нетология

Дипломная работа

Анализ данных на основе результатов матчей Английской Премьер Лиги с 2014 по 2021 гг.

Курс: Аналитик данных с нуля до middle DAU-28

Оглавление.

Введение.

Каждый момент в футболе имеет значение. Если вы смотрите матч и на секунду отвернетесь, вы можете пропустить важную игру и неожиданный гол. Вот что говорит английский комик Карл Доннелли:

> "Я смотрел матч чемпионата мира 2010 года между Испанией и Парагваем. Было смертельно скучно, поэтому я пошел в туалет и пропустил 3 пенальти, назначенных за эти 2 минуты".

Футбол – один из самых популярных видов спорта на планете, за которым следят миллионы людей. В последние годы было собрано огромное количество данных о многих матчах в разных странах и лигах. Эти данные включают в себя информацию о каждом ударе или пасе, сделанном в матче.

Также, с недавнего времени, в футбольной индустрии большое распространение получила метрика xG (expected goals), которая оценивает вероятность забить гол в зависимости от местоположения игрока на поле и других факторов. Эта метрика позволяет более точно анализировать эффективность команд, а также оценивать игроков, их стоимость и потенциал импакта в команде.

Существует множество возможностей использования этих данных, таких как анализ матчей, определение игровых стилей игроков, прогнозирование и предотвращение травм, прогнозирование результатов матчей и места в турнирной таблице, расчет коэффициентов ставок и многое другое.

Постановка задачи.

Цель данной работы заключается в проведении анализа данных матчей Английской Премьер Лиги за период с 2014 по 2021 гг., а также проверке гипотез, основанных на этих данных.

Для этого необходимо:

- Собрать и обработать данные о матчах Английской Премьер Лиги за период с 2014 по 2021 гг., включая информацию о командах, счетах, статистических показателях и других факторах, которые могут повлиять на исход матча.
- Провести статистический анализ данных, используя различные методы, такие как корреляционный анализ, множественную регрессию и др., для выявления закономерностей и зависимостей между исходами матчей.
- Визуализировать полученные результаты, используя различные графические методы, такие как графики рассеяния, диаграммы ящика с усами, гистограммы и другие, для более наглядного представления результатов анализа.
- Сделать выводы на основе проведенного анализа и проверки гипотез, определить наиболее важные факторы, влияющие на исход матчей.

Актуальность для заказчика.

- Для футбольных клубов, участвующих в АПЛ, важно обеспечивать стабильные и успешные результаты для удержания и привлечения фанатов, получения доходов от продажи билетов, спонсорских соглашений и участия в международных турнирах.
- Анализ данных и прогнозирование результатов матчей позволяют командам выявлять слабые и сильные стороны, оптимизировать стратегии и тактики, а также учесть факторы, влияющие на успех команды, для достижения наилучших результатов.
- Проверка гипотез о влиянии различных факторов на исход матчей может помочь в принятии решений о ставках на матчи букмекерскими конторами и футбольными болельщиками.

Круг стейкхолдеров по задаче.

- Футбольные клубы (тренерский штаб, аналитики, спортивные директоры)
- Футбольные болельщики и зрители
- Спонсоры и партнеры
- СМИ и журналисты
- Организаторы и регуляторы футбольных турниров (Английская Премьер Лига, ФИФА, УЕФА)

Сбор бизнес-требований по задаче.

- Создание системы анализа и визуализации данных, предоставляющей полезную информацию о командах, игроках и матчах для облегчения принятия решений.
- Разработка модели прогнозирования результатов матчей, учитывающей различные факторы, такие как статистика команд, индивидуальные показатели игроков, погодные условия и другие.
- Обеспечение возможности масштабирования и адаптации системы для анализа данных других футбольных лиг и турниров.

Основные гипотезы для проверки.

1. **Команды, играющие на домашнем стадионе, побеждают чаще, чем в гостях.**
2. **Существует сильная корреляция между ожидаемыми голами (xG) и фактическим количеством голов, забитых командами в матчах.**
3. **Команды, которые зарабатывают больше угловых ударов в матче, чаще выигрывают или набирают больше очков.**
4. **Команды, которые ведут в счете после первого тайма, чаще выигрывают матчи или набирают больше очков.**
5. **Команды с более высоким количеством желтых и красных карточек чаще теряют очки в матчах из-за снижения эффективности игры.**
6. **Команды, которые наносят большее количество ударов в створ ворот чаще выигрывают.**

Исследование аналогичных решений.

Проект "Home Advantage in Football Leagues Around the World", Дэвид Шихан:

- В этой статье исследуется общеизвестное, но малопонятное преимущество хозяев поля и то, как оно варьируется в футбольных лигах по всему миру.
- Результаты: Автор показал, что разные лиги имеют разные свойства, но в нигерийской лиге лучше не ставить на выездную команду.

Статья "Using Machine Learning techniques to predict the outcome of professional football matches" Корентин Эрбине (Corentin Herbinet)

- Статья посвящена применению методов машинного обучения для предсказания исходов профессиональных футбольных матчей.

- В рамках данной работы, автор изучает несколько алгоритмов машинного обучения для предсказания исходов футбольных матчей на основе статистических данных. Используемые алгоритмы включают логистическую регрессию, метод опорных векторов (SVM), k-ближайших соседей (KNN), и случайный лес (Random Forest).

- Автор представляет сравнительный анализ эффективности этих алгоритмов, используя набор данных, включающий информацию о прошлых матчах, таких как результаты, количество голов, угловых и фолов.

