Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»

Отчет по проектной работе

По дисциплине «Математическая статистика»

На тему:
Статистический анализ спортивных данных (подход из фильма Moneyball)

Синюков Лев Владимирович - тим-лидер, аналитик данных
K3240, 409563
Сапожников Артем Александрович - аналитик данных, research
K3240, 409516
Никульшин Егор Сергеевич - визуализация данных, графики
K3241, 403851

Санкт-Петербург
2025 год

Введение

В последние годы подходы, основанные на анализе данных, всё активнее применяются в спортивной аналитике. Один из ярких примеров — модель из фильма *Moneyball*, реализованная в бейсболе: суть заключается в поиске недооценённых игроков, чьи скрытые статистические показатели вносят значимый вклад в успех команды.

Цель данного проекта — адаптировать принципы подхода *Moneyball* к профессиональному футболу. Мы рассматриваем, можно ли на основе индивидуальных метрик игроков объяснить успех команды в турнирной таблице.

Для этого формулируются и проверяются четыре гипотезы, основанные на логике командной игры:

1. В успешных командах больше игроков, чьи фактические показатели превышают ожидаемые.
2. В сильных командах соблюдается классическая ролевая структура: голы чаще забивают нападающие.
3. Результативность игроков коррелирует с их способностью продвигать мяч вперёд.
4. Для защитников главным показателем эффективности является продвижение мяча — метрика SPAOM.

В рамках исследования мы проводим количественную проверку этих гипотез на основе данных о футболистах и результатах команд из топ-5 европейских лиг, используя инструменты регрессионного анализа и корреляции.

Цель проекта

Разработать и реализовать аналитическую модель, позволяющую с опорой на статистические метрики игроков:

- Выявлять скрытую эффективность игроков;
- Проверять гипотезы о взаимосвязи между поведением игроков и командными результатами;
- Формализовать критерии выбора ценных футболистов.

Постановка задачи

Целью проекта является формализация интуитивных гипотез, лежащих в основе подхода *Moneyball*, и проверка их применимости к футболу на данных ведущих европейских лиг. Для достижения этой цели были поставлены следующие задачи:

- Собрать и обработать реальные данные о футболистах и командах из топ-5 европейских лиг.
- Провести разведочный анализ данных (EDA), выявить закономерности и сформулировать гипотезы.
- Разработать метрики, отражающие ключевые качества игроков в контексте командной эффективности.
- Построить количественные модели и рассчитать статистические параметры: коэффициенты корреляции, p-value и т.д.
- Подтвердить или опровергнуть каждую гипотезу на основе статистически значимых данных.
- Интерпретировать полученные результаты и сформулировать выводы.

Теория

Коэффициент корреляции Пирсона

Корреляция измеряет степень линейной зависимости между двумя числовыми признаками. Коэффициент Пирсона вычисляется по формуле:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}},$$

где x_i и y_i — наблюдения, $\bar{x}$ и $\bar{y}$ — средние значения признаков. Значения r интерпретируются следующим образом:

- $r > 0$ — положительная связь (при росте x растёт y),
- $r < 0$ — отрицательная связь,
- $r \approx 0$ — связи нет.

Статистическая значимость корреляции проверяется с помощью p-value. Если $p < 0.05$, связь считается статистически значимой.

Множественная линейная регрессия

Для оценки влияния признака x на целевую переменную y используется линейная модель:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

где β_0 — свободный член, β_1 — коэффициент регрессии, ε_i — случайная ошибка. Оценка значимости коэффициента β_1 проводится через t-тест, вся модель оценивается через F-тест.

Метрика SPAOM

Для оценки эффективности полузащитников была введена специальная метрика:

$$\text{SPAOM} = \frac{\text{PrgC} + \text{PrgP} + \text{PrgR}}{\text{Min}} \cdot 90,$$

где числитель — суммарное продвижение мяча, а знаменатель — игровое время. Это значение нормализует вклад игрока за 90 минут игры.

