Факультет социальных наук

Вычислительные социальные науки

Москва
2025

Исследование взаимосвязи рейтинга игроков в настольный теннис с их характеристиками и уровнем популярности

Investigation of the Relationship between Table Tennis Players' Ratings, Their Characteristics and Popularity Level

Руководитель курсовой работы: Паклина София Николаевна,
кандидат экономических наук,
младший научный сотрудник Международной лаборатории экономики нематериальных активов
Выполнила: Резепова Софья Владимировна

Актуальность

Google Trends
открытые, регулярно обновляемые данные, широкий временной охват

Практическое применение:
Тренеры - оценка коммерческого потенциала
Маркетологи - стратегии продвижения
Агенты - обоснование контрактных решений

«Анна Курникова была второй по доходам среди теннисисток в 2003 году несмотря на то, что никогда не поднималась выше восьмого места в рейтинге»
(Yucesoy & Barabási, 2016)

Исследовательский вопрос

В какой степени показатель популярности способен предсказывать результативность?

Цель

Построение модели, позволяющей предсказывать результативность игроков в настольный теннис на основе их индивидуальных характеристик

Задачи

1. Провести обзор релевантной литературы для определения ключевых детерминант спортивной результативности.
2. Построить модели машинного обучения для предсказания результативности игроков на основе демографических и игровых характеристик
3. Оценить вклад показателя популярности, измеренного через Google Trends, в объяснение результативности
4. Проанализировать форму связи между популярностью и результативностью, включая возможные нелинейные эффекты
5. Сравнить предсказательное качество моделей с различными наборами переменных

Обзор литературы

Детерминанты результативности игроков в теннис

«Are performance trajectories associated with relative age in French top 100 youth table tennis players?»
Показывает, что спортсмены, родившиеся в первом квартале года, чаще добиваются успеха в юношеском возрасте (эффект относительного возраста)
(Faber et al., 2020)

«Performance indicators in table tennis: a review of the literature»
Обзорная статья, систематизирующая ключевые переменные (возраст, пол, хват, стиль игры и др.), используемые для анализа результативности
(Malagoli Lanzoni et al., 2012)

«Determinants for table tennis performance in elite Scottish youth players using a multidimensional approach»
Описывает, как возраст, пол и опыт тренировок влияют на рейтинг игроков, выявляя сильную корреляцию между стажем и результатами.
(Doherty et al., 2018)

Обзор литературы

Популярность как детерминанта спортивной результативности

Двойственность публичного внимания:
- эффект Пигмалиона
- эффект суперзвезды
- эффект аутсайдера

«Untangling Performance from Success»
взаимосвязь между популярностью и результативности нелинейна
(Yucesoy & Barabási, 2016)

«Social status and sport: A study of young Norwegians»
спортивная результативность повышает популярность как форму социального статуса среди подростков
(Seippel & Bergesen Dalen, 2024)

«The prognostic relevance of psychological factors with regard to participation and success in table-tennis»
ориентация на внешнее признание повышает риск выгорания у игроков в настольный теннис
(Martinent, 2018)

«Successful, sexy, popular: Athletic performance and physical attractiveness as determinants of public interest in male and female soccer players»
физическая привлекательность влияет на популярность в футболе, даже при прочих равных
(Mutz & Mutz, 2014)

Обзор литературы

Google Trends как инструмент оценки популярности

«Measuring the popularity of football players with Google Trends»
показатели Google Trends улучшают прогноз рыночной стоимости футболистов
« … ошибка модели 3 градиентного бустинга с использованием этих признаков составила 11,507,727 евро, тогда как при отсутствии любых показателей популярности она достигала 18,000,361 евро. В итоге, включение популярности позволило снизить ошибку более чем на 6,492,634 евро»
(Malagón-Selma, 2023)

«Using Google Trends as a proxy for occupant behavior to predict building energy consumption»
(Fu & Miller, 2022)

«Applying Google Trends' search popularity indicator to professional cycling»
(Genoe, Rousseau & Rousseau, 2021)

«Forecasting sports popularity: Application of time series analysis»
(Miller, Schwarz & Talke, 2017)

Гипотезы

H1: Модели машинного обучения позволяют предсказать результативность игроков на основе их демографических и игровых характеристик

H2: Учет результативности в предыдущем периоде повышает точность прогнозирования текущей результативности

H3: Добавление фактора популярности, измеренного по данным Google Trends, увеличивает предсказательную силу модели

H4: Влияние популярности на результативность носит нелинейный характер и может быть охарактеризовано квадратичной зависимостью