- Результаты: алгоритм случайного леса показывает наилучшую точность предсказания, составляющую около 60%. Также в статье обсуждаются возможные улучшения и расширения исследования, такие как использование дополнительных данных и применение более сложных алгоритмов машинного обучения или глубокого обучения.

Получение данных.

Используем Python и пакет Selenium для просмотра understat.com и извлечения послематчевых данных для всех необходимых матчей в лиге, по которым имеются данные.

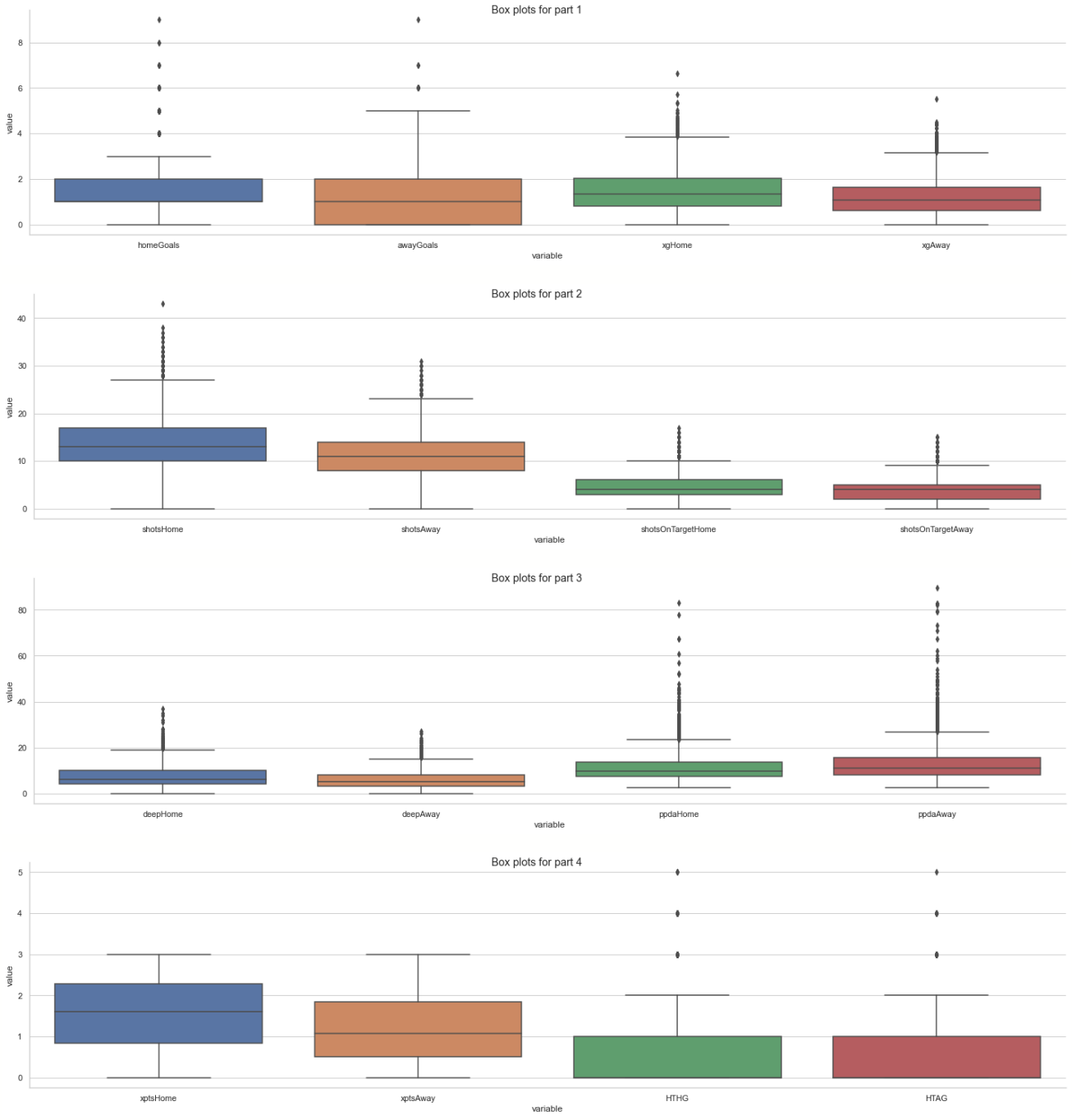
Считываем данные из CSV-файлов скачанных England Football Results и сохраняем их в переменные.

