



Блок рыночной привлекательности: 3 метрики

Ниже описано, какие признаки мы добавляем, что они означают и как они вычисляются.

1) Market Size Score — оценка размера целевого рынка

Что отражает:

Показывает, насколько большой коммерческий потенциал у заболевания — сколько денег потенциально приносит рынок препаратов для этой нозологии.

Откуда берётся:

Мы используем столбец `market_size_million`, который должен отражать эпидемиологию:

- распространённость болезни,
- объём продаж аналогичных препаратов,
- стоимость терапии.

Как считаем:

1. Нормализуем `market_size_million` в диапазон 0..1.
2. Разбиваем на категории:
 - 0 — маленький рынок
 - 1 — средний рынок
 - 2 — большой рынок

Почему это нужно:

Большой рынок = больше потенциальных пациентов = выше общий потолок прибыли.

2) Competition Score — уровень конкуренции на рынке

Что отражает:

Показывает, насколько сложным будет вывод препарата на рынок, если в этой болезни уже много аналогов.

Откуда берётся:

Зависит от того, **сколько других препаратов на рынке имеют ту же комбинацию:**

- **indication** — то же заболевание
- **target** — та же молекулярная мишень

Логика:

Если у болезни 10 препаратов с тем же механизмом действия → конкуренция высокая.

Если таргет редкий → конкуренция низкая.

Как считаем:

1. Считаем частоты по (**indication**, **target**) во всём df.
2. Нормализуем в диапазон 0..1.
3. Категоризируем на 0/1/2:
 - **0 — низкая конкуренция**
 - **1 — умеренная**
 - **2 — высокая**

Почему это важно:

Высокая конкуренция снижает шанс занять долю рынка и снизит финансовый потенциал.

Замечание:

Использование **uniqueness_score** здесь не нужно, потому что конкуренция — это **объективное количество существующих аналогов**, а уникальность — субъективная “новизна”.

3) Profit Potential — *потенциальная прибыльность препарата*

Что отражает:

Это итоговая финансовая метрика:
насколько выгодно создавать этот препарат.

Как считаем:

Используем формулу:

$$\text{profit} = \text{market_size} \times \text{market_share} - \text{dev_cost}$$
$$\text{profit} = \text{market_size} \times \text{market_share} - \text{dev_cost}$$

Где:

- **market_size** — объём рынка (в миллионах \$)
- **market_share** — доля рынка, которую можем занять
- **dev_cost** — стоимость R&D

Откуда берём market_share?

В симуляции мы используем простую формулу:

$$\text{market_share} = \frac{1}{1 + \text{competition}}$$
$$\text{market_share} = \frac{1}{1 + \text{competition}}$$

То есть:

- конкуренция низкая → препарат захватывает большую долю рынка
- конкуренция высокая → доля рынка маленькая

Пример:

competition_score	market_share
0 (низкая)	1.00
1 (средняя)	0.50

2 (высокая)

0.33

Как это используется:

Мы получаем столбец:

`profit_million_usd`

и можем превратить его в категории 0/1/2 для финальной интерпретации.

Как эти три метрики взаимодействуют

1. **Market Size Score** — определяет общую ёмкость рынка.
Большой рынок = больше денег.
 2. **Competition Score** — снижает потенциальную долю рынка, даже если рынок огромный.
 3. **Profit Potential** — это уже *реальная выгода*, учитывающая:
 - сколько денег на рынке в целом,
 - какую долю мы можем занять,
 - сколько стоит разработка.
-



Короткое резюме для PDF (максимально компактная версия)

Мы рассчитываем три метрики, важные для рыночной оценки лекарственного кандидата:

1. **Market Size Score** — показывает, насколько большой рынок заболевания.
Вычисляется по `market_size_million`, нормируется и разбивается на категории 0/1/2.

2. **Competition Score** — отражает уровень конкуренции среди аналогичных препаратов.
Основан на частоте встречаемости пары (indication, target).
Низкая конкуренция → 0, высокая → 2.
3. **Profit Potential** — итоговая финансовая привлекательность.
Считается как:
$$\text{profit} = \text{market_size} \times \text{market_share} - \text{dev_cost}$$
$$\text{profit} = \text{market_size} \times \text{market_share} - \text{dev_cost}$$
где $\text{market_share} = 1 / (1 + \text{competition})$.

Клинические испытания любого лекарственного препарата проходят **три основных фазы**, каждая из которых проверяет разные аспекты безопасности и эффективности.

Фаза I — безопасность (Safety)

Цель: проверить, безопасен ли препарат для человека.

Участники: 20–100 здоровых добровольцев.

Фокус:

- базовая токсичность,
- переносимость доз,
- побочные эффекты.

Как мы считаем вероятность успеха:

Используем уже вычисленный в системе показатель **toxicity_score** (0/1/2).

Чем выше токсичность — тем ниже шанс пройти Phase I.

Фаза II — эффективность (Efficacy)

Цель: показать, работает ли препарат для лечения конкретного заболевания.

Участники: сотни пациентов.

Фокус:

- подтверждение эффективности,

- подбор рабочей дозировки.

Как мы считаем вероятность успеха:

Используем показатель **efficiency_score** (0/1/2).

Чем выше он — тем выше вероятность успешно пройти Phase II.

Фаза III — масштабируемость и риски (Large trials)

Цель: доказать, что препарат эффективен и безопасен на больших выборках.

Участники: тысячи пациентов.

Фокус:

- сравнение с существующими аналогами (конкуренция),
- расширенная безопасность,
- устойчивость результата для разных групп.

Как мы считаем вероятность успеха:

Берём комбинацию:

- **efficiency_score**,
- **toxicity_score**,
- **competition_score** (сколько препаратов-аналога уже на рынке).

Phase III — самая дорогая и рисковая фаза, поэтому учитываются сразу несколько метрик.

Как мы преобразуем баллы в проценты успеха

Мы используем нормированную формулу:

Phase I

$p_{\text{phase1}} = (1 - \text{toxicity_score} / 2) * 0.55 + 0.45$

Диапазон примерно 45–100%.

Phase II

```
p_phase2 = (efficiency_score / 2) * 0.6 + 0.2
```

Диапазон примерно 20–80%.

Phase III

```
p_phase3 = (  
    0.45 * (efficiency_score / 2) +  
    0.35 * (1 - toxicity_score / 2) +  
    0.20 * (1 - competition_score / 2)  
)
```

Диапазон примерно 10–90%.

Почему проценты?

Такой формат интуитивно читается и хорошо подходит для дашборда / отчёта.

Почему мы считаем фазы именно так

- ✓ Токсичность определяет безопасность → Phase I
- ✓ Эффективность определяет рабочее действие → Phase II
- ✓ На больших выборках влияет всё: эффективность, токсичность и конкуренты → Phase III

Мы не обучаем три новые ML-модели, потому что:

1. **Это привело бы к повторному обучению токсичности и эффективности,** которые уже спрогнозированы ранее.
2. **Настоящие датасеты по фазам гигантские,** и в этом проекте их имитация бессмысленна.
3. **Проектная задача — построить симуляцию,** а не реальный ML-пайплайн FDA.

Мы используем уже имеющиеся предсказанные метрики (efficiency, toxicity, competition)
— это логично и соответствует реальным принципам оценки риска.