Дизайн исследования

Описание спецификаций

Спецификация	Переменные
1	Assoc, Gender, Age, Playing hand, Grip
2	+ lag_Points
3	+ lag_Popularity
4	+ lag_Popularity_sq

Алгоритмы машинного обучения:

1. Линейная регрессия
2. Дерево решений
3. Случайный лес
4. Дополнительные деревья
5. Градиентный бустинг

Методология

Метрики качества

MAE (Mean Absolute Error, средняя абсолютная ошибка)

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

R^2 (коэффициент детерминации)

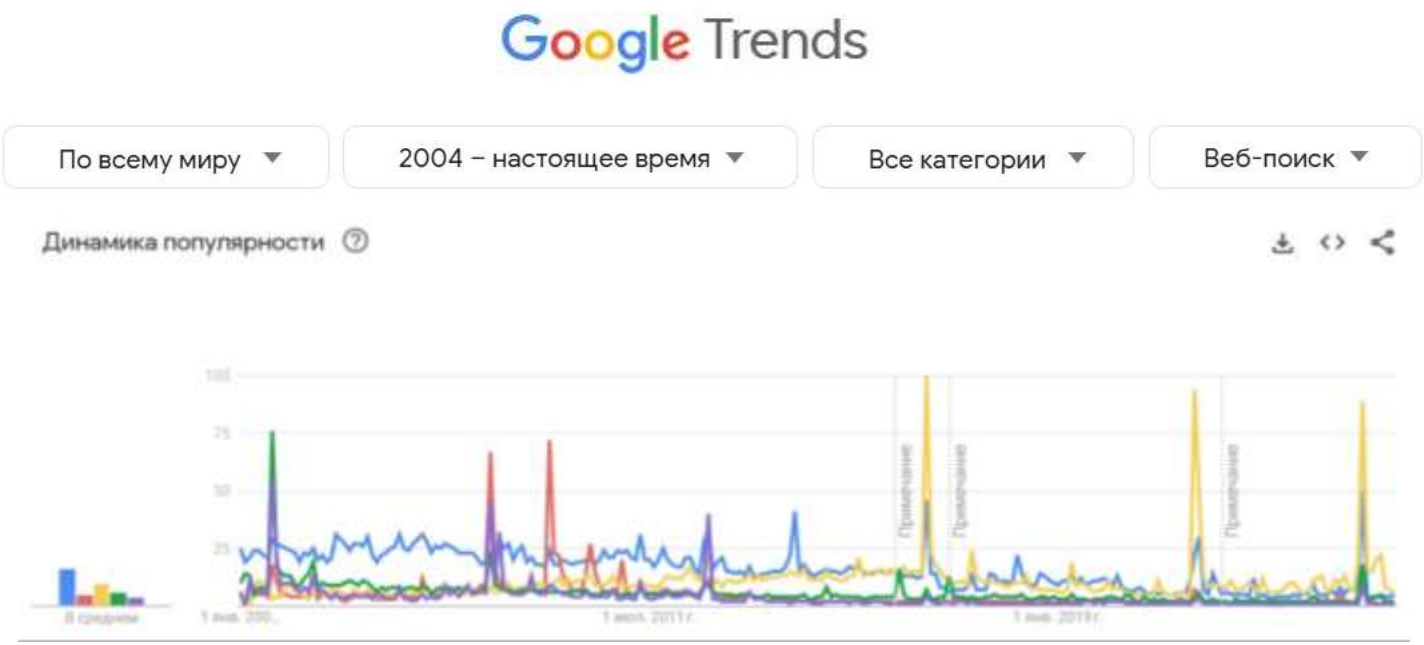
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Данные

Данные **Google Trends** в период с января **2004** года по февраль **2025** года

```python
import pandas as pd
df = pd.read_excel('BOOK.xlsx')
for i in range(224, 0, -5):
    if i <4:
        break
    norm_factor_1 = df[i] / df[i+1]
    df[i+1] = df[i+1] * norm_factor_1
    df[i+2] = df[i+2] * norm_factor_1
    df[i+3] = df[i+3] * norm_factor_1
    df[i+4] = df[i+4] * norm_factor_1
    df[i+5] = df[i+5] * norm_factor_1
df.to_excel('BOOK_updated.xlsx')
```

```python
import pandas as pd
df = pd.read_excel('BOOK_updated.xlsx')
for i in range(9, 225, 5):
    norm_factor = df[i] / df[i+1]
    df[i+1] = df[i+1] * norm_factor
    df[i+2] = df[i+2] * norm_factor
    df[i+3] = df[i+3] * norm_factor
    df[i+4] = df[i+4] * norm_factor
    df[i+5] = df[i+5] * norm_factor
df.to_excel('FINAL.xlsx')
```



