УДК 004.62, 330.43

ОЦЕНКА СТОИМОСТИ ОБЪЕКТОВ НЕДВИЖИМОСТИ ПРИ ПОМОЩИ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ: ИЗУЧЕНИЕ РЫНКА НЕДВИЖИМОСТИ ГОРОДА ДУБНЫ

Куликов Дмитрий Леонидович

Кандидат физико-математических наук, доцент;
ГБОУ ВО МО «Университет «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail:kulikov.undubna@gmail.com.

Данная работа посвящена использованию методов машинного обучения для анализа реальных данных об объектах недвижимости города Дубна (Московская область). Основной целью работы является сравнения прогностических характеристик различных методов машинного обучения (линейная регрессия, дерево решений, случайный лес, градиентный бустинг). База данных состоит из выборки из 800 записей цен на квартиры. Результаты анализа показывают, что градиентный бустинг и случайный лес показали лучшие результаты, чем другие модели при моделировании цен на жилье. В целом мы приходим к выводу, что методы машинного обучения могут предоставить полезный набор инструментов для получения информации о рынках жилья.

Ключевые слова: машинное обучение, оценка стоимости недвижимости, линейная регрессия, дерево решений, случайный лес, градиентный бустинг.

REAL ESTATE VALUATION WITH THE HELP OF MACHINE LEARNING ALGORITHMS: A STUDY OF THE REAL ESTATE MARKET OF THE CITY OF DUBNA

Dmitriy Kulikov

Candidate of physical and mathematical Sciences, associate professor;
Dubna State University,
Institute of the system analysis and management;
141980, Moscow region, Dubna, Universitetskaya str.,19;
e-mail: kulikov.undubna@gmail.com.

This work is devoted to the machine learning methods usage for on real estate data analysis in Dubna (Moscow region). The main purpose of the work is to compare the prognostic characteristics of various machine learning methods (linear regression, decision tree, random forest, gradient boosting). The database for our analysis consists of a sample of 800 apartment price records. The results of the analysis show that gradient boosting and random forest showed better results than other models in housing price modeling. In General, we conclude that machine learning methods can provide a useful set of tools for obtaining information about housing markets.

Keywords: machine learning, real estate price evaluation, linear regression, decision tree, random forest, gradient boosting.

Введение

В данной статье анализируются методы машинного обучения для анализа рынка жилой недвижимости в городе Дубна, Московская область. Обычно для исследования рынка жилья используются эконометрические подходы, то есть, по факту, это означает моделирование цен на недвижимость при помощи метода наименьших квадратов (МНК), который в некоторых руководствах рассматривается как «наиболее простой статистический метод, используемый экономистами» [5]. В большинстве исследований используется либо этот метод, либо его модификации.

В настоящее время подход с использованием методов машинного обучения для анализа рынков жилья используется для все большего числа городов по всему миру [2]. При этом используется одно из основных свойств методов машинного обучения к автоматическому поиску шаблонов и

закономерностей в исходных данных. При этом точность оценки может значительно варьироваться как в зависимости от исходных данных, так и от используемых алгоритмов машинного обучения.

Данная работа посвящена исследованию, направленному на проведение сравнительного анализа различных наиболее часто используемых методов машинного обучения для решения задачи прогнозирования стоимости жилья на конкретных данных (на примере оценки стоимости квартир в городе Дубна Московской области). Основываясь на существующих подходах, рассмотренных в разделе 3.2, мы будем использовать подход машинного обучения для следующей задачи: сравнить производительность дерева решений, случайного леса и градиентного бустинга с классической множественной линейной регрессией с при прогнозировании цен на жилую недвижимость в Дубне.

1. Обзор методов оценки недвижимости

Различные международные экономические события последнего десятилетия вызвали растущий интерес к исследованиям рынков жилой недвижимости. Кризис низкокачественных ипотечных кредитов в США и пузыри на рынке жилья в Европе подчеркнули необходимость получения более полной информации о поведении рынков жилья как для принятия краткосрочных, так и долгосрочных решений.

