УДК 004.855.5

Н.А. Выходцев

Использование искусственного интеллекта для оценки стоимости недвижимого имущества

Изложены результаты анализа информационных систем в сфере продажи недвижимого имущества с целью выявления систем, использующих искусственный интеллект. Проведен корреляционный анализ характеристик объектов недвижимости. Осуществлена обработка данных, проведен сравнительный анализ алгоритмов искусственного интеллекта с вычислением коэффициента детерминации, ошибки и отклонения. Проведен ряд экспериментов по подбору гиперпараметров алгоритма. Определены характеристики, оказывающие положительное влияние на адекватность модели.

Ключевые слова: искусственный интеллект, машинное обучение, прогностическая модель, недвижимость.
doi: 10.21293/1818-0442-2021-24-1-68-72

Вопрос сбережения и оптимального расхода денежных средств при покупке недвижимого имущества является одним из важнейших, возникающим перед человеком, который желает приобрести такое имущество. Большинство покупателей производят самостоятельный анализ цен на выбранный объект. Данная операция может занимать значительное время, точность же оценки не всегда находится на высоком уровне. Поэтому компании в сфере продажи недвижимого имущества начинают предлагать своим клиентам оценку объектов [1]. Информационные системы таких организаций содержат функциональные возможности, благодаря которым пользователь может ввести в форму поиска характеристики недвижимости, в частности, основной является адрес объекта, и получить рыночную цену с довольно высокой точностью около 95–99% [2]. Такое стало возможным вследствие использования алгоритмов искусственного интеллекта. Алгоритм делает предсказание цены на основе существующих аналогичных данных.

Анализ информационных систем в сфере продажи недвижимого имущества показал, что пользователь может оценить какой-то конкретный объект, но получить информацию обо всех объектах, находящихся в продаже в сжатом, но в то же время исчерпывающем и наглядном виде, пока он не может, поскольку такой функциональной возможности еще не представлено. Для анализа использовались популярные интернет-ресурсы для продажи недвижимости [3]. Результаты представлены в табл. 1.

Таблица 1

Оценка недвижимости в интернет-ресурсах

Интернет-портал	Интерактивная карта	Индивидуальная оценка	Отчет по объекту	Искусственный интеллект
Циан		+		+
Этажи	+			
Яндекс			+	
Твой адрес		+		+
Агентства		+		

Разработка программного обеспечения, способного дать потенциальному покупателю информацию о выгодности вложения денежных средств в тот или иной объект недвижимости посредством предоставления карты местности с отображенными на ней объектами с индексом привлекательности, где индекс – показатель, формирующийся на основе сравнения предсказанной цены объекта с ценой объекта, которую предлагает продавец, является актуальной, поскольку такое сочетание использования средств визуализации и алгоритмов искусственного интеллекта в сфере недвижимого имущества как на уровне страны, так и на региональном уровне отсутствует. Создание информационной системы ведется для Томского региона. Стоит также отметить, что ускорение принятия решения покупателем за счет быстрого анализа ведет к позитивному влиянию на движение денежных потоков [4] и как следствие экономику региона.

Ключевым моментом в разработке программного обеспечения является процесс оценки стоимости недвижимости, поэтому цель работы заключается в достижении оптимальных характеристик алгоритма машинного обучения для построения высокоточной прогностической модели.

Для реализации поставленной цели выполняется ряд задач: корреляционный анализ характеристик объектов недвижимости, анализ алгоритмов искусственного интеллекта для выбора оптимального под решение задачи прогнозирования цен на недвижимость [5, 6], подбор параметров алгоритма машинного обучения для достижения наилучших результатов.

В работе используется программный пакет «Anaconda». Язык программирования Python. Выбор данного языка программирования обусловлен его высокой частотой использования для решения задач, связанных с искусственным интеллектом [7]. «Anaconda» – современная среда разработки, удобство которой заключается в ее простоте и возможности использования в одной программе сразу нескольких библиотек для обработки и анализа данных, таких как NumPy, Pandas, Sklearn [8]. Наличие данных библиотек как единого согласованного комплекта позволяет избежать конфликтов, возникающих при одиночной установке. Программный код,

используемый в исследовании, написан с помощью трех вышеперечисленных библиотек, а также с использованием библиотек Matplotlib, Math и GeoPy.