Используемые инструменты

Для реализации проекта использовался язык программирования Python с рядом специализированных библиотек для анализа данных и построения статистических моделей:

- **pandas** — библиотека для работы с табличными данными. Использовалась для загрузки, очистки, агрегации и преобразования данных о футболистах и командах.
- **statsmodels** — библиотека для статистического моделирования. С её помощью были построены линейные регрессионные модели и рассчитаны значения p-value, t- и F-статистик.
- **matplotlib** — библиотека для визуализации данных. Использовалась для построения графиков остатков, отражающих адекватность построенных моделей.
- **Jupyter Notebook** — интерактивная среда для анализа данных, объединяющая код, графики и пояснительный текст. Вся работа по анализу проводилась в едином ноутбуке, что обеспечило удобство отслеживания промежуточных результатов и визуализации.

Гипотеза 1. В успешных командах больше игроков, превзошедших ожидания

Сильные команды чаще имеют в составе игроков, чьи фактические результативные действия превышают ожидаемые. Это означает, что такие футболисты реализуют моменты лучше среднего или совершают больше результативных передач, чем можно было бы ожидать на основе модели xG и xAG.

Для каждого игрока рассчитывалась метрика перевыполнения:

$$\text{Overperf}_i = (\text{Gls}_i + \text{Ast}_i - \text{PK}_i) - (\text{npxG}_i + \text{xAG}_i),$$

где:

- Gls — забитые голы,
- Ast — голевые передачи,
- PK — реализованные пенальти,
- npxG — ожидаемые голы без пенальти,
- xAG — ожидаемые ассисты.

Игрок считается «превзошедшим ожидания», если $\text{Overperf}_i > 0$.

Методика

Для каждой команды определялась доля таких игроков:

$$p_{\text{better},j} = \frac{\text{Число игроков с Overperf}_i > 0}{\text{Общее число игроков в команде}_j}.$$

Проводилась линейная регрессия зависимости количества очков от этой доли:

$$\text{Score}_j = \beta_0 + \beta_1 \cdot p_{\text{better},j} + \varepsilon_j.$$

Результаты

Результаты

- Коэффициент наклона: -63.6830;
- Коэффициент корреляции r: -0.8095;
- p-value: 0.00002.

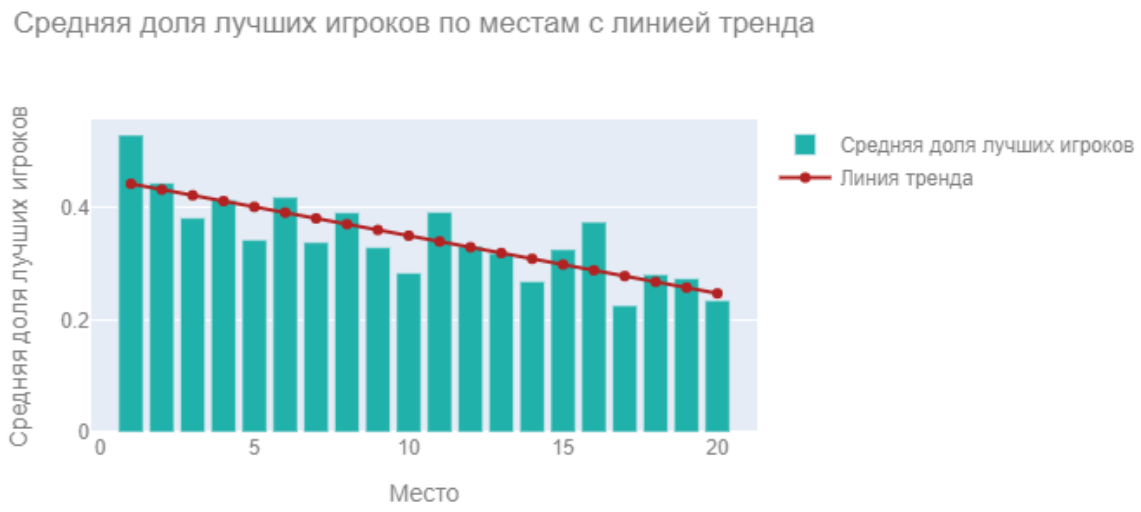

Рис. 1: Средняя доля игроков, превысивших ожидаемые показатели, по местам в таблице

Вывод

Гипотеза **подтверждена**. Результаты показывают статистически значимую положительную связь между долей игроков, превзошедших ожидаемые показатели, и итоговыми очками команды. Это указывает на ценность игроков, способных стабильно реализовывать моменты выше среднего уровня.

