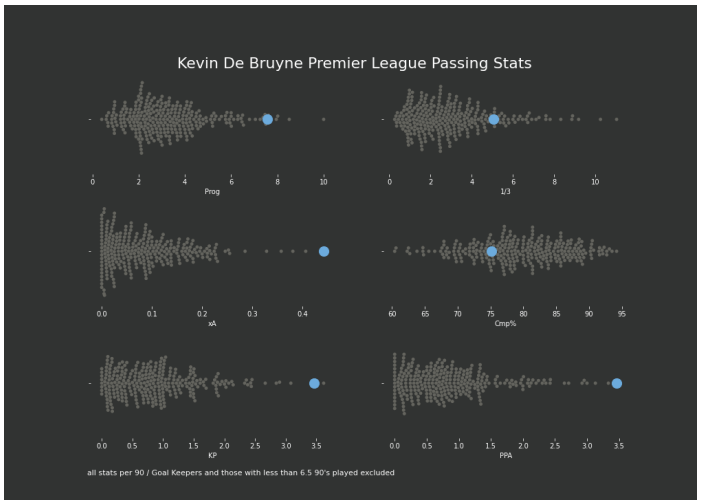

Рисунок 1.11 Показатели Кевина Де Брюйне относительно лиги

Здесь взяты основные показатели, по которым можно оценивать полузащитника(ожидаемы голевые передачи(xA), передачи в финальную треть). Читать его достаточно просто, однако он очень информативен. График объективно показывает превосходство Кевина Де Брюйне в целом по лиге в таких ключевых аспектах для атакующего полузащитника как ожидаемые ассисты и ключевые передачи(КР). Поэтому стоило бы вручать индивидуальные награды опираясь не только на общее впечатление, но и на статистические показатели, так как зачастую наше видение ситуации искажается.

ГЛАВА 2

Математические методы

2.1 Модель ожидаемых голов

На результат футбольного матча(больше, чем в любом другом виде спорта) может сильно влиять случайность. Эти эффекты усиливаются тем фактом, что голы случаются редко; в среднем за матч забивается 2,5 гола. Кроме того, подавляющее большинство матчей заканчивается вничью или разницей в один мяч, а это означает, что один гол может оказывать большое влияние на результат. Поэтому естественным образом становится задача устранения случайных событий при анализе матча.

Для того, чтобы забить гол, необходимо сначала попытаться ударить по воротам. Десять лет назад достаточно было просто взглянуть на общее количество ударов и попаданий в створ. Хотя это полезные инструменты для оценки созданых моментов, они не раскрывают всей истории, поскольку не все удары одинаковы. Здесь в силу вступает xG(ожидаемые голы). xG измеряет вероятность того, что после удара будет забит гол, исходя из ряда факторов. К таким факторам относят расстояние, с которого был произведен удар, угол относительно линии ворот, был ли это удар головой, был ли удар нанесен во время контратаки и другие факторы. Для простоты исследование стоит сосредоточиться только на двух основных(угол и расстояние). Мы можем использовать эту метрику для суммирования всех шансов в матче, чтобы определить, сколько голов должна была забить команда. Можно пойти еще дальше и применить это к серии игр, сезону или даже сроку пребывания тренера в должности.

Таким образом, xG может служить мерилом того, насколько сильна команда в атаке и насколько она надежна в обороне. Его также можно использовать для анализа способности игроков создавать возможности для ударов в опасных зонах и того, насколько хорошо они используют свои шансы. xG помогает исключить часть случайных факторов, связанных с голевыми возможностями, когда мы пытаемся количественно оценить способность команды забивать голы, что в конечном итоге является целью футбольного матча.

Перейдем к математической стороне вопроса. Цель модели — поставить в соответствие каждому удару вероятность забить гол. Если рассмотреть реальные данные, то график будет выглядеть как на рисунке 2.1.

