Слайд 1

Добрый день, уважаемая комиссия, разрешите представить вам мою магистерскую диссертацию на тему прогнозирования рыночной стоимости объектов недвижимости города Иваново методами машинного обучения.

Слайд 2

Целью данной работы, является исследование способа проведения оценки жилой недвижимости средствами машинного обучения. Для достижения этой цели нужно решить следующие задачи: ознакомиться с предметной областью, выбрать методы оценки недвижимости, собрать данные и подготовить их, выбрать алгоритмы МО и обучить их, а после протестировать полученные модели и проанализировать результаты. На основании этих результатов требуется выбор наиболее оптимально работающий метода для оценки при малом объеме данных в обучающей выборке.

Слайд 3

Для данной работы были проанализированы следующие документы, которые регламентируют процесс оценки недвижимости, а именно федеральный закон об оценочной деятельности в российской федерации, федеральный стандарт оценки №1 и федеральные стандарты оценки номер 7. На основании этих документов, было выявлено что оценка жилой недвижимости, а именно к ней относятся квартиры, применяется сравнительный подход, в рамках которого используется метод сравнения продаж/предложений жилой недвижимости.

Слайд 4

Для получения данных был создан простой парсер, который собирал данные с двух сайтов. C IvanovoDom.ru он собирал данные о продажах квартир в городе Иваново, а с сайта dom.mingkh.ru парсер собирал информацию о зданиях, в которых находились данные квартиры, после чего объединял эти данные и записывал в определенное хранилище.

Слайд 5

 После сбора стало видно, что было собрано около 6600 предложений по продаже квартир, а для каждого предложение было собрано 27 различных признаков, которые характеризуют его. После всех преобразований и обработки данных было оставлено всего 6 признаков. Общая выборка разбивалось на тренировочную и тестовую в соотношении 5 к 1.

Слайд 6

Изначально для каждого объекта имелось 27 различных признаков, но в ходе обработки часть из них были удалены, а часть преобразованы. Первой парой преобразованных признаков были площади квартиры, в этом случае общая площадь содержит в себе как жилую так и площадь кухни, поэтому признак 3 и 4 были удалены, оставляя только признак 2. То же самое произошло и с адресом размещенных квартир, то есть широта и долгота так же отражают координаты зданий, в которых находятся квартиры, они обозначают то же самое, что признаки 7, 8 и 9. Поэтому от них было принято решение отказаться, оставив только широту и долготу, которые имели больше возможностей для использования. Дальше были удалены, признаки у которых была найдена моя маленькая корреляции с целевым признаком, ценой. То есть в этом качестве выступили признак с количеством комнат, признак наличия мусоропровода в доме , признак наличия во дворе дома детской площадки и признак с наличием в дворе спортивной площадки. Также была удалена большая группа признаков, для которых, количество пропущенных значений в данных составило больше 60%, такие признаки не имели никакой ценности для обучения алгоритмов, и при этом могли навредить, уменьшив точность модели.

Слайд 7

После очистки данных остались следующие признаки: цена, по которой продается квартира; общая площадь этой квартиры; этаж, на котором располагается данная квартира; общее количество этажей в здании и координаты самого здания.  После этого некоторые признаки были преобразованы в следующие: цена преобразовалась в стоимость за квадратный метр, координаты преобразовались в удаленности квартиры от центра города и направлении относительно центра. Остальные признаки остались такими же какими и были. Из этих признаков стоимость за квадратный метр является целевым, остальные же являются независимыми. Проверим качество полученных данных, оценив корреляцию между оставшимися признаками. Слева представлена тепловая карта с корреляцией между признаками. Из нее видно, что наибольшая корреляция достигается между удаленностью от центра города и стоимостью за квадратный метр. Хуже всего корреляция между направлением и цена за квадратный метр. Из этого можно сделать вывод, что в городе Иваново нет зависимости цены от стороны света, но есть зависимость цены от удаленности здания от центра, что в принципе логично. Как видно из данного рисунка, корреляции между всеми возможными признаками достаточно низка, что говорит о достаточно плохом качестве данных.

Слайд 8

На данном слайде показана график зависимости стоимости квадратного метра от удаленности квартиры от центра. Работе отсев выбросов производится при помощи сужения границ используемых данных. Для этого была найдена медиана для стоимости квадратного метра, которая равна 41200,92 рублей за квадратный метр. И производилась в последствии производилось сужение при помощи квартилей. На нулевой итерации, использовались все доступные значения. После чего на шаг 1, а где брались квартили, равные 0,1 и 0,9 соответственно появились минимальная и максимальная границы данных, между которыми и брались предложения о продаже. После этого производилась вторая итерация, где квартили были уже 0,2 и 0,8. Потом 0,3 и 0,7, и наконец 0,4 и 0,6. Таким образом производилось постепенное сужение данных к медиане и отсев выбросов.