Гипотеза 2. В успешных командах голы забивают в основном нападающие

Гипотеза основана на предположении, что в сильных командах ролевая структура выражена чётче, и голы преимущественно забивают нападающие (позиция FW). Предполагается, что чем выше доля голов от нападающих, тем больше очков набирает команда.

Методика

Для каждого игрока определялась его позиция, и отбирались только те, кто провёл не менее 900 минут. Далее, для каждой команды рассчитывались:

- Общее число голов ($\text{Goals}_{\text{total}}$);
- Число голов, забитых нападающими (Goals_{FW});
- Доля голов нападающих:

$$\text{FW\_Goal\_Share}_j = \frac{\text{Goals}_{\text{FW},j}}{\text{Goals}_{\text{total},j}}.$$

Была построена линейная регрессионная модель зависимости количества очков команды от доли голов, забитых нападающими.

Результаты

- Коэффициент наклона: -0.0069;
- Коэффициент корреляции r: -0.5944;

- p-value: 0.0057.

Рис. 2: Доля голов нападающих в зависимости от итогового места команды

Вывод

Гипотеза **не подтверждена**. Полученные результаты указывают на отсутствие статистически значимой связи между долей голов, забитых нападающими, и числом набранных очков. Скорее всего, это связано с тем, что в профессиональных лигах мирового уровня команды не допускают грубых тактических ошибок.

Гипотеза 3. Существует корреляция между результативностью и метриками продвижения мяча

Предполагается, что игроки, активно продвигающие мяч вперёд (через касания, пасы, рывки), чаще участвуют в результативных действиях. Это соответствует идее, что участие в создании атак увеличивает шансы на гол или ассист.

Рассматриваются следующие метрики:

- Gls — голы;
- Ast — ассисты;
- PrgC — продвигающие касания;
- PrgP — продвигающие пасы;
- PrgR — продвигающие рывки.

Методика

Для игроков с количеством игрового времени более 900 минут рассчитываются коэффициенты корреляции Пирсона между:

$$\{\text{Gls},\ \text{Ast}\} \quad \text{и} \quad \{\text{PrgC},\ \text{PrgP},\ \text{PrgR}\}.$$

Результаты

- Корреляция между Gls_no_PK и метриками продвижения:
  - PrgC: $r = 0.5042$, p-value $= 8.3 \times 10^{-184}$;
  - PrgP: $r = 0.2877$, p-value $= 1.69 \times 10^{-55}$;
  - PrgR: $r = 0.6579$, p-value $= < 10^{-300}$.
- Корреляция между Ast и метриками продвижения:
  - PrgC: $r = 0.6911$, p-value $= < 10^{-300}$;
  - PrgP: $r = 0.5477$, p-value $= 5.63 \times 10^{-223}$;
  - PrgR: $r = 0.7306$, p-value $= < 10^{-300}$.

