ИССЛЕДОВАНИЕ ВАЖНОСТИ ХАРАКТЕРИСТИК ПРИ ПОСТРОЕНИИ ПРОГНОСТИЧЕСКОЙ МОДЕЛИ ДЛЯ ОЦЕНКИ СТОИМОСТИ НЕДВИЖИМОСТИ

Н.А. Выходцев, аспирант гр. А9-39.
Томский политехнический университет
E-mail: vyh.dtsev@mail.ru

Введение

В рамках научно-исследовательской работы проводится исследование базы данных объектов недвижимости г. Томска, находящихся в продаже и проданных за последние 5 лет. Разработано программное обеспечение для оценки стоимости объектов недвижимости и проведен ряд экспериментов по улучшению эффективности работы алгоритма машинного обучения [1]. Коэффициент детерминации алгоритма «Случайный лес» составляет 93.9% на используемом наборе данных при медианном абсолютном отклонении (MAD) - 8.8%, средней абсолютной ошибке (MAPE)- 13.9%.

Анализ статей по машинному обучению показал, что улучшение адекватности модели также достигается за счет создания новых характеристик, имеющих уникальные значения для каждого объекта недвижимости [2]. Поэтому принято решение разработать соответствующее программное обеспечение и провести ряд экспериментов для изучения его эффективности и влияния на итоговые результаты модели.

Основная часть

В качестве новых характеристик выбраны «Дистанция» - расстояние от центра города до объекта недвижимости и «Азимут» - градусная мера дуги, между направлением на север от центра города и направлением на объект от центра города.

Данные характеристики в своей совокупности позволяют уникально идентифицировать каждый объект в базе. Поскольку наибольшее влияние на точность модели оказывает площадь помещения с коэффициентом корреляции Пирсона равным 0.92 и значения площади у большинства объектов различны, то использование таких характеристик как «Азимут» и «Дистанция» может улучшить точность модели.

Выдвинута гипотеза о влиянии новых характеристик на значение коэффициента детерминации в сторону его увеличения и уменьшения MAD на любом наборе данных, связанном с продажей недвижимого имущества.

Поскольку база данных уже содержит столбцы с широтой и долготой по каждому объекту, то «Азимут» и «Дистанция» вычисляются с их помощью. Для вычисления «Дистанции» используется функция geodesic из библиотеки geopy Python. «Азимут» вычисляется с помощью функции «get_azimuth» [3].

Выявлены и удалены выбросы, проведена нормализация данных. В качестве алгоритма для построения прогностической модели используется «Случайный лес» со следующими значениями гиперпараметров: «n_estimators» = 2000, «max_features» = 6, «max_depth» = 55 [4].

В модели исследуется 10 характеристик: «Площадь помещения», «Серия дома», «Тип дома», «Этаж», «Количество этажей», «Азимут», «Дистанция», «Широта», «Долгота», «Количество комнат». Проведена оценка корреляции характеристик, результат представлен на рисунке 1.

Корреляционная матрица:

	price
price	1.000000
area_value	0.927917
floors_total	0.315999
location_latitude	-0.131830
location_longitude	-0.208674
floor	0.239265
rooms	0.847646
building_type	-0.018805
building_series	0.004838
distance	-0.090402
azimuth	0.146231

Рис. 1. Корреляционная матрица

Коэффициент детерминации при данных параметрах составил 90%, MAD - 12%. MAPE - 16%. Полученный результат ниже, чем достигнутый без новых характеристик.

Исследование характеристик показало, что в модели присутствуют коллинеарный характеристики, которые сильно коррелируют между собой и тем самым уменьшают точность прогноза. Удалены «Широта» и «Долгота», так как эти характеристики сильно коррелируют с «Азимут» и «Дистанция». После удаления коллинеарных характеристик в итоговый набор вошли: «Площадь помещения», «Серия дома», «Количество этажей», «Азимут», «Дистанция», «Количество комнат». Коэффициент детерминации при данных параметрах составил 93.2%, MAD - 11.4%. MAPE - 15%. Диаграмма важности характеристик представлена на рисунке 2.

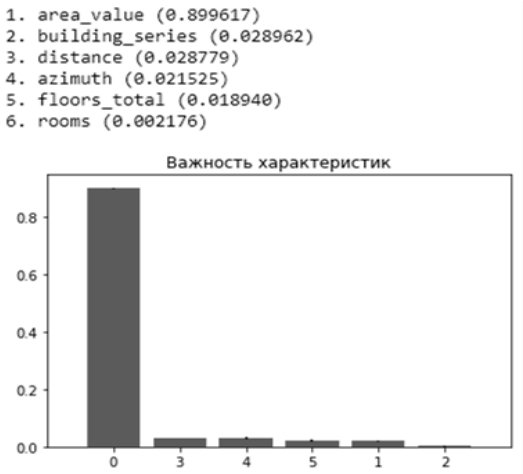

Рис. 2. Диаграмма важности характеристик.

Проведено сравнение алгоритмов машинного обучения с целью выявления алгоритма, наиболее адекватно формирующего прогностическую модель. Результаты представлены в таблице 1.

Таблица 1. Сравнение алгоритмов машинного обучения

Название алгоритма	R^2	MAD, %	MAPE, %
RandomForestRegressor	0.932	11.4	15
Polynomial Regression	0.91	22.2	15
Linear Regression	0.863	18.5	21
Ridge Regression	0.86	18.5	22
Lasso Regression	0.857	18.5	22
DecisionTree	0.82	18.5	22

Лучшие показатели точности и наименьшего отклонения у алгоритма «Случайный лес».

Заключение

Выдвинутая гипотеза о влиянии новых характеристик на значение коэффициента детерминации в сторону его увеличения и уменьшения MAD на используемом наборе данных не подтвердилась. Значения коэффициента детерминации практически идентичны 93.2% и 93.9%, отклонение MAD увеличилось с 8.81% до 11.4%.

Полученные результаты свидетельствуют о том, что влияние характеристик «Азимут» и «Дистанция» на точность прогноза зависит от используемой базы данных и характеристик объектов недвижимости [5].

Список использованных источников

1. Выходцев Н.А. Использование искусственного интеллекта для оценки стоимости недвижимого имущества. Доклады ТУСУР. – 2021. – Т. 24. – №. 1.
2. Machine Learning и оценка недвижимости. [Электронный ресурс]. – URL: https://medium.com/@max.bobkov/machine-learning-moscow-flats-appraising-25a1e9f171db (дата обращения 15.01.2021).
3. Географические информационные системы и дистанционное зондирование [Электронный ресурс]. – URL: https://gis-lab.info/qa/great-circles.html (дата обращения 19.01.2021).
4. Шолле Ф. Глубокое обучение на Python. – СПб.: Питер. – 2018. – 400 с.
5. Рашка С. Python и машинное обучение. – М.: ДМК Пресс. – 2017. – 420 с.