Искусственный интеллект как система включает в себя множество подсистем. В исследовании используется раздел искусственного интеллекта, называемый машинным обучением. Машинное обучение, в свою очередь, включает в себя четыре подсистемы: классическое обучение, нейросети и глубокое обучение, обучение с подкреплением и ансамблевые методы [9, 10]. В исследовании используется классическое обучение с учителем, это означает, что на вход алгоритму подаются данные, на которых он учится выявлять взаимосвязи между характеристиками объекта и строит прогностическую модель, которая, в свою очередь, в дальнейшем используется для предсказания цены объекта. Для решения поставленной задачи используется алгоритм регрессии, который способен предсказать дискретное значение [11]. Данные – это база данных с объектами недвижимости города Томска. База данных содержит информацию, включающую более двадцати характеристик о десяти тысячах объектов города и окрестностей, находящихся в продаже и проданных за последние три года. Чем больше данных в базе, тем выше точность предсказания, при этом существуют и другие факторы, влияющие на предсказание цены. Это характеристики объекта (тип объекта, адрес, площадь, количество комнат и др.). Благодаря наличию разнообразных входных данных, машине проще обучиться и выявить закономерности, соответственно и точнее результат.

После загрузки базы данных в среду разработки произведена обработка данных для последующего изучения. Удалены строки с пропущенными значениями методом «dropna()», затем произведен поиск одинаковых строк и удаление одной из них с помощью метода «drop_duplicates()».

На следующем этапе произведена оценка данных на наличие точек наблюдения, удаленных от других наблюдений. Как правило, такие данные называют выбросами, и они влияют на точность модели в сторону ее уменьшения. На рис. 1 представлена диаграмма метода оценки IQR (interquartile range), используемого для поиска выбросов.

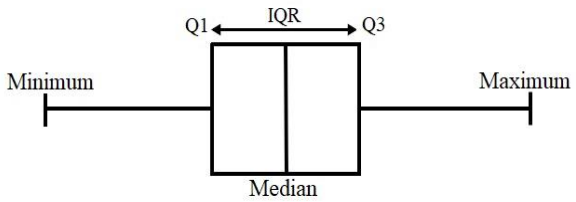

Рис. 1. Диаграмма метода оценки IQR

Median – это второй квартиль; Q1 – первый квартиль данных, т.е. 25% данных находится между Minimum и Q1; Q3 – третий квартиль данных, т.е. 75% данных находится между Minimum и Q3.

$$IQR = Q3 - Q1.$$

Для оценки выбросов используется нормальное распределение данных, т.е. распределение Гаусса. Диапазон 0,25–0,75 считается оптимальным, если мы увеличим его, то соответственно добавим нормальные рабочие данные к выбросам, если же уменьшим, то оставим часть выбросов с нормальными данными.

В результате анализа найдено и удалено 72 строки с выбросами. Общее количество данных составляет 8 213 строк. Таким образом, удалено 0,9% данных.

Для последующей работы с данными проведена их нормализация, т.е. приведение к виду, способному быть воспринятым программным обеспечением. Данные типа «Object» приведены к числовому типу.

Выбор нужных характеристик значительно влияет на итоговый результат, поэтому занимает, как правило, больше времени, чем само обучение модели. Первичная оценка характеристик проведена на основе корреляционной матрицы. Коэффициенты корреляции определены с помощью метода корреляции Пирсона [12] (рис. 2).

Корреляционная матрица:

	price
price	1,000000
area_value	0,928225
floors_total	0,315381
location_latitude	–0,153215
location_longitude	–0,209891
floor	0,238655
rooms	0,847558
building_type	–0,017732
building_series	–0,008270

Рис. 2. Корреляционная матрица характеристик недвижимого имущества

Из 24 характеристик на формирование цены влияют 8. Остальные 16 характеристик имеют оценку корреляции меньше 0,001 в сравнении с ценой (price), поэтому не используются в исследовании.

Как видно из рис. 2, на цену наибольшее влияние оказывают площадь помещения (area_value) и количество комнат (rooms). Такие характеристики, как материал, из которого построен дом (building_type), и тип помещения (building_series), имеют незначительную оценку корреляции и могут как увеличить точность итоговой модели, так и уменьшить ее. Поэтому для того чтобы определить необходимость наличия этих признаков, нужно построить модель и вычислить коэффициент детерминации.