Большинство методов оценки стоимости недвижимости на основе параметров объекта недвижимости прямо или косвенно используют гедонистическую гипотезу, сформулированную Розеном [4] в 1974 году, которая предположила, что товары оцениваются потребителями на основе их характеристик или атрибутов. То есть, фактически, исследователями используется многофакторная регрессионная модель, где в качестве независимой переменной (целевой) выступает стоимость объекта недвижимости, а в качестве зависимых: количественные (площадь, число комнат, этаж и т.д.) и качественные (престижность района, наличие инфраструктуры, соседи и т.д.) характеристики. Наиболее простые линейные модели [3] учитывают такие числовые характеристики, как площадь объекта, возраст здания, площадь участка, расстояние до центра города и т. д. При этом предположение о том, что зависимость между параметрами объекта недвижимости и ценой является линейной, обычно неверно. И, как следствие, многие исследователи используют другие модели (например, логарифмическую [6]). При этом вопрос выбора той или иной модели зависимости между переменными является до сих пор открытым - так, в частности в работе [1] утверждается, что «одна из самых важных проблем в эконометрике состоит в том, что мы никогда не уверены в форме и / или спецификации уравнения, которое мы хотим оценить».

Таким образом, в существующих исследованиях рассматриваются различные методы регрессии для прогнозирования цен на жилье. В текущей работе мы пытаемся выяснить, какие из существующих методов, основанных на машинном обучении, лучше всего подходят для конкретной реальной (не модельной) задачи оценки объектов недвижимости на рынке города Дубны по существующим параметрам.

2. База исследования

Мы строим различные модели с использованием решающего дерева, случайного леса, метода градиентного бустинга и классической множественной линейной регрессии и сравниваем их прогностические характеристики.

База данных для нашего анализа состоит из выборки из 812 записей цен с 10 июля 2018 года по 09 июля 2019 года. Все данные взяты с Портала недвижимости города Дубны [10].

Прогнозируемая зависимая переменная является числовой и представляет собой стоимость объекта недвижимости в рублях. Независимые переменные перечислены и определены в Таблице 1. Типы переменных указаны для языка программирования Python, на котором были реализованы описанные в данной работе методы.

Таблица. 1. Описание исходных данных (переменные и их типы)

Имя переменной	Описание	Тип данных
last-update-date	Дата обновления объявления (Unix-время)	int64
town_reg	Район города (варианты: ЧР, БВ, ИЧ, ЛБ, другое)	object
value	Стоимость квартиры (руб)	int64
area_value	Общая площадь квартиры ($м^2$)	float64
living-space_value	Жилая площадь квартиры ($м^2$)	float64
kitchen-space_value	Площадь кухни ($м^2$)	float64
rooms	Количество комнат в квартире	int64
balcony	Наличие балкона (есть/нет)	object
floor	Этаж, на котором находится квартира	int64
floors-total	Всего этажей в доме	int64
house_type	Тип дома (кирпичный, блочный, монолитный, деревянный, монолитно-кирпичный, панельный)	object
street	Улица	object

Таким образом, в исходном наборе содержится 12 независимых переменных. Из них 8 являются числовыми (типы *int*64 и *float*64) и 4 категориальными (тип *object*: район города, тип дома, наличие балкона и название улицы). Для работы большинства методов машинного обучения необходимо, чтобы все признаки являлись числовыми. Для выполнения данного требования заменим исходные категориальные признаки на соответствующие им числовые аналоги, проведя простую операцию введения взаимного однозначного соответствия между текстовыми и числовыми значениями переменных. Новые дополнительные переменные причислены в таблице 2.

Таблица. 2. Вспомогательные переменные

Имя переменной	Описание	Тип данных
street_cat	Улица (перевод категории в числовое значение)	int64
reg	Район (перевод в числовое значение)	int64
balc	Наличие балкона (перевод в числовое значение)	int64
h_t	Тип дома (перевод в числовое значение)	int64

3. Обработка данных и результаты

3.1 Анализ исходных данных

Прежде чем строить модели зависимости целевой переменной (стоимости объекта недвижимости) от признаков рассмотрим возможные зависимости данных признаков между собой. Для этого построим матрицу взаимной корреляции (рис. 1.)