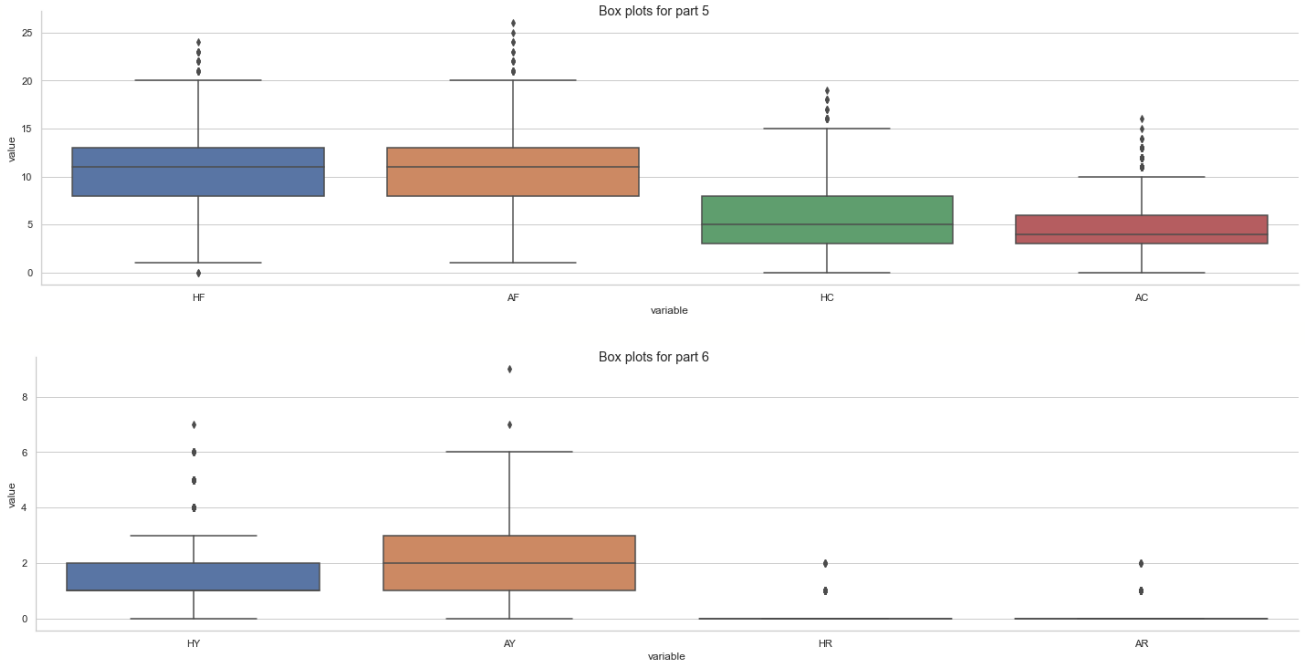
Подготовка и очистка данных.

Прежде чем приступать к анализу данных и проверке гипотез, важно убедиться, что данные корректны, полны и консистентны.

Даже если данные взяты из надежных источников, проверка на аномалии и выбросы все еще имеет смысл, так как ошибки могут возникнуть в процессе сбора, обработки и представления данных.

Создаем боксплоты с выбросами для всех столбцов.





Выбросы, которые выходят за боксплот, могут указывать на наличие в данных необычных или аномальных значений. Однако, эти значения не всегда являются ошибками или неточностями и могут быть результатом реальных процессов или явлений.

В таком случае, решение о том, что делать с выбросами, зависит от конкретной ситуации и целей анализа данных. Если выбросы являются результатом естественных процессов или явлений, то они могут оставаться в данных. (исключение выбросов может привести к потере важной информации.)

В конечном итоге мы получили таблицу в которой собрана статистика матчей с 2014 по 2021 гг (3040 матчей).

- В таблице нет пропущенных значений или NaN.
- Все типы данных в каждом столбце соответствуют ожидаемым.
- Дублирующиеся строки отсутствуют.
- Данные весьма однородны и аномалий в них не обнаружено.

Данная таблица содержит информацию о футбольных матчах, проведенных между домашней и гостевой командами. Структура таблицы включает в себя 31 столбец:

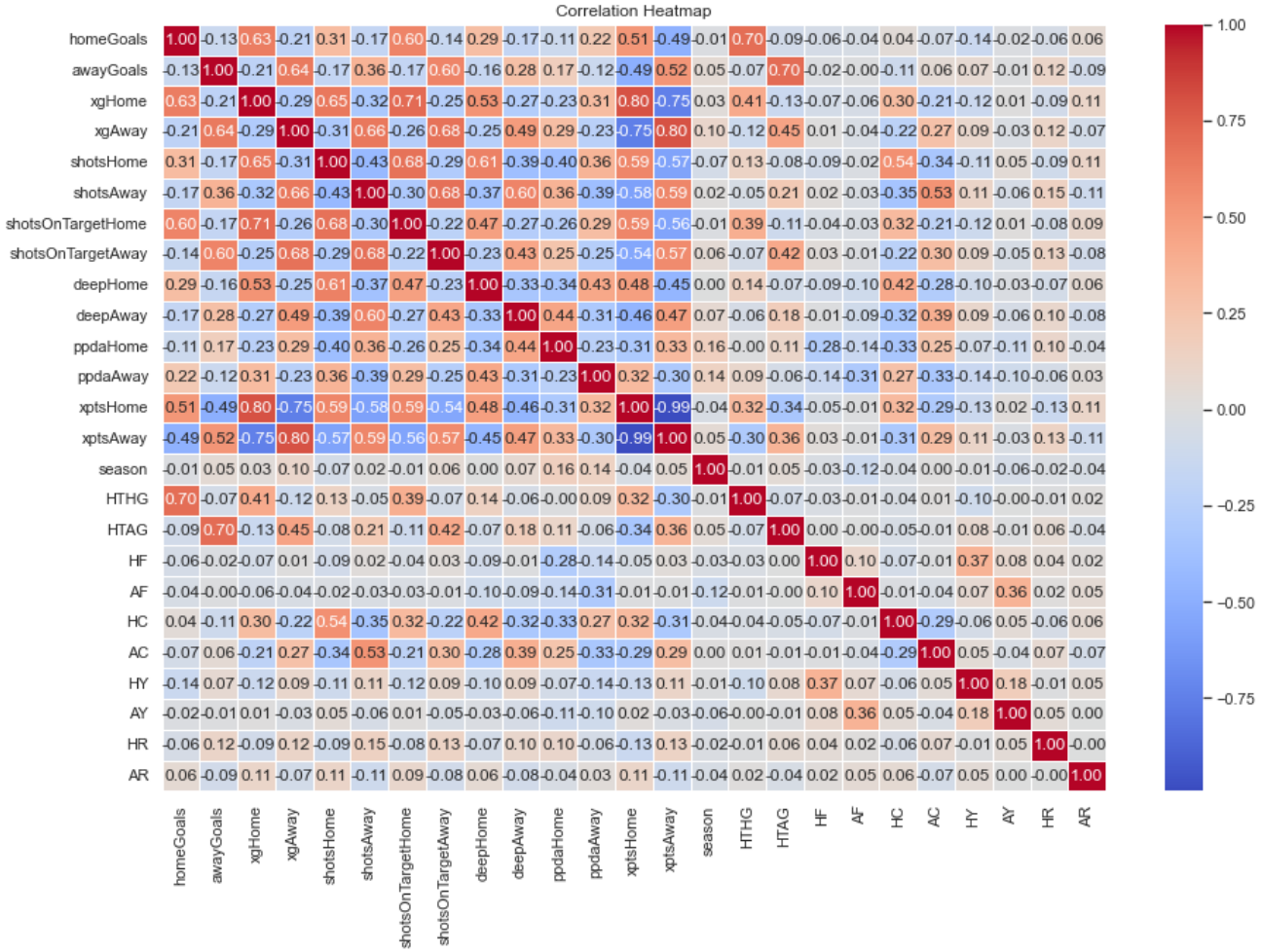
1. date - дата проведения матча.
2. homeTeam - название домашней команды.
3. awayTeam - название гостевой команды.
4. homeGoals - количество забитых голов домашней командой.

