

Шихов Андрей Сергеевич, магистрант
ФГБОУ ВО «Вятский государственный университет»

СРАВНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ СТОИМОСТИ ЖИЛОЙ НЕДВИЖИМОСТИ

Аннотация. В статье рассматриваются алгоритмы машинного обучения как инструмент для повышения точности прогнозирования стоимости жилой недвижимости. Рассмотрены основные алгоритмы машинного обучения, такие как XGBoost, CatBoost и нейронные сети.

Ключевые слова: Машинное обучение, прогнозирование стоимости недвижимости, алгоритмы прогнозирования, нейронные сети, математические модели.

Анализ и прогнозирование стоимости недвижимости являются ключевыми задачами в сфере экономики и финансов, находящими применение в широком спектре областей, начиная от инвестиционного анализа и заканчивая разработкой государственной жилищной политики. Стоимость недвижимости – это комплексная характеристика, зависящая от множества факторов, таких как местоположение, инфраструктура, экономическая ситуация в регионе, демографические показатели и другие. Прогнозирование стоимости недвижимости позволяет не только оценивать текущее состояние рынка, но и предсказывать его динамику, что является важным инструментом для принятия решений как для частных инвесторов, так и для бизнеса и государственных органов.

Традиционные методы прогнозирования стоимости недвижимости, такие как регрессионный анализ или методы на основе экспертных оценок, обладают рядом ограничений, связанных с точностью и способностью учитывать сложные нелинейные зависимости между множеством факторов. С развитием машинного обучения и нейронных сетей появляется возможность значительно повысить точность и адаптивность прогнозных моделей, благодаря их способности к обучению на больших объемах данных и обработке сложных взаимосвязей между различными параметрами.

В наши дни в моделях прогнозирования стоимости недвижимости используются современные алгоритмы машинного обучения. Каждый из алгоритмов имеет свои особенности, преимущества и недостатки, которые важно учитывать при выборе подхода для решения задачи. Рассмотрим основные алгоритмы подробно.

XGBoost (Extreme Gradient Boosting) – это мощный алгоритм градиентного бустинга, который строит деревья решений последовательно, минимизируя ошибку на каждом шаге. Он использует методы регуляризации (L1 и L2) для предотвращения переобучения и автоматически обрабатывает пропущенные значения. XGBoost также поддерживает пользовательские функции потерь и метрики оценки. Преимуществами данного алгоритма являются: высокая точность и производительность, поддержка регуляризации для борьбы с переобучением, возможность работы с пропущенными значениями, гибкость в настройке и использовании пользовательских функций. Недостатками данного алгоритма являются: сложная настройка гиперпараметров, требование больших вычислительных ресурсов, меньшая интерпретируемость результатов.

CatBoost – это еще один алгоритм градиентного бустинга, специально разработанный для работы с категориальными признаками. Он автоматически обрабатывает категориальные переменные без необходимости их предварительного кодирования, что упрощает подготовку данных. CatBoost также использует упорядоченный boosting для борьбы с переобучением и улучшения обобщающей способности модели. Преимуществами данного алгоритма являются: эффективная работа с категориальными признаками, автоматическая обработка пропущенных значений, устойчивость к переобучению благодаря упорядоченному бустингу, высокая точность на задачах с разнородными данными. Недостатками данного алгоритма являются: медленная скорость работы по сравнению с XGBoost, требование большей памяти из-за хранения истории данных, меньшая гибкость в настройке по сравнению с XGBoost.



Нейронные сети – это мощный класс алгоритмов, вдохновленный работой человеческого мозга. Они состоят из слоев нейронов, которые обрабатывают входные данные и передают информацию через веса и активационные функции. Нейронные сети [1] могут быть использованы для решения широкого круга задач, включая классификацию, регрессию, обработку изображений, текста и звука. К преимуществам нейронных сетей можно отнести: способность моделировать сложные нелинейные зависимости, высокая точность на больших объемах данных, подходят для задач любого типа, возможность автоматического извлечения признаков. К недостаткам нейронных сетей можно отнести: требуют больших объемов данных для обучения, высокая вычислительная сложность и затраты на обучение, сложность интерпретации результатов, требуют тщательной настройки гиперпараметров и архитектуры.

Несмотря на широкое применение машинного обучения в оценке недвижимости, большинство моделей сталкиваются с рядом ограничений [3]. Во-первых, зависимость от качества данных – пропуски, аномалии и субъективность в описании снижают точность прогнозов [4]. Во-вторых, модели не учитывают экономические факторы, разработанные для конкретной страны, они не могут быть использованы для прогнозирования стоимости недвижимости других стран. Кроме того, многие модели игнорируют пространственные зависимости (например, ценовую динамику в конкретных районах), что критично для недвижимости [5]. В условиях постоянно меняющейся экономической обстановки в стране модели быстро теряют актуальность [6] и требуют обновления обучающих данных с учетом текущей ситуации на рынке [7].