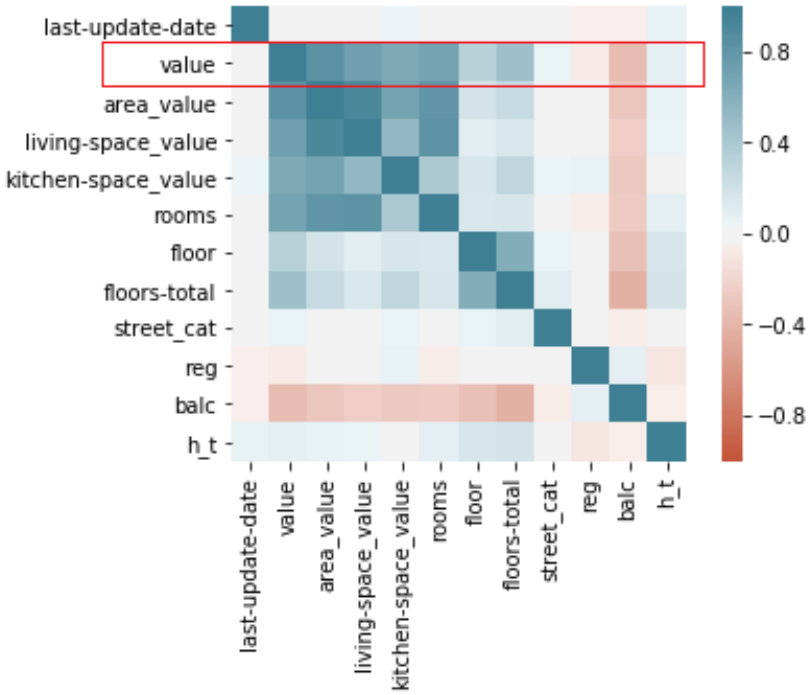

Рис. 1. Матрица взаимной корреляции входных переменных

Анализируя полученные результаты, можно прийти к некоторым выводам (часть из которых была очевидна до проведения данного анализа):

- Стоимость квартиры сильно зависит от площади квартиры.
- Общая площадь зависит от жилой площади, площади кухни и количества комнат.
- Есть зависимость между этажом квартиры и количеством этажей.
- Также существует отрицательная корреляция между стоимостью квартиры и районом города, а также наличием балкона – данный феномен объясняется особенностью сортировки числовых данных в дополнительных введенных переменных и исходными категориальными значениями (так, например, районы города сортируются по алфавиту, а не по стоимости в них жилья).

По итогам проведённого анализа было принято решения для построения первоначальных моделей оценки стоимости объекта недвижимости использовать все исходные переменные.

3.2 Используемые методы

Методы восстановления предполагаемой зависимости целевой переменной от набора независимых переменных, используемые в данной статье, хорошо изучены и активно используются в различных задачах. Приведём их краткое описание.

Линейная регрессия – зависимости одной переменной от другой или нескольких других переменных с линейной функцией зависимости.

Дерево решений: структура представляет собой «листья» и «ветки». Каждый лист представляет собой значение целевой переменной, изменённой в ходе движения от корня по листу. Каждый внутренний узел соответствует одной из входных переменных. Для решения задачи регрессии используется спуск по дерево, где в каждом листе (на основе его условия) принимается решение о переходе в левую или правую часть дерева.

Случайный лес – это множество решающих деревьев. В случае исходной задачи регрессии – ответы деревьев усредняются. Все деревья строятся независимо по следующей схеме: выбирается подвыборка обучающей выборки по ней строится дерево, далее данное дерево расщепляется на основе случайно выбранных признаков из множества всех доступных признаков, выбирается наилучший признак и расщепление по нему. Как правило, процесс останавливается до исчерпания выборки.

Градиентный бустинг – это метод машинного обучения, который создает модель прогнозирования в виде ансамбля слабых моделей прогнозирования, как правило, деревьев решений. Он строит модель поэтапно, как и другие методы бустинга, и обобщает их, позволяя оптимизировать произвольную дифференцируемую функцию потерь.