1. Сортировка по среднему рейтингу
2. Разделение на группы по 5 человек, с частичным перекрытием между соседними группами
3. Нормализация – приведение к размерности первой, самой сильной, группы

Данные

Данные с **ITTF*** в период с **2001** года по апрель **2020** года

Переменная	Категория	Количество
Пол	Мужчина (male)	182
	Женщина (female)	182
Игровая рука	Правая (right–handed)	282
	Левая (left–handed)	82
Хватка	Европейская (shakehand)	331
	Пенхолд (penhold)	33
Ассоциация	Китай (CHN)	63
	Япония (JPN)	49
	Корея (KOR)	31
	Германия (GER)	28
	Гонконг (HKG)	19

Переменная	Минимум	Максимум	Среднее	Медиана
Возраст	10.00	57.00	26.03	25.00
Очки	9.00	17915.00	4186.65	2304.00

Итог:

44 559 наблюдений

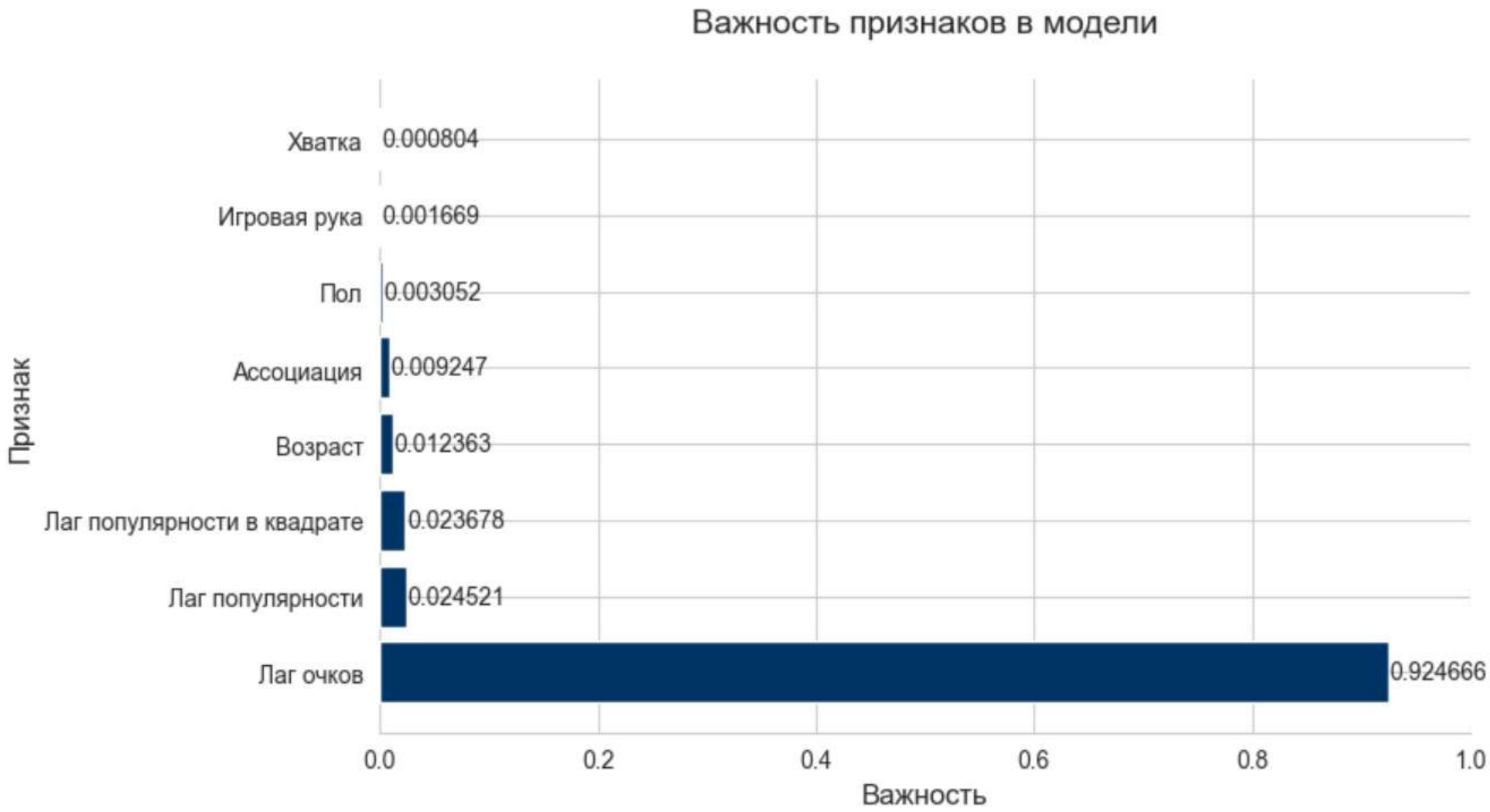
364 игрока

*International Table Tennis Federation URL: https://www.ittf.com/

Результаты

Спецификация	Переменные	Лучшая модель	MAE	R^2
1	Ассоциация, пол, возраст, игровая рука, хватка	Дополнительные деревья	1783.85	0.44
2	+ Лаг очков	Случайный лес	321.86	0.90
3	+ Лаг популярности	Случайный лес	250.91	0.9358
4	**+ Лаг популярности в квадрате**	**Случайный лес**	**251.39**	**0.9375**

Результаты


Важность признаков в модели
Признак
Хватка
0.000804
Игровая рука
0.001669
Пол
0.003052
Ассоциация
0.009247
Возраст
0.012363
Лаг популярности в квадрате
0.023678
Лаг популярности
0.024521
Лаг очков
0.924666
0.0
0.2
0.4
0.6
0.8
1.0
Важность

Заключение

Ограничения:

1. Проблема эндогенности
2. Особенности данных Google Trends
3. Неучтенные факторы

Перспективы дальнейших исследований:

1. Причинно-следственный анализ

Инструментальные переменные:

- Блокировка Google/соцсетей в Китае (с 2014 года) – как экзогенный шок
- Смена гражданства игроками – влияние на доступ к глобальным медиа

2. Модели одновременных уравнений (simultaneous equations models) для глубокого понимания структуры взаимозависимостей