5. awayGoals - количество забитых голов гостевой команды.
6. xgHome - ожидаемое количество голов, которое должна была забить домашняя команда.
7. xgAway - ожидаемое количество голов, которое должна была забить гостевая команда.
8. shotsHome - количество ударов по воротам домашней команды.
9. shotsAway - количество ударов по воротам гостевой команды.
10. shotsOnTargetHome - удары в створ ворот, которые нанесла домашняя команда.
11. shotsOnTargetAway - удары в створ ворот, которые нанесла гостевая команда.
12. deepHome - это показатель, для обозначения количества передач домашней команды, выполненных вблизи ворот соперника (20 – 25 метров).
13. deepAway - это показатель, для обозначения количества передач гостевой команды, выполненных вблизи ворот соперника (20 – 25 метров).
14. ppdaHome - PPDA (Passes Allowed Per Defensive Action) — футбольный статистический показатель, который позволяет определить интенсивность прессинга в матче. Чем меньше значение PPDA, тем выше интенсивность игры домашней команды в обороне.
15. ppdaAway - показатель, который позволяет определить интенсивность прессинга в матче. Чем меньше значение PPDA, тем выше интенсивность игры гостевой команды в обороне.
16. xptsHome - (xPoints) – «ожидаемые набранные очки домашней команды». По формуле полной вероятности рассчитывается возможность спортивного коллектива победить, проиграть или сыграть вничью. Цифры получаются дробными – это символическое ожидание значения, а не сколько очков наберёт команда в результате (победа – 3, ничья – 1, поражение – 0).
17. xptsAway – «ожидаемые набранные очки гостевой команды».
18. season - сезон, период который определяется двумя датами (год начала и год завершения сезона).
19. FTR - Full Time Result - результат игры в конце ее основного времени. FT result может быть победой домашней команды, ничьей или победой гостевой команды. (H=Home Win, D=Draw, A=Away Win)
20. HTHG - Half Time Home Team Goals - количество голов, забитых домашней командой в первом тайме матча.
21. HTAG - Half Time Away Team Goals - количество голов, забитых гостевой командой в первом тайме матча.
22. HTR - Half Time Result - результат игры к концу первого тайма. (H=Home Win, D=Draw, A=Away Win)
23. Referee - имя рефери.

24. HF - Home Team Fouls Committed - количество фолов, совершенных домашней командой в течение игры.
25. AF - Away Team Fouls Committed - количество фолов, совершенных гостевой командой в течение игры.
26. HC - Home Team Corners - количество угловых ударов, исполненных домашней командой во время игры.
27. AC - Away Team Corners - количество угловых ударов, исполненных гостевой командой во время игры.
28. HY - Home Team Yellow Cards - количество желтых карточек, полученных игроками домашней команды во время игры.
29. AY - Away Team Yellow Cards - количество желтых карточек, полученных игроками гостевой команды во время игры.
30. HR - Home Team Red Cards - количество красных карточек, полученных игроками домашней команды во время игры.
31. AR - Away Team Red Cards - количество красных карточек, полученных игроками гостевой команды во время игры.

Анализ данных.

Построим тепловую карту корреляции между различными показателями, чтобы определить наиболее связанные переменные



- Положительная корреляция: Если увеличение признака A приводит к увеличению признака B, то они положительно коррелируют. Значение 1 означает идеальную положительную корреляцию.
- Отрицательная корреляция: Если увеличение признака A приводит к уменьшению признака B, то они коррелируют отрицательно. Значение -1 означает идеальную отрицательную корреляцию.