3.3 Предобработка исходных данных и результаты

Почти всегда при решении задач машинного обучения на реальных (не модельных) данных необходимо сначала провести «очистку» исходных данных от «шума» (от выбросов). Для иллюстрации данного тезиса рассмотрим визуализацию редуцированного набора исходных данные, оставив только две переменные (стоимость и площадь объекта недвижимости) (рис. 2).

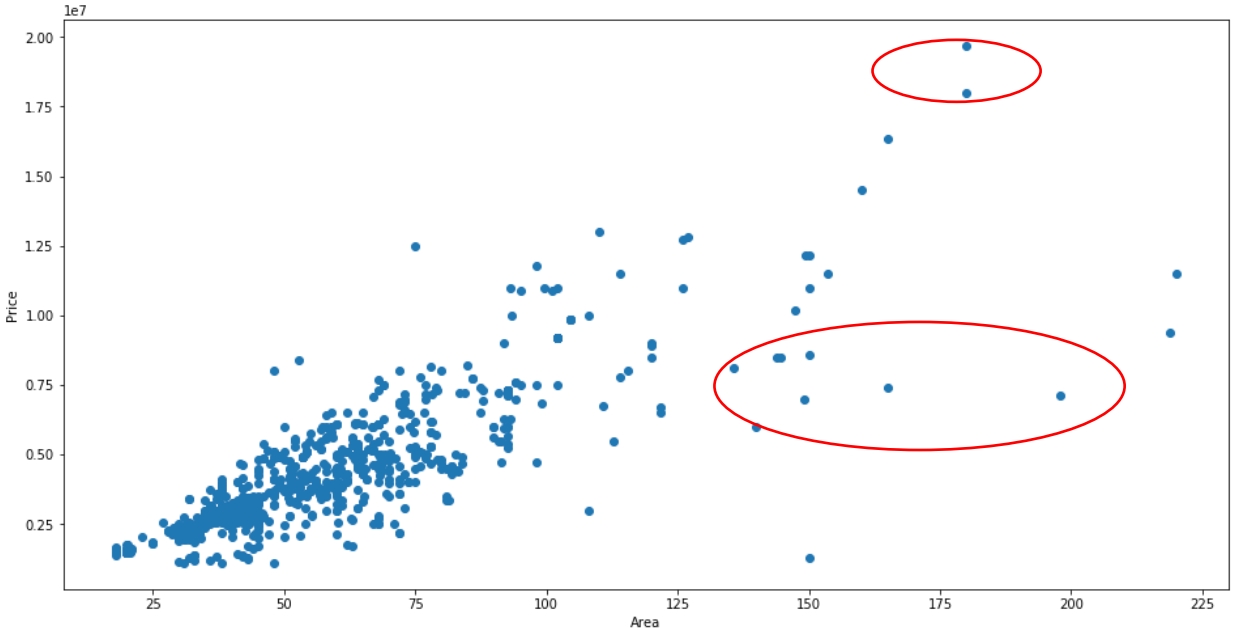

Рис. 2. Исходные данные (зависимость цены от площади)

На данной визуализации видно, что часть исходной выборки является выбросами. Это особенно заметно для объектов недвижимости большой площади и/или с высокой стоимостью. Этот факт может быть объяснён естественными причинами: стоимость нетипичных квартир (с большой площадью) может сильно отличаться от большинства квартир на рынке, так как эти объекты являются эксклюзивными. Но для задачи моделирования зависимости стоимости квартиры от её параметров данные выбросы являются «шумовыми» и от них необходимо избавиться.

Для проведения такой предобработки (удаления «шумных» значений) можно использовать стандартизованнаю оценку (z-оценку) исходных данных (для всех исходных признаков). И удалять те объекты обучающей выборки, для которых (хотя бы по одному из признаков) полученная z-оценка превышает пороговое значение. Для исходной задачи таким пороговым значением является 3 – при этом исходная выборка прорежена не сильно (осталось 727 объектов из исходных 812), при этом точность работы всех моделей повысилась на 20-30% (см. Таблица 2.).

На рис. 3 приведены обработанные данные. На данной визуализации видно, что большая часть выбросов, заметных на рис. 2, была удалена.