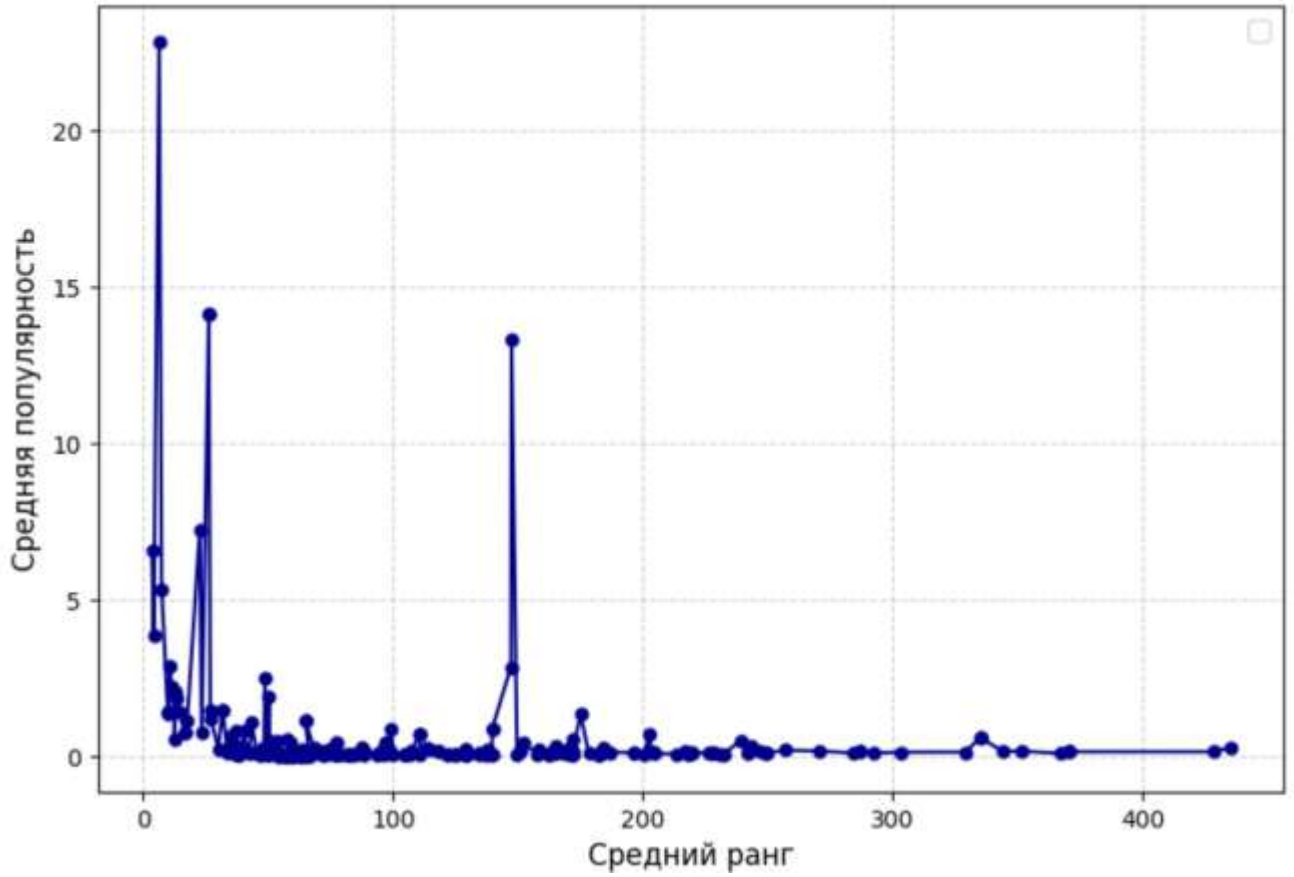
Список использованной литературы

1. Doherty S. A. P. et al. Determinants for table tennis performance in elite Scottish youth players using a multidimensional approach: A pilot study // High Ability Studies. – 2018. – Т. 29. – № 2. – С. 241–254.
2. Fu C., Miller C. Using Google Trends as a proxy for occupant behavior to predict building energy consumption // Applied Energy. – 2022. – Т. 310. – С. 118343.
3. Genoe A., Rousseau R., Rousseau S. Applying Google Trends' search popularity indicator to professional cycling // Journal of Sports Economics. – 2021. – Т. 22. – № 4. – С. 459–485.
4. Malagoli Lanzoni I. et al. Performance indicators in table tennis: a review of the literature // International Journal of Table Tennis Sciences. – 2012. – Т. 7. – С. 71–75.
5. Martinent G. et al. The prognostic relevance of psychological factors with regard to participation and success in table-tennis // Journal of Sports Sciences. – 2018. – Т. 36. – № 23. – С. 2724–2731.
6. Miller R., Schwarz H., Talke I. S. Forecasting sports popularity: Application of time series analysis // Academic Journal of Interdisciplinary Studies. – 2017. – Т. 6. – № 2. – С. 75–82.
7. Mutz M., Meier H. E. Successful, sexy, popular: Athletic performance and physical attractiveness as determinants of public interest in male and female soccer players // International Review for the Sociology of Sport. – 2014. – Т. 51. – № 5. – С. 567–580.
8. Pradas F. et al. Analysis of specific physical fitness in high-level table tennis players – sex differences // International Journal of Environmental Research and Public Health. – 2022. – Т. 19. – № 9. – С. 5119.
9. Seippel Ø., Bergesen Dalen H. Social status and sport: A study of young Norwegians // International Review for the Sociology of Sport. – 2024. – Т. 59. – № 3. – С. 343–360.