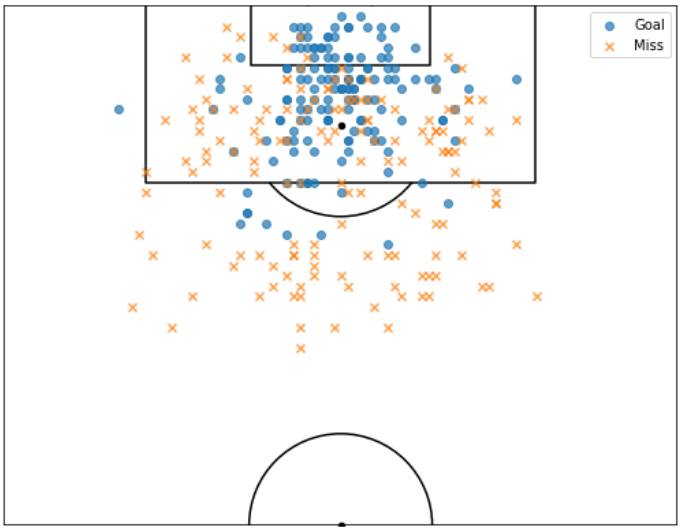

Рисунок 2.1 Распределение голов

Данная задача напоминает задачу классификации, то есть нужно разбить удары по воротам на более или менее опасные. По рисунку 2.1 легко заметить, что вероятность сильно зависит от расстояние и угла. Существует множество функций, которые отображают вероятности и соответствуют неразделимым данным, но мы используем логистическую функцию (также известную как сигмоида) из-за ее простоты. Выглядит она так:

$$G(y) \equiv \frac{1}{1-e^{y}} \equiv \frac{1}{1-e^{\alpha \cdot x+\beta}}$$

Логистическая функция принимает наши параметры(к примеру расстояние) и выводит число от 0 до 1(вероятность гола). Она представляет собой S-образную кривую, которая меняет наклон и траекторию в зависимости от значений коэффициентов. Теперь вопрос в том, как мы можем использовать логистическую функцию для моделирования наших данных по ударам? Для каждой используемой предикторной переменной мы оптимизируем соответствующий коэффициент (α, β и т. д.), чтобы лучше всего соответствовать данным. Оптимизация происходит с помощью логарифмического правдоподобия(Оценка максимального правдоподобия включает в себя рассмотрение проблемы задачи

оптимизации или поиска, где мы ищем набор параметров, который дает наилучшее соответствие для совместной вероятности выборки данных). Подгонка обучающих данных к логистической функции даст коэффициенты для наших предикторов. Если мы начнем только с подбора переменной расстояния, мы должны прийти к оптимальным параметрам для описания обучающих данных:

$$G(y) \equiv \frac{1}{1 - e^{0.146 \cdot distance - 0.097}}$$

Следующий график показывает насколько хорошо функция отображает данные:

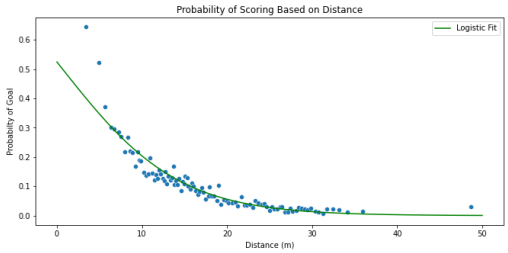

Рисунок 2.2 Первая логистическая модель

Из рисунка 2.2 видно, что модель хорошо предсказывает данные для значений больше 6 метров, но недооценивает вероятность голов с более близких дистанций. Это своего рода преимущество графического подхода. Можно попытаться лучше предсказать удары ближе к цели, добавив квадратичный член к логистической функции.

$$G(y) \equiv \frac{1}{1 - e^{0.216 \cdot distance - 0.002 \cdot distance^2 - 0.584}}$$

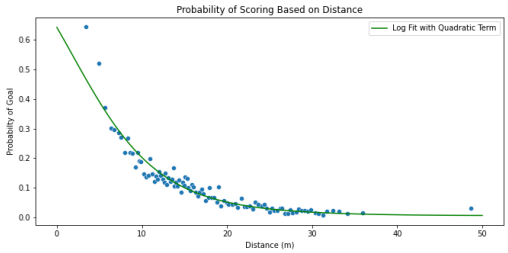

Рисунок 2.3 Вторая логистическая модель

Далее, добавив параметр угла и подогнав эти две переменные к данным, нарисуем вероятность с помощью контурных карт на поле:

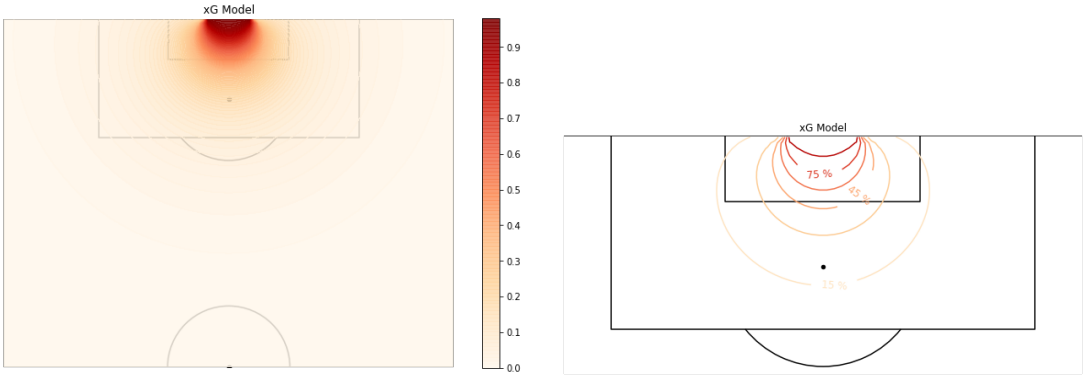

Рисунок 2.4 Контурные карты модели xG

Построенную модель я бы назвал наивной, так как она учитывает всего лишь два фактора. Однако ее можно улучшать и улучшать, добавляя новые предикторы. При этом модель дает все-таки куда более качественный показатель эффективности команды в атаке, нежели просто количество ударов, что еще раз подтвержает преимущество математического подхода.

2.2 Цепи Маркова. Оценка эффективности

Цепь Маркова — последовательность случайных событий с конечным или счётным числом исходов, где вероятность наступления каждого события зависит только от состояния, достигнутого в предыдущем событии. На основании цепей Маркова была разработа модель для тактического анализа и оценки индивидуальных атакующих действий футболистов.

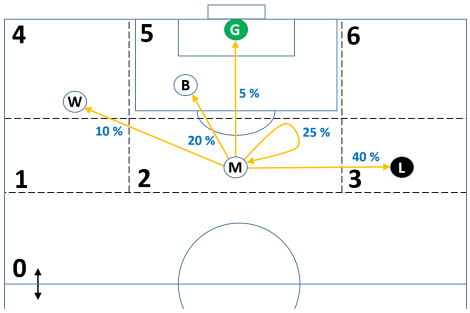

Рисунок 2.5 Разбиение на зоны

Основной акцент уделяется финальной трети поля, для которой были выделены 6 зон. Оставшаяся часть поля была помечена отдельной зоной под номером 0.

Далее вводится понятие состояния игры. Весь игровой процесс рассматривается как последовательность переходов между различными состояниями. Рассматривают такие последовательности переходов как марковские цепи, что позволяет использовать основное свойство данного подхода - отсутствие памяти о прошлых состояниях (основное свойство марковских моделей). Другими словами можно сказать, что для марковских процессов вероятности будущих состояний определяются только текущим состоянием процесса и не зависят от прошлых состояний. Безусловно, может смутить применение данного определения к описанию действий на футбольном поле, т.к. мы знаем, что иногда прошлые состояния в игре могут существенно влиять на то, что будет происходить на поле в следующие моменты времени, но принятое допущение позволяет довольно легко оценивать вероятности будущих состояний игры, в частности вероятность гола (xG), опираясь только на текущее состояние, что в свою очередь упрощает процесс оценки действий футболистов.

Всего в оригинальной модели выделялось 39 состояний:

- 2 состояния, которые характеризуют окончание процесса владения мячом (гол и потеря мяча)

- 7 состояний, которым предшествует остановка игры (то что называется Set play или Set pieces). В данный набор были включены: пенальти, навес с углового, розыгрыш углового, навес со штрафного, розыгрыш штрафного, короткие и длинные вбрасывание из аута.