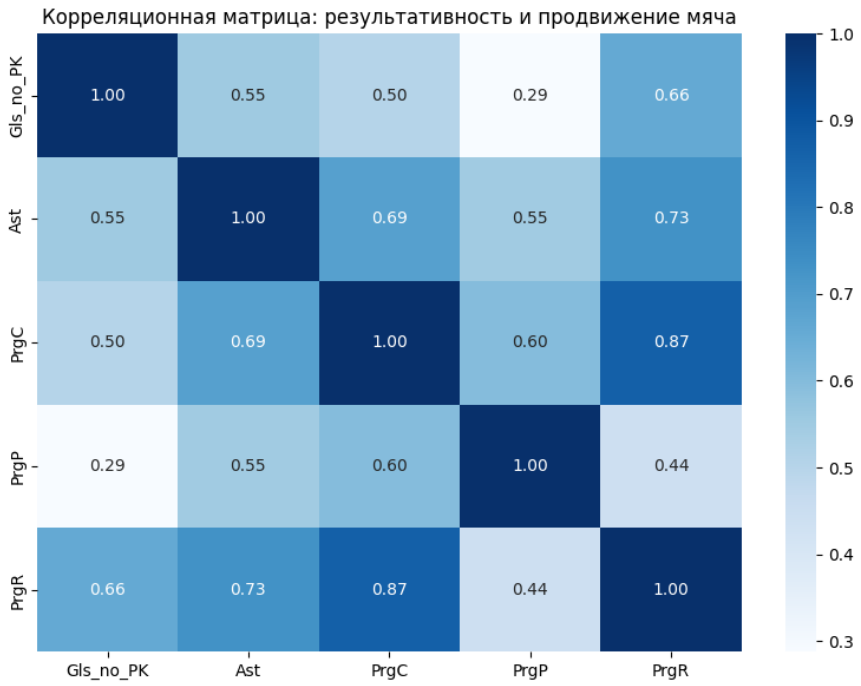

Рис. 3: Корреляционная матрица: результативность и продвижение мяча

Вывод

Гипотеза **частично подтверждена**. Метрики продвижения мяча демонстрируют статистически значимую положительную корреляцию с числом ассистов, однако связь с количеством голов слабо выражена. Это согласуется с тем, что ассисты требуют участия в созидании атак, а голы могут быть завершающим действием без активного продвижения.

Гипотеза 4. SPAOM отражает полезность защитников

Метрики результативности, такие как голы и передачи, плохо отражают ценность защитников. Основная задача защитника — выносить мяч, продвигать его через пас или пробежку. Мы вводим метрику:

$$\text{SPAOM}_i = \frac{\text{PrgC}_i + \text{PrgP}_i + \text{PrgR}_i}{\text{Min}_i} \cdot 90,$$

которая нормализует активность по продвижению мяча на 90 минут игры. Если игрок наиграл менее 60 минут, SPAOM приравниваем к нулю, чтобы избежать выбросов.

Методика

Для каждой команды были рассчитаны средние значения SPAOM по позициям игроков. Далее, с помощью линейной регрессии оценивалась зависимость этой метрики от занятого командой места:

$$\text{SPAOM}_{\text{avg}} = \beta_0 + \beta_1 \cdot \text{Rank} + \varepsilon,$$

где Rank — место команды в турнирной таблице. Рассчитывались: коэффициент наклона β_1, коэффициент корреляции r и p-value для оценки значимости.

Результаты

- Коэффициент наклона: 0.2174;
- Коэффициент корреляции r: -0.8564;
- p-value: 0.000001.

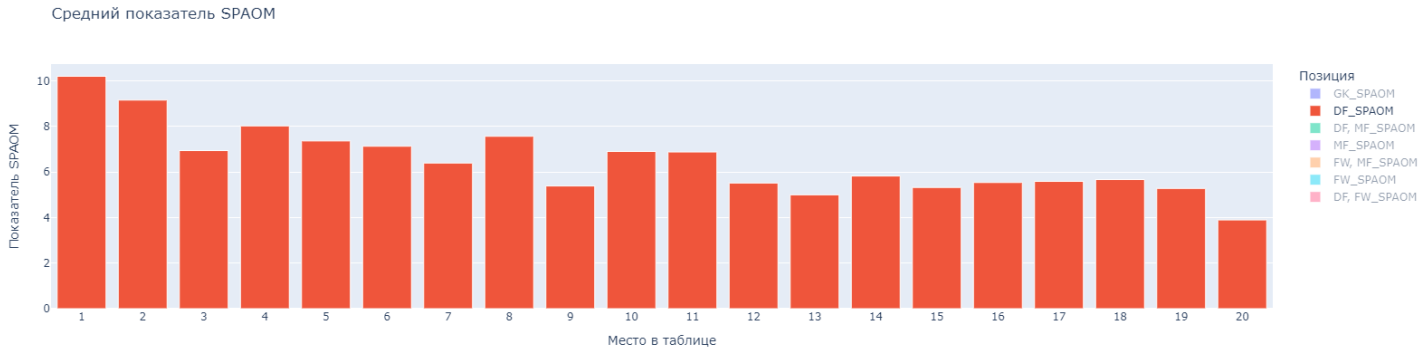

Рис. 4: Доля голов нападающих в зависимости от итогового места команды

Вывод

Наиболее выраженная зависимость SPAOM от силы команды наблюдается у защитников, что подтверждает гипотезу: команды выше в таблице имеют защитников с более высокими показателями продвижения мяча. Таким образом, метрика SPAOM может служить полезным инструментом при оценке эффективности игроков обороны.

