Алгоритм определения цены на недвижимость

Выполнил:
Молчанов Артём
371 гр.
Научный руководитель:
Невоструев Константин

Введение

- Потеря в среднем 14% от стоимости недвижимости при покупке/продаже
- Машинное обучение позволяет достаточно точно определять цену по набору признаков

Задачи

- Собрать достаточную для работы базу данных, используя методы интеллектуального анализа данных
- Обработать полученные данные
- Построить регрессионную модель
- Построить классификационную модель
- Провести апробацию данных моделей на реальных данных
- Провести измерение полученных результатов и сравнить с имеющимися

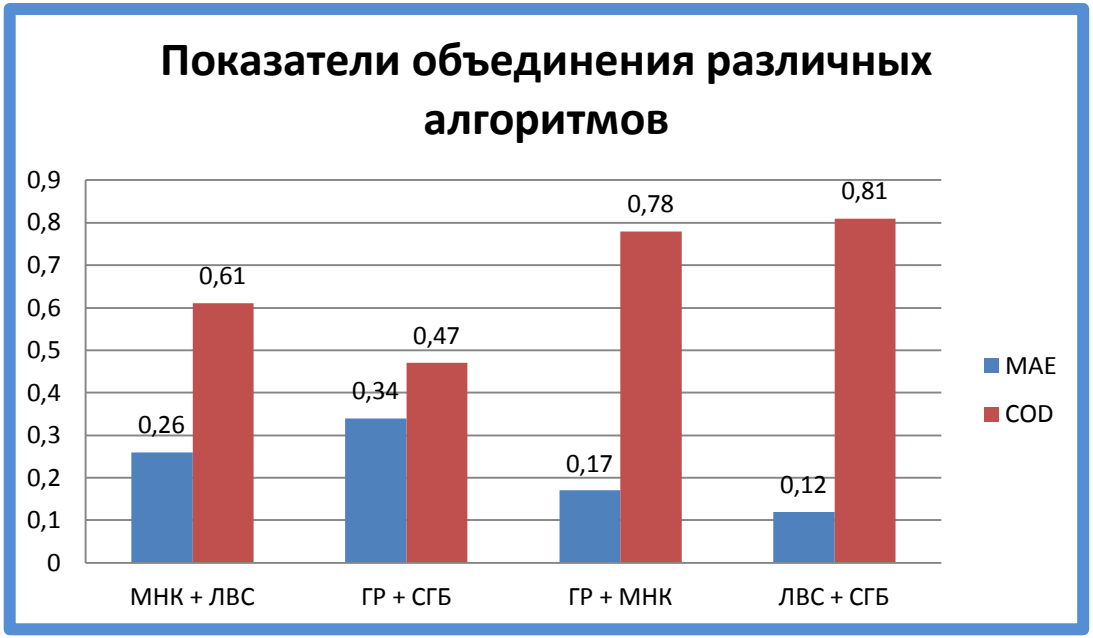
Обзор

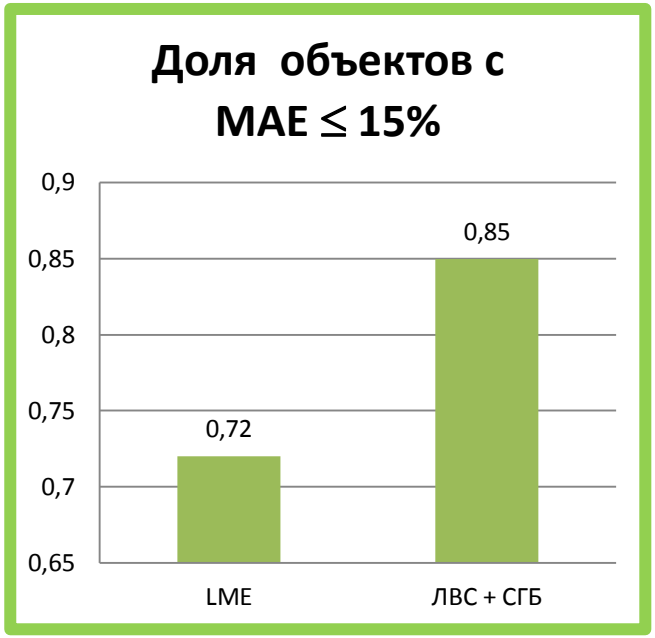
- Крупнейший американский портал Zillow
- Сервис Flatorial
- Статья «Discovering the Hidden Structure of House Prices with a Non-Parametric Latent Manifold Model»
  - Лучший результат: при MAE ≤ 0.15 верно определено 72%