В ходе подготовки данных также осуществлено разделение данных на тестовую и обучающие выборки в отношении 25:75% соответственно.

Чтобы провести оценку недвижимости, необходимо выбрать алгоритм машинного обучения для этой задачи. Для создания и обучения прогностической модели в исследовании используется алгоритм регрессии. Анализ алгоритмов искусственного интеллекта проведен на основе изучения результатов решения задач прогнозирования цен, имеющихся в

свободном доступе в сети Интернет, в частности, анализ цен на жилье в Бостоне [13].

Поставленную задачу можно решить, используя разные алгоритмы регрессии. От выбора алгоритма зависят точность предсказания, скорость работы и размер модели. Однако стоит отметить, что если исходная база данных содержит мало информации по объему или по качеству, то даже правильно подобранный алгоритм не сможет справиться с задачей.

Всего исследуется 6 алгоритмов: «Дерево решений», «Случайный лес», «Линейная регрессия», «Ridge регрессия», «Lasso регрессия» и «Полиномиальная регрессия» [14]. Из самих названий алгоритмов «Дерево решений» и «Случайный лес» уже можно понять, что «Случайный лес» состоит из множества «Деревьев решений». «Случайный лес» – это ансамблевая модель. Ансамблевые модели сегодня используются повсеместно из-за своей эффективности и быстроты [15]. Данные методы подходят для всего, где используется классическое обучение, но точность выше, чем у отдельных методов. Часто их сравнивают с нейросетью по своей эффективности. Алгоритм «Случайный лес» строит деревья решений, раз за разом исправляя ошибки предыдущего дерева, в итоге качество полученных предсказаний намного выше в сравнении с использованием алгоритма «Дерево решений».

Такие методы, как «Линейная регрессия», «Ridge регрессия», «Lasso регрессия» и «Полиномиальная регрессия», применяются непосредственно в регрессионном анализе. При этом «Полиномиальная регрессия» так же, как и «Случайный лес», работает на основе более простого алгоритма «Линейной регрессии».

Проведен анализ точности алгоритмов. Работа каждого алгоритма оценена с помощью вычисления коэффициента детерминации (R-квадрат), медианного абсолютного отклонения (MAD) и средней абсолютной ошибки (MAPE) [16]. Результаты представлены в табл. 2.

Таблица 2

Сравнение алгоритмов машинного обучения

Название алгоритма	R^2	MAD, %	MAPE, %
RandomForestRegressor	0,939	8,81	13,88
Ridge Regression	0,925	16,48	22,61
Lasso Regression	0,924	15,8	22,61
Linear Regression	0,922	22,53	16,64
Polynomial Regression	0,881	15,6	22,17
DecisionTree	0,807	14,29	23,24

В табл. 2 коэффициенты детерминации у первых четырех алгоритмов очень близки, поэтому дополнительно высчитываются MAD и MAPE. В результате проведения экспериментов по оценке эффективности работы алгоритмов выявлено, что наивысший показатель коэффициента детерминации и наименьшие показатели MAD и MAPE у алгоритма «Случайный лес».

На итоговое значение коэффициента R^2 влияют также значения гиперпараметров алгоритма «Случайный лес». На рис. 3 представлена функция, отображающая зависимость адекватности модели от количества «деревьев». При количестве «деревьев» ,равном 250, коэффициент R^2 составляет 0,937, время обработки данных и формирования модели составляет 5 с. Максимальное значение коэффициента R^2 равно 0,939, достигается при количестве «деревьев» – 2000 за время 10 с. Дальнейшее увеличение гиперпараметра ведет к увеличению времени обработки, но коэффициент детерминации остается на том же уровне.

