Сравнение эффективности методов машинного обучения в задаче оценки стоимости недвижимости

Е.О. Агафонова
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
super.kia.140401@gmail.com

А.А. Белоусов
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
adark@narod.ru

***Аннотация*—В данной статье рассматривается проблема оценки стоимости жилой недвижимости. Были собраны и обработаны данные о продаже трехкомнатных квартир в г. Самара. Для решения проблемы оценки были использованы методы машинного обучения Random Forest и Gradient Boosting, среди которых выбран наиболее эффективный.**

Ключевые слова— машинное обучение, оценка стоимости недвижимости, Random Forest, Gradient Boosting, обработка данных

1. Введение

Рынок недвижимости России является одной из самых динамичных сфер российской экономики. Учитывая огромное количество как внутренних, так и внешних факторов, влияющих на стоимость объектов недвижимости, вопрос оценки жилого имущества играет ключевую роль в сделке купли-продажи [1]. Однако расчет оценки недвижимости достаточно трудоемкая задача. Для решения данной проблемы могут использоваться методы машинного обучения, которые позволяют с высокой точностью определить стоимость недвижимости. Что подтверждает актуальность данного исследования.

Целью исследования является сравнение методов машинного обучения для задачи оценки стоимости жилой недвижимости в г. Самара и определение их эффективности.

2. Особенности подготовки данных для методов машинного обучения

Для исследования была выбрана вторичная недвижимость, расположенная в городе Самара. Данные для исследования были получены с интернет-сервиса для размещения объявлений о недвижимости «Авито Недвижимость». Данный сервис был выбран в силу своей популярности у пользователей. На нем размещено наибольшее количество интересующих нас объявлений.

Методом скрапинга были получены данные о продаже 1705 трехкомнатных квартир. Сбор данных был осуществлен при помощи расширения для Google Chrome “Web Scraper”. Данный инструмент позволяет создавать карту сайта из различных типов селекторов, извлекать данные с сайтов с несколькими уровнями навигации.

Точность оценки зависит от набора ценообразующих параметров, с помощью которых происходит идентификация объекта оценки. Определение состава параметров для описания каждого объекта является значимой составляющей формирования исходных данных. Выделим существенные характеристики объекта, определяющие его рыночную стоимость: числовые переменные – площадь квартиры, жилая площадь, площадь кухни, этаж расположения, количество этажей в доме, высота потолков, год постройки; категориальные переменные – балкон/лоджия, тип санузла, качество ремонта, вид из окна, район, тип дома.

Существенную роль в оценке недвижимости играет не только количество параметров, но и степень влияния каждого из них на рыночную стоимость квартиры. Проведенный анализ данных позволил установить значимость каждого из используемых признаков с точки зрения его влияния на стоимость объекта недвижимости (рисунок 1). Значимость признаков считается при помощи встроенного в алгоритм построения ансамбля деревьев метода feature_importances. Он основан на вычислении суммарного уменьшения минимизируемого функционала ошибки с помощью ветвлений по рассматриваемому признаку.

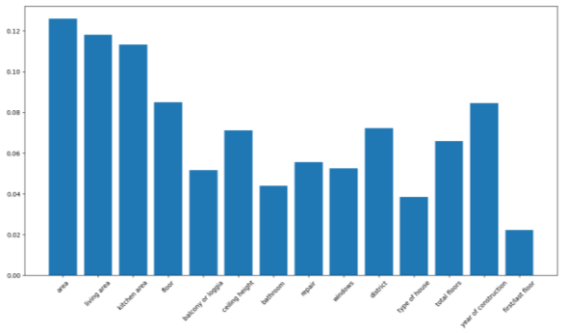

Рис. 1. Значимость признаков с точки зрения его влияния на стоимость объекта

Анализирую полученные результаты можно сделать выводы: наибольшее влияние на стоимость недвижимости оказывает площадь квартиры, жилая площадь и площадь кухни; менее значимыми оказались признаки расположения квартиры на первом или последнем этаже, тип дома и тип санузла.

После создания датасета его необходимо подготовить для методов машинного обучения. Целевой переменной является price. Для корректной работы методов машинного обучения необходимо восстановить пропущенные значения, в нашем случае ставится среднее значение или строка удаляется полностью, если характеристика является качественной. Потом находятся выбросы (результат измерения, сильно выбивающийся из выборки) и удаляются из набора данных. Значения числовых признаков подвергаются типизации, а категориальные признаки – стандартизации (убираются лишние символы и знаки препинания, исправляются

орфографические ошибки, используются только строчные буквы). Затем происходит кодирование категориальных данных методом маркировки – процесс, который ставит в соответствие числовым порядковые значения. Именно маркировка позволяет сохранить порядок, присвоив целочисленные значения, начинающиеся с 0 для значения самого низкого порядка, 1 для следующего порядка и так далее. Методы, которые будут применяться не требуют нормализации данных. Последним шагом в подготовке данных к обучению является разделение выборки на две части: обучающую (80%) и тестовую (20%). Обучающая выборка использовалась для обучения моделей, а тестовая - для определения качества их предсказания.