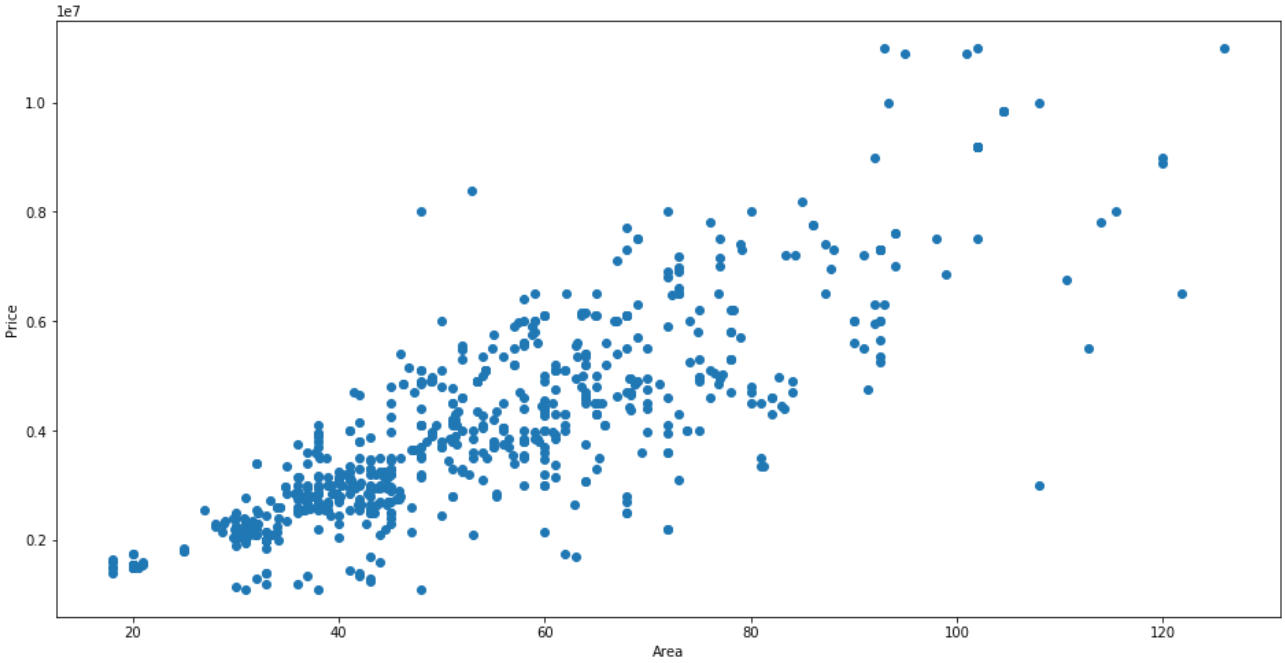

Рис. 3. Исходные данные (зависимость цены от площади) после обработки

Для проверки качества работы различных методов прогнозирования стоимости объекта недвижимости исходная выборка была разделана на обучающую (80% объектов) и тестовую подвыборки (20% объектов). На обучающей проводилось обучение моделей, а на тестовой они оценивались.

В качестве меры качества прогнозирования была выбрана *MAE* (*mean absolute error*) – средняя абсолютная ошибка, как наиболее легко интерпретируемая для условия исходной задачи оценки стоимости объекта недвижимости.

В таблице 3 приведены результаты работы методов на исходном наборе данных и на предобработанном.

Таблица. 3. Результаты работы методов

Название метода	**MAE до предобработки**	**MAE после предобработки**	**% снижения ошибки**	**% ошибки на предобработанных данных**
Линейная регрессия	800 546	645 457	19,4%	15,1%
Дерево решений	691 047	579 637	16,1%	13,6%
Случайный лес	549 813	387 601	29,5%	9,1%
Градиентный бустинг	537 373	410 291	23,6%	9,6%

Как видно из данной таблицы ансамблированные методы (случайный лес и градиентный бустинг) показали более высокие результаты по сравнению с методом дерева решений и линейной регрессией.