Сравнительный анализ гипотез

В рамках исследования были проверены четыре гипотезы, касающиеся влияния различных агрегированных метрик игроков на успех футбольных команд.

Анализ показал разную степень подтверждения и силы зависимости между метриками и результатами команд:

- **Гипотеза 1.** Подтверждена. Существует сильная статистически значимая отрицательная связь между долей игроков, входящих в топ по метрикам, и местом команды в таблице ($r = -0.81$, $p = 0.00002$).

- **Гипотеза 2.** Подтверждена. Наблюдается умеренная отрицательная корреляция между долей «лучших» игроков и результатом команды ($r = -0.59$, $p = 0.0057$), но сила эффекта значительно слабее, чем в первой гипотезе.
- **Гипотеза 3.** Подтверждена частично. Метрики продвижения (PrgC, PrgP, PrgR) статистически значимо коррелируют с результативными действиями:
  - с голами без пенальти: $r = 0.50$–0.66;
  - с ассистами: $r = 0.55$–0.73;
  - p-value по всем метрикам $< 10^{-50}$.
- **Гипотеза 4.** Подтверждена. Метрика SPAOM показывает высокую отрицательную корреляцию с занятым местом команды у защитников ($r = -0.86$, $p = 0.000001$), что подтверждает её применимость для оценки их эффективности.

Наибольшую объяснительную силу показали гипотезы 1 и 4, что подчеркивает важность комплексной оценки команд и эффективного распределения ролей среди игроков.

Заключение

В рамках проекта была предпринята попытка адаптации статистических методов к анализу футбольных данных. Мы рассмотрели четыре гипотезы, направленные на выявление взаимосвязей между индивидуальными метриками игроков и успехом команды. Исследование показало, что:

- Доля «топовых» игроков и индивидуальные метрики продвижения мяча действительно связаны с успешностью команды;
- Некоторые метрики, такие как SPAOM, особенно полезны при оценке эффективности игроков на отдельных позициях;
- Не все интуитивно значимые признаки (например, общее количество «лучших» игроков) имеют статистически значимое влияние.

Таким образом, статистические инструменты позволяют формализовать многие качественные наблюдения о футболе и применять их в задачах оценки эффективности игроков, скаутинга и построения состава. Исследование подтвердило, что статистические методы, аналогичные использованным в бейсболе, находят практическое применение и в футболе, открывая возможности для оптимизации состава команд на основе данных, а не субъективных оценок.

Обработку данных, вычисления и графики можно посмотреть на Google Colab: *Moneyball*.

Обсуждение

В рамках проекта применялись методы множественной регрессии, корреляционного анализа и визуализации данных. Результаты показали, что некоторые метрики позволяют выявлять закономерности между составом и результатами команд, однако:

- Эффективность метрик зависит от позиции: SPAOM показал значимость для защитников, но не для нападающих;
- Линейные модели не учитывают взаимодействия между игроками, контекст матчей и тактические особенности;

- Дальнейшие исследования могут включать:
  - Моделирование с учётом взаимодействий между метриками;
  - Использование моделей машинного обучения (например, градиентного бустинга);
  - Расширение выборки на несколько сезонов.

Несмотря на упрощения, предложенный подход позволяет перейти от интуитивных оценок к количественному анализу состава команд.

Кроме того, хочется отметить, что было очень интересно окунуться в мир реальных данных на столь прикладном уровне и проверить математические методы и теории. Изначально мы планировали использовать данные из бейсбола, как это было в фильме, но пришли к тому, что никто из нас не разбирается в этом виде спорта, поэтому взяли что-то более народное, а именно футбол)