3. Результаты работы методов

Была написана программа на языке программирования Python, реализующая методы Random Forest и Gradient Boosting. Данные модели были обучены на тестовой выборке, которая составляет 20% от выборки. Для оценки качества моделей использовались следующие метрики: R^2 – коэффициент детерминации, MAE – средняя абсолютная ошибка, MSE – средняя квадратичная ошибка [2].

Для получения наибольшей точности предсказаний в моделях необходимо настроить гиперпараметры. Опытным путем установлено, что наилучший результат предсказания на обучающей выборке был получен при следующих значениях гиперпараметров: max_features = 0,75, min_samples_leaf = 1, n_estimators = 150 для Random Forest и learning_rate = 0,1, max_depth = 3, n_estimators = 200 для Gradient Boosting.

Результаты расчетов точности на тестовой выборки для методов Random Forest и Gradient Boosting представлены в таблице I.

Таблица I. Результаты расчетов точности на тестовой выборке для разных методов

Метод	R^2	MAE	MSE
Random Forest	0,991212	182577,08	209110839446,94
Gradient Boosting	0,990235	213013,84	232343126319,58

Анализируя таблицу, можно заметить, что Random Forest превосходит по точности Gradient Boosting по всем метрикам.

Определим зависимость точности оценки от параметров метода обучения, показавшего лучшие результаты - Random Forest.

Минимальное количество объектов в листе способствует борьбе с переобучением. Дерево строится до тех пор, пока количество объектов в листьях остается более заданного пользователем минимального числа. Когда в листе остается один объект, это может привести как к очень точной оценке, так и к большой ошибке. В то же время чем больше объектов в листе, тем больше их разнородность, что тоже может привести к ошибочной оценке. На рисунке 2 заметим, что высокое качество оценки может быть достигнуто, когда в листе не более шести объектов.

Чем больше деревьев решений, тем точнее должен быть результат. Однако на рисунке 3 заметно, что при достижении числа деревьев, равного десяти значение метрик качества меняться незначительно.

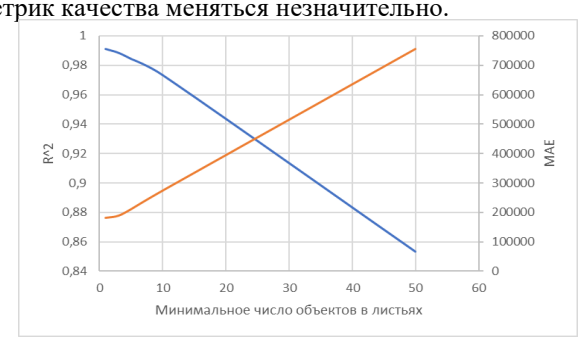

Рис. 2. Зависимость R^2 и MAE от числа от числа объектов в листьях

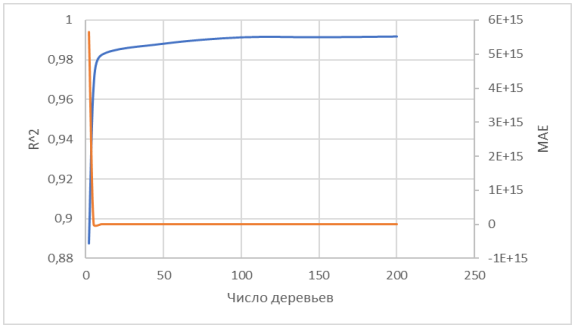

Рис. 3. Зависимость R^2 и MAE от числа деревьев

Заключение

В статье была рассмотрена проблема оценки стоимости жилой недвижимости по ее характеристикам. Для решения этой задачи были реализованы методы машинного обучения Random Forest и Gradient Boosting. Сравнение методов по качеству, которое измерялось с помощью метрик регресии, показало, что наибольшая точность предсказания у Random Forest. Были определены зависимости точности оценки от параметров метода Random Forest. Графики зависимостей показали, что высокое качество оценки может быть достигнуто, когда в листе не более шести объектов и при числе деревьев равном десяти. В ходе исследований были собраны и проанализированы данные, размещенные на сайте «Авито Недвижимость» о продаже жилой недвижимости в г. Самара, которые были использованы для обучения методов машинного обучения. Научная новизна исследования состоит в применении методов машинного обучения для решения задачи оценки стоимости трехкомнатных квартир в г. Самара.

Литература

[1] Сурков, Ф.А. Сравнение временных рядов и нейросетевых методов в задаче прогнозирования стоимости и оценки недвижимости / Ф.А. Сурков, Н.В. Петкова, С.Ф. Суховский // Моделирование, оптимизация и информационные технологии. – 2018. – Т.6, №3.

[2] Chugh, A. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? [Electronic resource]. — Access mode: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e (15.10.2021).