10. Yucesoy B., Barabási A. L. Untangling performance from success //EPJ Data Science. – 2016. – Т. 5. – №. 1. – С. 1-10.


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •

Приложение. Нелинейность

График зависимости между средним рангом и средней популярностью



Приложение. Алгоритмы машинного обучения

Дерево решений (Decision Tree Regressor). Алгоритм бинарного дерева решений начинается с вычисления предиката в корневой вершине – функции, которая возвращает 0 или 1. Если результат равен 0, переход осуществляется в левую дочернюю вершину; если 1 – в правую. На каждой вершине проверяется предикат, процесс продолжается до достижения листовой вершины, в которой записан итоговый ответ модели – прогноз (численное значение) в случае регрессии или вектор вероятностей по классам в случае классификации. Параметры модели включают min_samples_split=2, max_depth=None – это означает, что дерево может расти до тех пор, пока не исчерпаются данные для разделения. Однако деревья склонны к переобучению, что приводит к высокой дисперсии модели. Для решения этой проблемы используются ансамбли деревьев.

Random Forest: ансамбль деревьев, обучающихся на бутстрап-выборках (bootstrap=True) и случайных подмножествах признаков (max_features=1.0). Снижает дисперсию и переобучение.

Extra Trees Regressor: похож на RF, но без бутстрапа (bootstrap=False) и с полностью случайными порогами при разбиении. Более устойчив к шуму, имеет меньше дисперсии, но выше смещение.

Градиентный бустинг (Gradient Boosting Regressor) – это метод последовательного построения ансамбля деревьев. В отличие от случайного леса, где деревья обучаются независимо, бустинг строит деревья последовательно. Каждый новый элемент ансамбля обучается на остатках предыдущих моделей, то есть на разности между фактическими значениями Points и предсказанными. На каждом шаге m строится дерево $hm(x)$, минимизирующее градиент функции потерь.