3.4 Значимость признаков

После построения моделей зависимости стоимости квартиры от её параметров, можно провести анализ важности этих параметров (признаков), то есть, оценить степень их влияния на результирующее значение. На рис. 4 отображены наиболее значимые с точки зрения модели признаки и их значимость.

Рис. 4. Важность признаков

На основе анализа этих данных видно, что, как и ожидалось, площадь объекта недвижимости является основным наиболее значимым параметром для оценки стоимости, при этом следующими по важности являются также площади (жилая и кухни), далее – улица (фактически район города) и этаж. А такие параметры, как тип дома, наличие или отсутствие балкона, общее количество этажей в доме слабо влияют на стоимость квартиры. Также слабая зависимость от даты обновления объявления означает, что за последний год стоимость квартир в городе Дубна менялась слабо.

Заключение

Для задачи оценки стоимости недвижимости на основе параметров объекта недвижимости (особенно на большом наборе данных) крайне важна правильная предобработка исходных данных – удаление выбросов. Данная предобработка позволяет повысит точность методов оценки на 20-30% в зависимости от выбранного метода.

Наилучшие решения (наименьшие значения ошибок) показали два метода: случайный лес и градиентный бустинг. При этом случайный лес показал более высокие результаты, чем градиентный бустинг на очищенных (предобработанных) данных, а градиентный бустинг – наоборот, оказался лидером на зашумлённых данных.

С данным набором параметров процент ошибки достигает 9-10%, что приемлемо для первоначальной оценки стоимости недвижимости, но для профессиональной работы необходимо повышение точности (снижение процента ошибки до 2-5%).

Для повышения точности работы методов возможны дальнейшие изменения описанных в работе подходов, а именно: анализ адресов через координаты и применение других методов, таких как: метод опорных векторов и нейронные сети. Для дальнейших исследований планируется применить методы машинного обучения для оценки инвестиционных проектов [8], кадрового потенциала инновационных организаций [7] и оценки перспективности инновационных площадок [9], где задачи имеют изначально сходные условия (многокритериальная оценка), но также ряд небольших исключений – отсутствие размеченных данных (обучение без учителя) для части задач.

Список литературы

1. Asteriou D., Hall S.G. Applied econometrics. 2nd Edn., Palgrave Macmillan. — 2011. — Pp: 172-196.
2. Mather B. Artificial Intelligence in Real Estate Investing: How Artificial Intelligence and Machine Learning Technology Will Cause a Transformation in Real Estate Business, Marketing and Finance for Everyone". — [Электронный ресурс]. URL: https://books.google.ru/books?id=fqV6wgEACAAJ.
3. Miller N.G., Geltner D.M. Real estate principles for the new economy. Mason, Ohio: Thomson South-Western. — 2005. — Pp. 251-330.
4. Rosen S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. — 1974. — J.P.E. 82 (1). — Pp. 34-55.
5. Triplett J. Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products // OECD Science, Technology and Industry Working Papers. — Paris: OECD Publishing, 2004. — No. 2004/09. — [Электронный ресурс]. URL: https://doi.org/10.1787/643587187107.
6. Iturra V., Paredes D. Construction of a spatial housing price index by estimating an almost ideal demand system // Tijdschrift voor economische en sociale geografie. — 2014. — Vol. 105(3). — Pp. 301-314.
7. Куликов Д.Л., Курсова Н.В., Пузова С.В. Исследование кадрового обеспечения инновационного территоривльного кластера «Дубна» // Федерализм, 2013. — № 3 (71). — С. 189-200.
8. Куликов Д.Л., Кучеров А.А. Становление и развитие методов оценки эффективности инновационных проектов // Современные проблемы науки и образования. — 2015. — № 1-1. — [Электронный ресурс]. URL: http://www.science-education.ru/ru/article/view?id=19451.
9. Куликов Д.Л., Лыкова В.А. Многокритериальный выбор инвестиционной площадки методом анализа иерархий // Фундаментальные исследования. — 2017. — № 7. — С. 151-155. — [Электронный ресурс]. URL: http://fundamental-research.ru/ru/article/view?id=41602.
10. Портал недвижимости города Дубны. — [Электронный ресурс]. URL: http://domvdubne.ru.