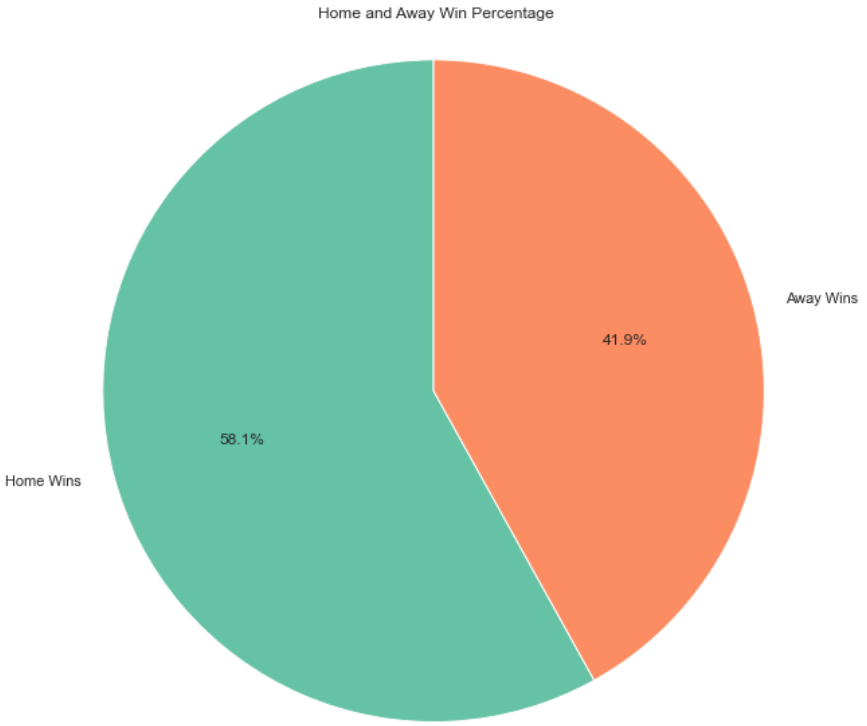
Построим столбчатую диаграмму.сгруппированных и агрегированных данных по сезонам (голы, удары, угловые, желтые карточки, красные карточки).

Выводы:

- По результатам анализа графика можно сделать вывод, что показатели в указанные сезоны имели относительно стабильные значения без существенных колебаний.
- Однако, в сезоне 2014-2015 было выдано наибольшее количество красных карточек в сравнении с последующими сезонами, где среднее количество выданных красных карточек составило 43,5.
- Следует отметить, что количество красных карточек за сезон может колебаться случайным образом из-за различных факторов, таких как форма команд, индивидуальная агрессия игроков и т.д.
- Возможно, сезон 2014-2015 был аномальным в этом плане, и после этого количество выданных красных карточек вернулось к более нормальным значениям.
- Данные по красным карточкам за предыдущие сезоны
    - 2009 - 67
    - 2010 - 64
    - 2011 - 64
    - 2012 - 52
    - 2013 - 53.

Проверка гипотезы "Команды играющие на домашнем стадионе побеждают чаще, чем в гостях"

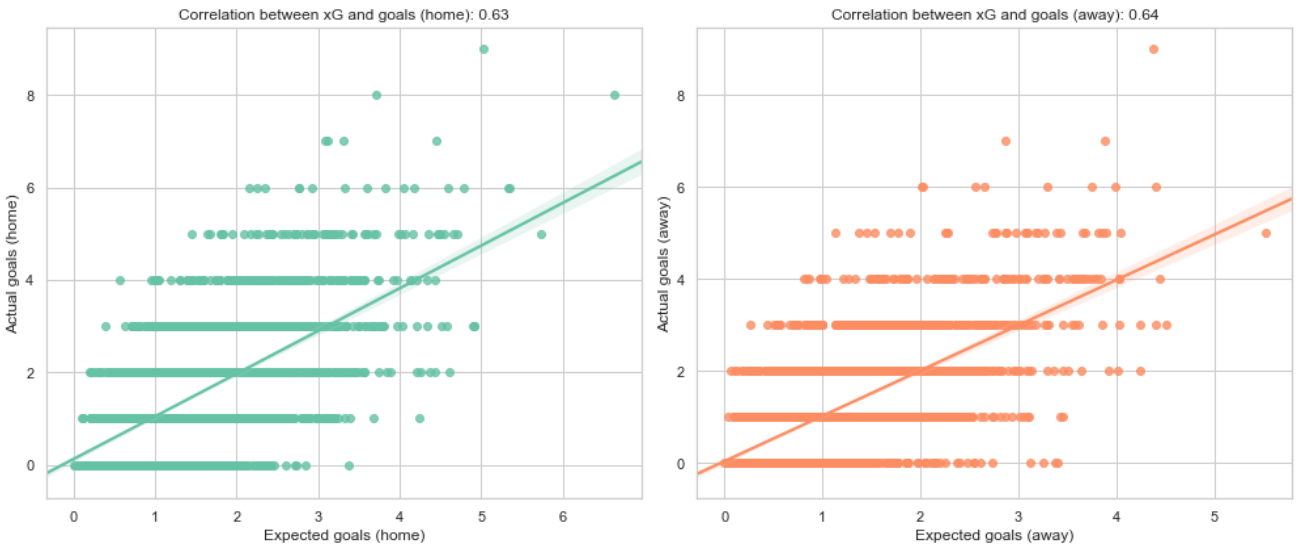
Построим круговую диаграмму с процентами побед дома и в гостях.



Выводы:

- В большинстве сезонов с 2014 по 2021 год домашние команды в АПЛ чаще всего одерживали победы, что свидетельствует о том, что домашний стадион и поддержка болельщиков могут оказывать положительное влияние на результаты команд.
- В 2020 году произошло отклонение от общей тенденции - такое изменение в результатах может быть связано с влиянием пандемии COVID-19 на спортивные мероприятия и состояние команд. Возможно, отсутствие или ограничение числа зрителей на стадионах снизило преимущество домашних команд, что привело к росту проигрышей

Проверка гипотезы "Существует сильная корреляция между ожидаемыми голами (xG) и фактическим количеством голов, забитых командами в матчах."