Для оценки алгоритмов машинного обучения был проведен эксперимент, в котором все три модели были обучены на одном наборе данных. Набор данных содержит в себе 4995 строк и содержит факторы единицы жилой недвижимости и внешние макроэкономические факторы. В качестве сравнительной метрики был выбран коэффициент детерминации [8]. R-квадрат – коэффициент детерминации. Этот показатель оценивает долю дисперсии зависимой переменной, которая может быть объяснена моделью. Высокое значение R-квадрата указывает на то, что модель хорошо объясняет вариацию цен на недвижимость.

В результате эксперимента были получены следующие значения коэффициента детерминации. Для алгоритма XGBoost коэффициент детерминации на тренировочной выборке составил 0,951, а на тестовой 0,674. У алгоритма Catboost коэффициент детерминации на тренировочной выборке составил 0,895, а на тестовой 0,793. Для эксперимента был использован многослойный перцептрон и его результат на тренировочной выборке составил 0,795, а на тестовой 0,648. По результатам эксперимента можно сделать выводы, что на тестовой выборке лучший результат показывает алгоритм XGBoost, а на тренировочной лучший результат показывает алгоритм Catboost. Так как результат на тренировочной выборке более важен чем на тестовой, то можно смело утверждать, что алгоритм Catboost наиболее предпочтителен в задачах прогнозирования стоимости жилой недвижимости.

В заключение можно отметить, что в сравнении с альтернативами (XGBoost, XGBoost, нейросети) CatBoost показывает сопоставимую или лучшую точность при меньших затратах на предобработку данных, что подтверждается тестами на реальных датасетах с параметрами недвижимости и макроэкономическими показателями (ипотечные ставки, инфляция). Кроме того, CatBoost обеспечивает быстрое обучение и предсказание даже на больших объемах данных, поддерживает GPU-ускорение и легко интегрируется в производственные пайплайны через экспорт обученной модели. Быстрое обучение позволяет проводить эксперименты с обновлением модели: подбирать веса, менять глубину обучения и добавлять новые правила в модель. Это позволяет быстро повышать эффективность модели.

Список литературы:

1. Методы машинного обучения в исследовании рынка жилой недвижимости. – URL: <https://cyberleninka.ru/article/n/metody-mashinnogo-obucheniya-v-issledovanii-rynka-zhiloy-nedvizhimosti/viewer> – электронный.



2. Стерник, Г. М. Развитие и совершенствование методов прогнозирования на рынке жилой недвижимости / Г. М. Стерник, С. Г. Стерник, А. В. Свиридов // Урбанистика и рынок недвижимости. – 2014. – № 1. – С. 53–93.
3. Стерник Г.М., Стерник С.Г. Пять макроэкономических законов функционирования рынка недвижимости как неотъемлемой составляющей глобального финансового рынка в транзитивной экономике // Финансовая аналитика: проблемы и решения. – 2010. – № 11 (35). – С. 15–29.
4. Статистический анализ динамики стоимости квартир на вторичном рынке жилой недвижимости города Москвы. – URL: <https://cyberleninka.ru/article/n/statisticheskiy-analiz-dinamiki-stoimosti-kvartirna-vtorichnom-rynke-zhiloy-nedvizhimosti-goroda-moskvy/viewer> – Текст: электронный. – (Дата обращения 16.03.2025)
5. Стерник Г. М., Стерник С. Г. Методология прогнозирования российского рынка недвижимости. Ч. 1. Основные допущения, ограничения и рабочие гипотезы // Механизация строительства. – 2013. – № 8 (830). – С. 53–63.
6. Массовая оценка объектов недвижимости на основе технологий машинного обучения. Анализ точности различных методов на примере определения рыночной стоимости квартир. – URL: <https://cyberleninka.ru/article/n/massovaya-otsenka-obektov-nedvizhimosti-naosnove-tehnologiy-mashinnogo-obucheniya-analiz-tochnosti-razlichnyh-metodovna-primere/viewer> – Текст: электронный. – (Дата обращения 18.03.2025)
7. Mohd, T. Rapid Modelling of Machine Learning in Predicting Office Rental Price / T. Mohd, M. Harussani, S. Masrom – Text: direct // International Journal of Advanced Computer Science and Applications. 2022. Vol. 13. № 12.
8. R2_score – Gallery examples: L1-based models for Sparse Signals Linear Regression Example Non-negative least squares Failure of Machine Learning to infer causal effects Effect of transforming the targets in r... – URL: https://scikitlearn/stable/modules/generated/sklearn.metrics.r2_score.html – Текст: электронный. (Дата обращения 12.03.2025)