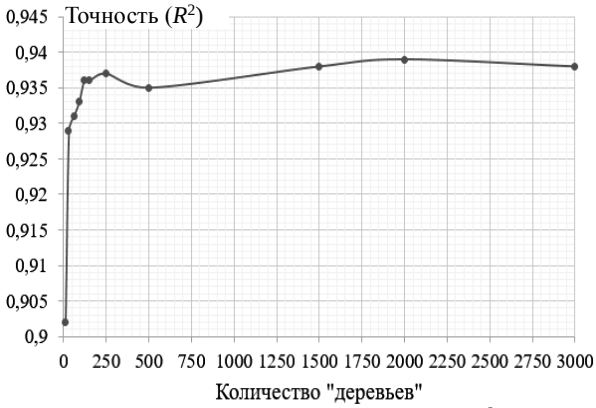

Рис. 3. Зависимость коэффициента R^2 от количества «деревьев»

В табл. 3 представлена зависимость коэффициент R^2 от глубины «леса». Лучшее значение достигается при глубине «леса», равной 55. Гиперпараметр «min_samples_split», обозначающий минимальное количество выборок, необходимое для разделения внутреннего узла, установлен равным 2. При его уменьшении точность уменьшается, при увеличении – не изменяется.

Таблица 3

Определение максимальной глубины «леса»

Глубина «леса»	R^2
5	0,918
10	0,938
30	0,937
55	0,939
65	0,937
100	0,937

Проведена заключительная оценка характеристик (табл. 4). Для этого исследовано, как влияет наличие последних двух характеристик из корреляционной матрицы на рис. 2 с наименьшими показателями корреляции на адекватность модели. Вычислены коэффициент детерминации, ошибка (MAPE) и отклонение (MAD).

Таблица 4

Оценка характеристик

Характеристики	Наличие/отсутствие			
Building_type	–	–	+	+
Building_series	–	+	+	–
Оценки	Значения			
R^2	0,872	0,88	0,939	0,87
MAPE	17,61	14,87	13,92	15,65
MAD	9,65	9,5	8,81	9,12

Исходя из результатов экспериментов, лучшее значение коэффициента R^2 достигается при наличии обеих характеристик, следовательно, в исследовании все характеристики важны, и их общее количество равно 8.

Проведенные выше эксперименты показали, что оптимальным алгоритмом является «Случайный лес» с глубиной «леса» – 55, количеством «деревьев» – 2000, «min_samples_split» – 2 и количеством характеристик – 8. Значение коэффициента детерминации составило 0,939, медианное абсолютное отклонение – 8,81%.

Выводы

Таким образом, в рамках исследования проведено изучение информационных систем по продаже недвижимости, в результате выявлена потребность в выработке алгоритма машинного обучения, необходимого для анализа цен на недвижимость. Для достижения поставленной цели проведен анализ характеристик объектов недвижимости и выбраны оказывающие положительное влияние на адекватность формируемой модели, осуществлена предварительная обработка данных путем удаления выбросов повторяющихся и незаполненных данных, проведен сравнительный анализ алгоритмов искусственного интеллекта с вычислением коэффициента детерминации, ошибки и отклонения, выбран алгоритм «Случайный лес», проведен ряд экспериментов по подбору гиперпараметров алгоритма. Достигнутые характеристики и параметры позволяют сделать вывод об адекватности модели.

Литература

1. Шолле Ф. Глубокое обучение на Python. – СПб.: Питер, 2018. – 400 с.

2. Turing A.M. Computing Machinery and Intelligence // Mind. – 1950. – Vol. 59, No. 236. – P. 433–460.

3. Cortes C. Support-Vector Networks / C. Cortes, V. Vapnik // Machine Learning. – 1995. – Vol. 20, No. 3. – P. 273–297.

4. Mukhopadhyay S. Advanced Data Analytics Using Python. – Kolkata: Apress, 2018. – 195 p.

5. Embarak O. Data Analysis and Visualization using Python. – Abu Dhabi: Apress, 2018. – 390 p.

6. Nelli F. Python Data Analitics. – Rome: Apress, 2018. – 576 p.

7. Milovanovich I. Python Data Visualization Cookbook. – Birmingham: Packt Publishing, 2013. – 280 p.

8. Levantesi S. The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach / S. Levantesi, G. Piscopo // Risks. – 2020. – Vol. 8, No. 4. – P. 112–129.

9. Randal S. Python Machine Learning. – Birmingham: Packt Publishing, 2015. – 454 p.

10. Madhavan S. Mastering Python for Data Science. – Birmingham: Packt Publishing, 2015. – 294 p.

11. Squire M. Mastering Data Mining with Python – Find patterns hind in your data. – Birmingham: Packt Publishing, 2016. – 269 p.