Выводы:

- Коэффициенты корреляции 0.63 и 0.64 указывают на среднюю положительную корреляцию между ожидаемыми голами (xG) и фактическим количеством голов для домашних и гостевых команд. Это означает, что в целом команды, имеющие более высокий показатель xG, часто забивают больше голов.

- Полученные результаты подтверждают гипотезу о существовании корреляции между ожидаемыми голами (xG) и фактическим количеством голов, забитых командами в матчах. Однако стоит отметить, что корреляция средняя, а не сильная.

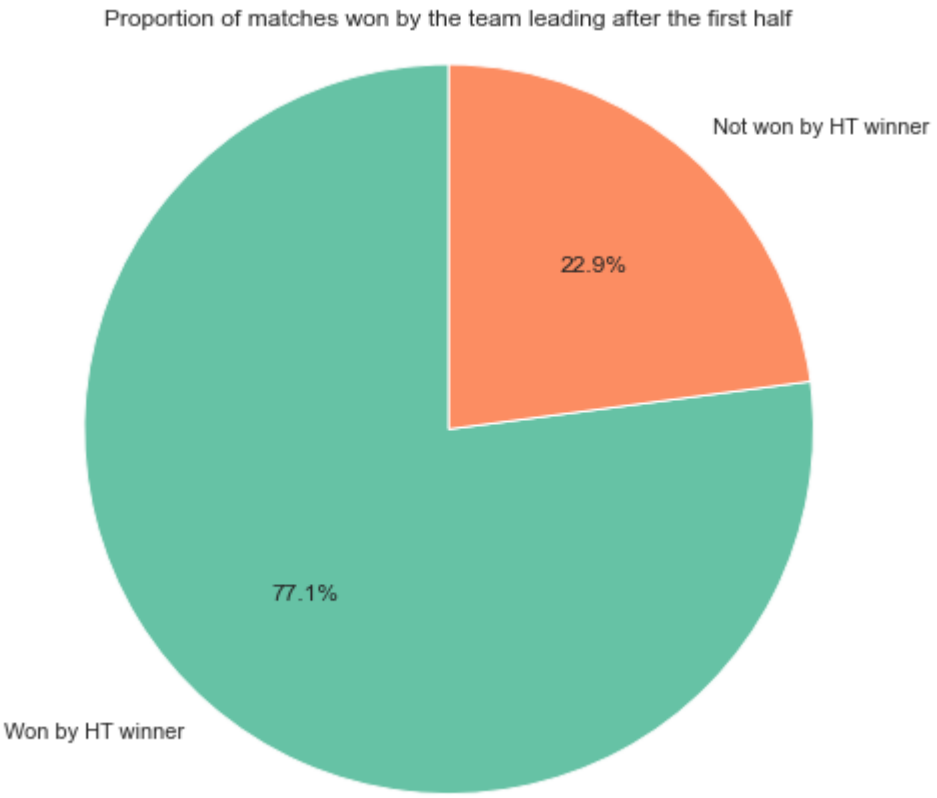
Проверка гипотезы "Команды, которые зарабатывают больше угловых ударов в матче, чаще выигрывают или набирают больше очков."

Выводы:

- Доля побед команды с большим количеством угловых ударов составляет примерно 38%.
- Коэффициент корреляции между числом угловых ударов и количеством забитых голов для домашних матчей составляет 0.018.
- Коэффициент корреляции между числом угловых ударов и количеством забитых голов для выездных матчей составляет -0.0047.
- На основе этих результатов можно сделать вывод, что количество угловых ударов не имеет существенного влияния на результаты матчей, как для команд, играющих дома, так и для команд, играющих на выезде.

Проверка гипотезы "Команды, которые ведут в счете после первого тайма, чаще выигрывают матчи или набирают больше очков."