- 30 состояний игры, определяемые зоной, в которой находится атакующий игрок с мячом, и расположением обороны соперника (некоторые состояния определяются исключительно зоной, некоторые зоной и оборонительной линией) В рассматриваемом примере используются три состояния, которые определяются исключительно зоной, в которой находится атакующий игрок: M - игрок в центре поля, W - игрок на фланге, B - игрок в штрафной. Также в предлагаемом примере указаны два ключевых состояния: G - гол и L - потеря мяча. Оранжевые линии, выходящие из состояния M, указывают

все возможные переходы из текущего состояния и соответствующую вероятность (шансы) данного перехода. Т.е. из состояния M доступно 4 перехода в другие состояния и один "переход"обратно в текущее состояние, который соответствует сохранению мяча атакующим игроком. Общая вероятность всех возможных переходов равна 1.

В таблице ниже приведены значения вероятностей для различных конечных состояний (колонки с оранжевыми заголовками) в зависимости от исходного состояния (колонки с желтыми заголовками). Данная таблица называется матрицей переходов, значения для которой рассчитываются исходя из статистических данных за рассматриваемый промежуток времени (для рассматриваемого примера статистика может выглядеть следующем образом - все игроки всех команд получали мяч в состоянии M (в центре поля) 100 раз, при этом: 25 раз игрок в данной позиции не отдавал передачи и не бил по воротам (совершал движение с мячом), 20 раз отдавал передачи в штрафную в позицию B, 10 раз на фланг в позицию W, 5 раз бил по воротам и забивал гол и 40 раз команда теряла владение в результате передачи на правый фланг в позицию L. Данный пример - условный, как я говорил ранее, в реальности матрица переходов имела размерность 39 на 39 и учитывала 1521 различный переход.

Матрица переходов

Переход из / в	Центр поля (Midfield)	Штрафная (Box)	Фланг (Wing)	Гол (Goal)	Потеря (Lost)
Центр поля (Midfield)	25%	20%	10%	5%	40%
Штрафная (Box)	10%	25%	20%	15%	30%
Фланг (Wing)	10%	10%	25%	5%	50%
Гол (Goal)	0%	0%	0%	100%	0%
Потеря (Lost)	0%	0%	0%	0%	100%

Рисунок 2.6 Матрица переходов

Первая строка соответствует состоянию M и описывает вероятности возможных переходов в другие состояния. Если предположить, что в предыдущий момент времени мяч был отправлен из состояния M в штрафную в состояние B и теперь исходное состояние игры - B, то распределение вероятностей всех воз-

можных конечных состояний, доступных из текущего состояния, можно наблюдать во второй строке, причем мы можем оценить как изменилась вероятность гола (xG) в результате данного действия - xG для нового состояния B минус xG для предыдущего состояния M, что равно 0.15 - 0.05 = 0.1. Т.е. в результате паса из M в B - вероятность гола увеличилась на 0.1. По аналогии мы можем оценить изменение вероятности гола для любой пары состояний и как следствие оценить соответствующие действие игрока по увеличению или уменьшению xG в результате данного действия.

Разработанная модель стала первым инструментом в футбольной аналитике, с помощью которого удалось проводить количественную и качественную оценку атакующих действий футболистов с учетом игрового контекста. Модель на основе цепей Маркова позволяет:

- оценить действие каждого участника атаки в финальной трети поля, а не только двух последних игроков.

- провести количественную и качественную оценку для всех пасов в финальной трети поля, т.е. разработанная модель, основываясь на знаке разности xG между двумя состояниями, позволила выделять пасы, которые увеличивают вероятность гола и которые наоборот, снижают опасность.

Цепи Маркова используют в футболе и по сей день. Также на основе данной модели разрабатываются новые, что еще раз подтвержает ее эффективность.

ЗАКЛЮЧЕНИЕ

В заключение хотелось бы сказать, что анализ данных проник в индустрию футбола относительно недавно. Были заимствованы подходы из других индустрий, однако нельзя отрицать специфичность спортивных и, в частности, футбольных данных. В данной работе были рассмотрены лишь самые основные подходы и модели, однако даже они не всегда используются при развитии футбольных клубов. Но так как спорт это тоже большой бизнес, я считаю, что в будущем без этого не обойтись.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

[1] Н. В. Лазакович, С.П. Сташуленок, О.Л. Яблонский. Теория вероятностей

[2] A Gentle Introduction to Logistic Regression With Maximum Likelihood Estimation [Электронный ресурс] —
Режим доступа: https://machinelearningmastery.com/logistic-regression-with-maximum-likelihood-estimation/