Слайд 9

В работе рассматривалось 3 различных алгоритма машинного обучения для решения задачи регрессионного анализа, которая появляется при оценке жилой недвижимости сравнительным подходом. Для этих целей были выбраны алгоритмы: случайного в леса, который основан на деревьях решений, градиентного бустинга на деревьях и линейная регрессия.

Слайд 10,11

Для оценки полученных алгоритмов, а было, были выбраны 2 критерия. Первый - этот коэффициент детерминации, который отражает общую универсальности полученной модели и ее соответствие данным. Второй является средней абсолютной ошибкой в процентах, который показывает, насколько точно алгоритм может определять стоимости той или иной квартиры. Соответственно, чем меньше данный параметр, тем точнее алгоритм должен определять стоимость квартиры. Выбран этот параметр, потому что, а он является наиболее подходящим для экономических операций в силу того, что, уменьшает штраф при выбросах.

Слайд 12

На данном слайде представлены результаты обучения и тестирования алгоритма случайного леса. Для данного алгоритма использовалось 2000 деревьев, а глубина построения каждого дерева равна 50-ти узлам, с количеством вариантов для разделения равным 5-ти. Наиболее оптимальный результат по точности и универсальности, был достигнут при 53% коэффициента детерминации и ошибки равной 2,12%.

Слайд 13

Алгоритм бустинга был также построен на 2000 деревьях с глубиной построения в 50 узлов и для данного алгоритма была получена оптимальная точность равная 47% и ошибка равная 2,15%.

Слайд 14

Линейная регрессия из всех результатов показал себя хуже всего, поскольку ее коэффициент детерминации составил всего 9%, ошибка была 3,5%.

Слайд 15

На слайде представлена таблица, с результатами обучения алгоритмов машинного обучения при разном количестве отсечённых данных. Из таблицы видно, что сужение границ используемых данных, увеличивает точность модели, но ухудшает ее универсальность. В данном случае можно предположить, что если оценщик найдет достаточно большое количество аналогов для оценки квартиры, то он сможет получить модель, которая с хорошей точностью предскажет цену на квартиру.

Слайд 16

Для практического эксперимента, была выбрана квартира сильно приближенная по стоимости за квадратный метр к медиане, то есть к 41 000 рублей за квадратный метр. Производилась оценка данной квартиры при помощи методов машинного обучения, которые были обучены при максимальном усреднения данных, то есть при квартилях равных 0,4 и 0,6.  Для каждого из алгоритмов были получены оценка стоимости недвижимости без учета погрешности и диапазон цен с учетом погрешности. Также была произведена оценка данной квартиры при помощи встроенного калькулятора на циане. С теми же данными что использовались и для обучения. Наиболее близким по диапазону цен является линейная регрессия, также она наиболее близкую оценила квартиру к ее оригинальной стоимости. Но это достигнуто за счет более высокой о погрешности вычислений в 1,7%, из - за чего ее диапазон цен расширился. В данном эксперименте стоит отметить, что предоставленный на сайте калькулятор, оценил квартиру в 2 000 000 рублей. Этот результат является более надежным, потому что калькулятор работу на основе большего количества данных. Такое неоднозначное. Значение в цене можно объяснить тем, что цена самого предложения нас сайте казана не совсем точно относительно рынка, в котором она находится. Поэтому линейной регрессии оказалось к стоимости, указанной на сайте намного ближе, чем остальные методы, которые стремятся также к 2 000 000 рублей. Исходя из этого. Делается вывод, что наиболее оптимальным является алгоритм случайного леса, который показывает результаты, удовлетворительное относительно. И цены указаны на сайте, и цены, предсказанной калькулятором, а так же на основании коэффициента детерминации и абсолютной ошибки.

Слайд 17

В качестве заключения для данной работы хотелось бы привести следующие результаты. В работе были собраны данные по продажам жилой недвижимости в городе Иваново за определенный срок; были применены методы машинного обучения к оценке жилой недвижимости, а так же был выбран метод случайного леса, как наиболее оптимальный из рассмотренных для проведения оценки жилой недвижимости после обучения на малом объеме данных. Был сделан вывод, что для получения наиболее точных результатов, в рамках оценки конкретного объекта недвижимости требуется жертвовать универсальностью моделей машинного обучения, за счет усреднения обучающей выборки к параметрам оцениваемого объекта. Было сделано основные заключение, что использовании методов машинного обучения в оценочной деятельности в рамках небольшого сегмента рынка, такого как в городе Иваново не целесообразно.