Рассчитаем долю матчей, выигранных командой, ведущей после первого тайма: 0.77%

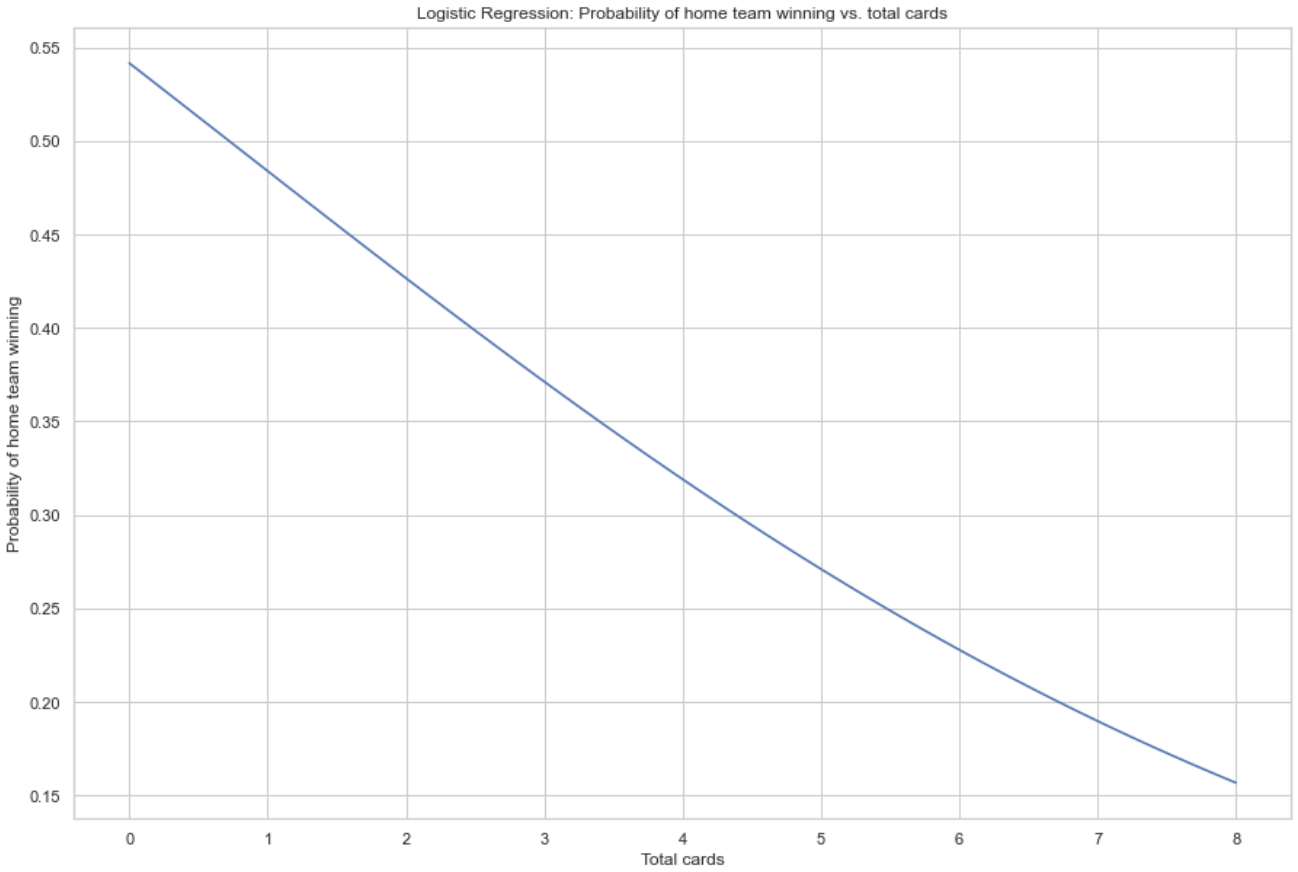


Вывод:

- Доля матчей, выигранных командой, которая ведет в счете после первого тайма составляет примерно 77.1%. Исходя из этого, можно сделать вывод, что команды, которые ведут после первого тайма, действительно чаще выигрывают матчи.

Проверка гипотезы "Команды с более высоким количеством желтых и красных карточек чаще теряют очки в матчах из-за снижения эффективности игры"

Возьмем только домашние команды.
Коэффициент логистической регрессии: -0.23124761

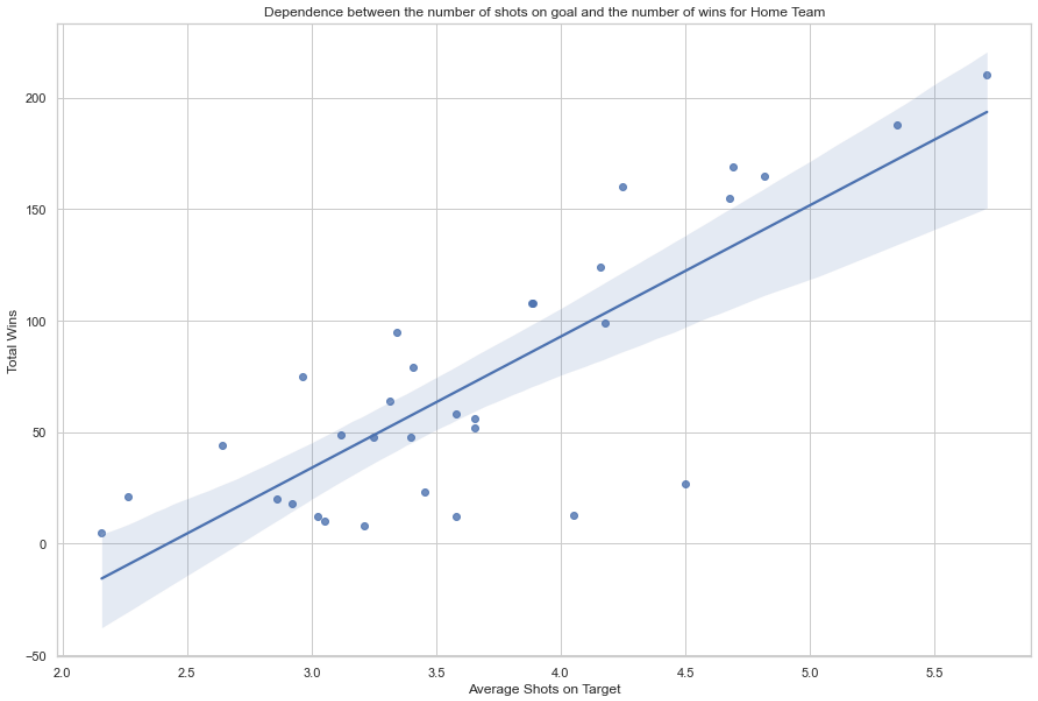


Выводы:

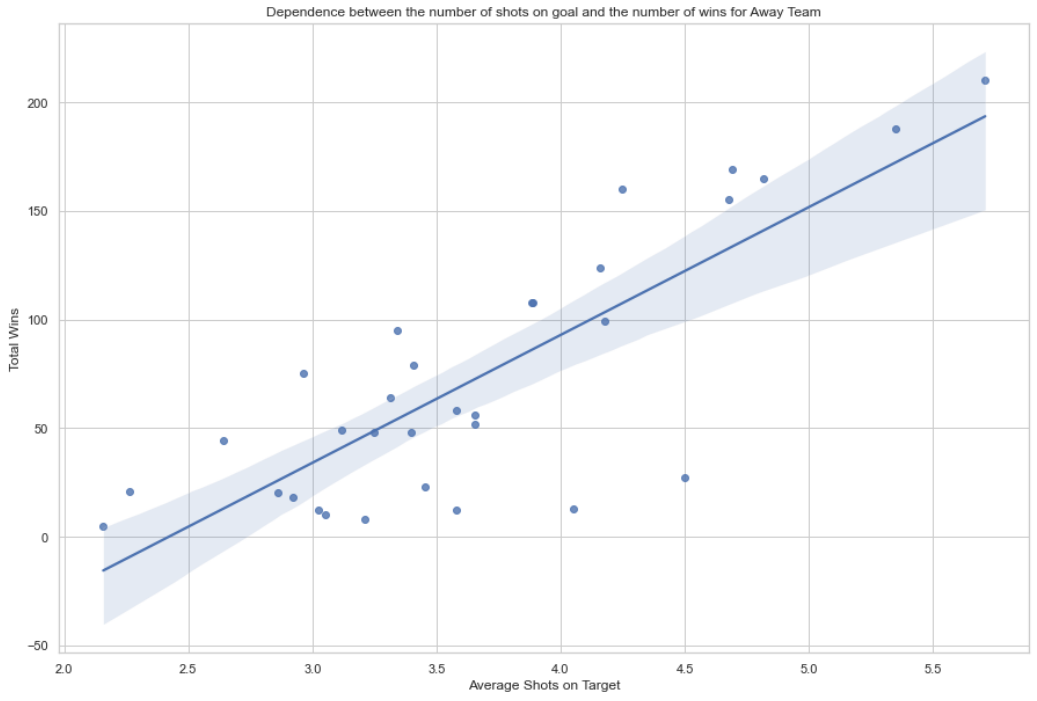
- Точность модели составляет примерно 57.6%. Учитывая, что точность бинарного классификатора составляет примерно 50%, модель логистической регрессии незначительно превосходит бинарный классификатор.
- Коэффициенты логистической регрессии: -0.231. Отрицательный коэффициент указывает на то, что с увеличением количества карточек вероятность победы команды уменьшается. Однако коэффициент достаточно мал, что может говорить о слабом влиянии количества карточек на исход матча.
- В целом, результаты анализа говорят о том, что влияние количества карточек на исход матча есть, но оно достаточно слабое.

Проверка гипотезы "Команды, которые наносят большее количество ударов в створ ворот чаще выигрывают"

Коэффициент корреляции Пирсона для домашних команд: 0.882



Коэффициент корреляции Пирсона для гостевых команд: 0.811



Выводы:

- Значения коэффициентов корреляции Пирсона равные 0.88 для домашних команд и 0.81 для гостевых команд подтверждает нашу гипотезу о том, что команды, которые наносят большее количество ударов в створ ворот, чаще выигрывают. (однако следует помнить, что корреляции не подразумевает причинно-следственную связь)

Общие выводы.

В ходе анализа данных были проверены и подтверждены следующие гипотезы:

1. **Команды, играющие на домашнем стадионе, действительно побеждают чаще, чем в гостях.**
2. **Ожидаемое количество голов (xG) средне положительно коррелирует с фактическим количеством голов, что указывает на то, что команды с более высоким показателем xG, в целом, забивают больше голов.**
3. **Команды, которые ведут после первого тайма, чаще выигрывают матчи (доля таких матчей составляет примерно 77.1%).**
4. **Команды, наносящие большее количество ударов в створ ворот и имеющие больше атакующих действий вблизи ворот соперника, чаще выигрывают матчи.**

Однако гипотеза о том, что команды, зарабатывающие больше угловых ударов, чаще выигрывают или набирают больше очков, не подтвердилась.

Также было выявлено, что влияние количества карточек на исход матча есть, но оно достаточно слабое.

Области внедрения результатов.

- Результаты анализа могут быть полезны для тренеров, аналитиков и людей, занимающихся футбольными ставками. Они могут использовать эту информацию для разработки стратегий игры, определения ключевых факторов, влияющих на исход матча, и принятия более обоснованных решений о ставках. Также результаты могут быть полезны для составления прогнозов на исход футбольных матчей и определения вероятности победы каждой из команд.
- В целом, данное исследование показывает, что определенные показатели, такие как домашний стадион, ожидаемое количество голов (xG), результаты первого тайма, количество ударов в створ ворот и атакующих действий вблизи ворот соперника, имеют существенное влияние на исход футбольных матчей и могут быть использованы для предсказания результатов игр.

Пути развития и улучшения решения.

1. Расширение набора данных: Включение данных из других лиг, сезонов и турниров может помочь улучшить обобщающую способность модели и повысить точность прогнозов.
2. Дополнительные переменные: Исследование и включение дополнительных переменных, таких как статистика игроков, состояние погоды, местоположение стадиона, состояние газона и т.д., может помочь улучшить точность модели и выявить новые факторы, влияющие на исход матча.
3. Применение машинного обучения: Разработка моделей машинного обучения, таких как логистическая регрессия, случайный лес или градиентный бустинг, может повысить точность прогнозов исходов матчей, а также помочь выявить более сложные зависимости между переменными.
4. Интеграция с другими данными: Совмещение данных о матчах с данными о ставках может помочь лучше понять, какие факторы учитываются на рынке ставок и использовать эту информацию для определения наиболее выгодных ставок.

Список используемых источников.

1. understat.com
2. England Football Results