12. Martins L. Mastering Python Data Analysis / L. Martins, M. Persson. – Birmingham: Packt Publishing, 2016. – 282 p.

13. Кормен Т. Алгоритмы. Построение и анализ / Т. Кормен, Ч. Лейзерсон, Р. Ривест. – М.: Вильямс, 2013. – 1328 с.

14. Chervonenkis A. A Note on One Class of Perceptrons / A. Chervonenkis, V. Vapnik // Automation and Remote Control. – 1964. – Vol. 25, No. 16. – P. 103–109.

15. Рашка С. Python и машинное обучение. – М.: ДМК-Пресс, 2017. – 420 с.

16. Layton R. Learning Data Mining with Python. – Birmingham: Packt Publishing, 2015. – 344 p.

Выходцев Никита Андреевич
Аспирант отделения информационных технологий (ОИТ)
инженерной школы информационных технологий
и робототехники (ИШИТР)
Национального исследовательского
Томского политехнического университета (НИ ТПУ)
Ленина пр-т, 30, г. Томск, Россия, 634050
Тел.: +7-913-823-33-88
Эл. почта: vyh.dtsev@mail.ru

Vykhodtsev N.A.

Artificial intelligence in price estimation of real estate

The article contains results of information systems analyses that are used for real estate estimation and are based on artificial intelligence. A correlation analysis of the characteristics of real estate objects has been carried out. Data processing, as well as comparative analysis of artificial intelligence algorithms with computation of accuracy, error and deviation were implemented. A number of experiments were realized to select the hyper parameters of the algorithm. The characteristics that have a positive effect on adequacy have been determined.

Keywords: artificial intelligence, machine learning, forecasting model, real estate.
doi: 10.21293/1818-0442-2021-24-1-68-72

References

1. Sholle F. *Glubokoe obuchenie na Paiton* [Deep learning on Python]. Saint Petersburg, Piter, 2018. 400 p. (in Russ.).

2. Turing A.M. Computing Machinery and Intelligence. *Mind*, 1950, vol. 59, no. 236, pp. 433–460.

3. Cortes C., Vapnik V. Support-Vector Networks. *Machine Learning*, 1995, vol. 20, no. 3, pp. 273–297.

4. Mukhopadhyay S. *Advanced Data Analytics Using Python*. Kolkata, Apress, 2018, 195 p.

5. Embarak O. *Data Analysis and Visualization using Python*. Abu Dhabi, Apress, 2018, 390 p.

6. Nelli F. *Python Data Analitics*. Rome, Apress, 2018, 576 p.

7. Milovanovich I. *Python Data Visualization Cookbook*. Birmingham, Packt Publishing, 2013, 280 p.

8. Levantesi S., Piscopo G. The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach. *Risks*, 2020, vol. 8, no. 4, pp. 112–129.

9. Randal S. *Python Machine Learning*. Birmingham, Packt Publishing, 2015, 454 p.

10. Madhavan S. *Mastering Python for Data Science*. Birmingham, Packt Publishing, 2015, 294 p.

11. Squire M. *Mastering Data Mining with Python – Find patterns hind in your data*. Birmingham, Packt Publishing, 2016, 269 p.

12. Martins L., Persson M. *Mastering Python Data Analysis.* Birmingham, Packt Publishing, 2016, 282 p.

13. Cormen T., Layzerson Ch., Rivest R. *Algoritmy. Postroenie i analis* [Algorithms. Constructions and analysis]. Moscow, Viliams, 2013, 1328 p. (in Russ.).

14. Chervonenkis A., Vapnik V. A Note on One Class of Perceptrons. *Automation and Remote Control*. 1964, vol. 25, no. 16, pp. 103–109.

15. Rashka S. *Paiton i mashinnoe obuchenie* [Python and machine learning]. Moscow, DMK Press, 2017. 420 p. (in Russ.).

16. Layton R. *Learning Data Mining with Python*. Birmingham, Packt Publishing, 2015, 344 p.

Nikita A. Vykhodtsev
Postgraduate student, Department of Information Technologies (OIT), School of Engineering of Information Technologies and Robotics (ISHITR),
National Research Tomsk Polytechnic University (NR TPU)
30, Lenin pr., Tomsk, Russia, 634050
Phone: +7-913-823-33-88
Email: vyh.dtsev@mail.ru