Сбор и обработка данных

- Рассматривался г. Санкт-Петербург и часть Ленинградской области
- Собраны данные со специализированной базы данных (~5000 объектов)
- Получена информация с 5-и информационных порталов и 12-и отраслевых сайтов и 2-х статей
- Исключены 4 признака как мало влияющие (< 0,05%) на результат
- Добавлены признаки:
  - Различные статистические показатели
  - Средняя прогнозируемая рентабельность
  - Расстояние до КАД
  - Средняя цена объектов находящихся в доме/корпусе/участке
- Всего проанализировано 140000 объектов
- По каждому объекту получено 43 признака

Регрессия





Коэффициент смешанной корреляции (COD или R^2) — статистический показатель, оценивающий соответствие модели данным.

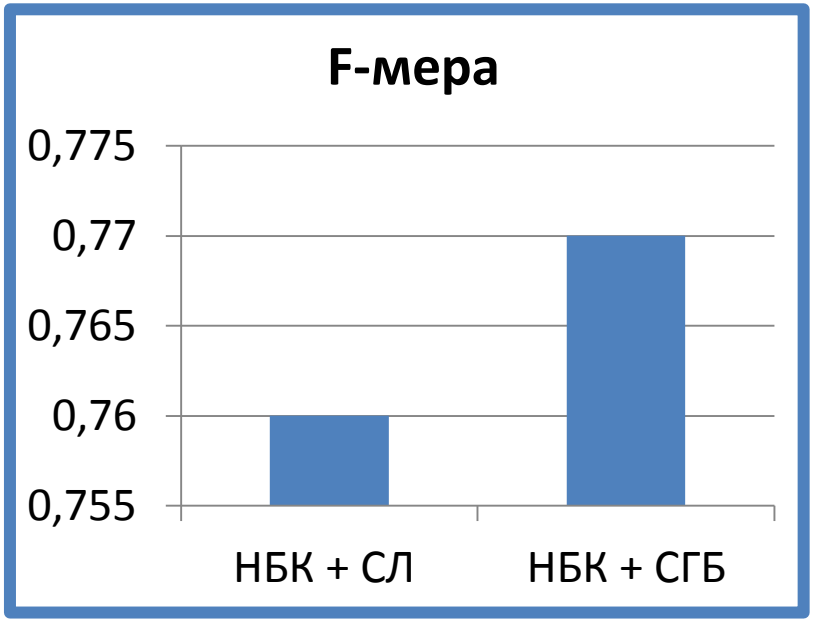
Среднее арифметическое отклонение (MAE) — среднее значение арифметических отклонений (*отклонение* — это разница между спрогнозированным значением и его фактическим значением).

Рассмотренные алгоритмы: Метод наименьших квадратов (**МНК**), Стохастический градиентный бустинг (**СГБ**), Гребневая регрессия (**ГР**), Локально взвешенное сглаживание (**ЛВС**), Latent Manifold Estimation (**LME**) или Множественная скрытая оценка

Классификация

40 классов по 400 тыс. руб. (от 1 до 17 млн.) и 2 класса (от 0 до 1 млн. и от 17 млн.)

Ансамбль	Recall	Precision	Accuracy	F
k Ближайших Соседей (**kБС**) + Наивный байесовский классификатор (**НБК**)	0,7	0,75	0,87	0,72
kБС + Стохастический градиентный бустинг (**СГБ**)	0,62	0,81	0,85	0,7
k**БС** + Случайный лес (**СЛ**)	0,64	0,72	0,81	0,68
НБК + **СЛ**	0,71	0,81	0,92	0,76
СГБ + **СЛ**	0,56	0,68	0,74	0,61
НБК + **СГБ**	0,73	0,81	0,9	0,77



Recall (avg) –цена действительно принадлежит ценовому диапазону

Arecision(avg) – цена действительно не принадлежит диапазону

Accuracy – корректные предсказания, относительно всей выборки

F-мера - характеристика, дающая оценку одновременно по точности и полноте

Технологии

- Язык программирования:
  - Python
- Библиотеки:
  - Beautiful Soup
  - Grab
  - Lxml
  - Pandas
  - Scikit-Learn
  - Numpy
  - SciPy
  - Matplotlib

Результаты

- Используя методы интеллектуального анализа данных, была собрана достаточная для работы база данных
- Полученные данные были обработаны
- Построена регрессионная модель
- Построена классификационная модель
- Проведена апробация данных моделей на реальных данных
- Проведены измерения полученных результатов, которые были сравнены с имеющимися (на 18% повышена точность регрессионной модели)