Санкт-Петербургский Государственный Университет
Математико-механический факультет

Программная инженерия
Системное программирование

Молчанов Артём Андреевич

Алгоритм определения цены на недвижимость

Курсовая работа

Научный руководитель:
ведущий разработчик ООО «НМТ» Невоструев К. Н.

Санкт-Петербург
2016

Оглавление

1. Введение

Определение справедливой цены на недвижимость является одной из основных проблем как для покупателей, так и для продавцов, потому что это очень многофакторная задача. На стоимость объекта влияют как объективные параметры (удаленность от метро, жилая площадь, этаж), так и более субъективные (экология района, качество отделки и надёжность застройщика). Решением этой проблемы могут выступать такие методы машинного обучения с учителем как регрессионный анализ и классификация. Модели на их основе будут исходя из анализа объективных и необъективных параметров определять близкую к реальной[1] цену на недвижимость.

1.1. Формирование базы данных

Для непосредственного тестирования алгоритмов необходима база данных, содержащая необходимое количество объектов, а также достаточный набор признаков, который позволил бы оценить влияние каждого из них на результат. Вследствие того, что удовлетворительной готовой базы данных не было обнаружено, было принято решение поставить дополнительную интересную и трудоёмкую в практическом смысле задачу сбора данных.

При решении поставленной задачи были использованы такие методы интеллектуального анализа данных, предназначенные для автоматического обнаружения веб-документов и извлечения информации из веб-ресурсов (Web Mining) как:

- Web Content Mining (Извлечение веб-контента) — процесс извлечения знаний из контента документов или их описания, доступных в Интернете
- Web Structure Mining (Извлечение веб-структур) — процесс обнаружения структурной информации в Интернете

[1] Не завышенную или заниженную относительно похожих объектов

1.2. Машинное обучение

Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Один из его основных разделов - обучение с учителем, который предназначен для решения следующей задачи. Имеется некоторое множество объектов и определённое множество вероятных ответов. Существует некоторая зависимость между объектами и ответами, однако она неизвестна. Известна лишь конечная совокупность прецедентов — пар «объект, ответ», называемая обучающей выборкой. На основе этих данных нужно восстановить зависимость, то есть построить алгоритм, который для любого объекта мог бы выдать достаточно точный ответ. Учитель - это сама обучающая выборка, либо тот, кто указал на заданных объектах верные ответы.

Пусть X — множество описаний объектов, Y — множество допустимых ответов. Существует неизвестная целевая зависимость — отображение $y^* : X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Требуется построить алгоритм $a : X \to Y$, который приближал бы неизвестную целевую зависимость как на элементах выборки, так и на всём множестве X.

Основные типы задач, рассматриваемых данным разделом - задачи регрессии и классификации, решаемые с помощью регрессионного анализа и классификационного подхода.

Регрессия Регрессия — зависимость математического ожидания (например, среднего значения) случайной величины от одной или нескольких других случайных величин (свободных переменных), то есть $E(y|\mathbf{x}) = f(\mathbf{x})$. Регрессионным анализом называется поиск такой функции f, которая описывает эту зависимость. Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих.

$$y = f(\mathbf{x}) + \nu,$$

где f — функция регрессионной зависимости, ν — аддитивная случайная величина с нулевым математическим жиданием.

Основные виды регресионных моделей, рассматриваемые в данной работе:

- Линейная регрессия предполагает, что функция f зависит от параметров $\mathbf{w}$ линейно. При этом линейная зависимость от свободной переменной $\mathbf{x}$ необязательна,

$$y = f(\mathbf{w}, \mathbf{x}) + \nu = \sum_{j=1}^{N} w_j g_j(\mathbf{x}) + \nu.$$

  В случае, когда функция $g \equiv \mathrm{id}$ линейная регрессия имеет вид

$$y = \sum_{j=1}^{N} w_j x_j + \nu = \langle \mathbf{w}, \mathbf{x} \rangle + \nu,$$

  здесь x_j — компоненты вектора $\mathbf{x}$.

  Многомерная линейная регрессия — это линейная регрессия в n-мерном пространстве (объекты и признаки являются n-мерными векторами).

- Нелинейная регрессионная модель — модель вида

$$y = f(\mathbf{w}, \mathbf{x}) + \nu,$$

  которая не может быть представлена в виде скалярного произведения

$$f(\mathbf{w}, \mathbf{x}) = (\mathbf{w}, \mathbf{g}(\mathbf{x})) = \sum_{i=1}^{n} w_i g_i(\mathbf{x}),$$

  где $\mathbf{w} = [w_1, \ldots, w_n]$ — параметры регрессионной модели, $\mathbf{x}$ — свободная переменная из пространства R^n, y — зависимая переменная, ν — случайная величина и $\mathbf{g} = [g_1, \ldots, g_n]$ — функция из некоторого заданного множества.

Классификация Пусть X — множество описаний объектов, Y — конечное множество номеров (имён, меток) классов. Существует неизвестная целевая зависимость — отображение $y^* : X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Требуется построить алгоритм $a : X \to Y$, способный классифицировать произвольный объект $x \in X$.

Основные виды задач классификации, разделённые по типам классов:

- Двухклассовая классификация. Наиболее простой в техническом отношении случай, который служит основой для решения более сложных задач.

- Многоклассовая классификация. Когда число классов достигает многих тысяч (например, при распознавании иероглифов или слитной речи), задача классификации становится существенно более трудной.

1.3. Измерение результатов

Один из основных методов оценки качества алгоритмов на основе регрессии или классификации - перекрёстная проверка или кросс-валидация. Это тип исследования качества алгоритма, при котором уже сохраненные, статические данные разбиваются на k частей. Одно часть называется тестовой — на ней проводится измерение качества работы алгоритма по метрике. Остальные части называются тренировочными. На них производится обучение алгоритма.

Метрики качества работы Существует несколько метрик для оценки качества работы алгоритма на основе регрессионной модели, среди них можно выделить:

- **Среднеквадратичная ошибка** (**MSE**)

$$\mathbf{E}(\hat{\theta}_i - \theta_i)^2,$$

где $\hat{\theta}$ — предсказанные, а θ — истинные значение

- **Средняя абсолютная ошибка** (**MAE**)

$$\mathbf{E}|\hat{\theta} - \theta|,$$

где $\hat{\theta}$ — предсказанные, а θ — истинные значение

- **Относительное арифметическое отклонение** (**RAE**)

$$\frac{\mathbf{E}|\hat{\theta} - \theta|}{\mathbf{E}|\theta_i - \bar{\theta}|},$$

где $\hat{\theta}$ — предсказанные, θ_i — фактические, а $\bar{\theta}$ — среднее арифметическое фактических значений

- **Коэффициент смешанной корреляции** (**COD или** R^2)

$$\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2}{\sum_{i=1}^{n}(\theta_i - \bar{\theta})^2},$$

где $\hat{\theta}$ — предсказанное, $\bar{\theta}$ — среднее значение самой величины, а θ — истинное значение

Множество образцов в результате классификации разбивается на четыре множества:

- **True Positive** — образцы из положительного класса определенные в положительный класс
- **True Negative** — образцы из отрицательного класса определенные в отрицательный класс
- **False Positive** — образцы из отрицательного класса определенные в положительный класс
- **False Negative** — образцы из положительного класса определенные в отрицательный класс

Определим диапазон, которому принадлежит определённая цена как положительный класс, а остальные в совокупности - как отрицательный. На основе четырёх вышеперечисленных множеств определены следующие метрики:

- **Precision** — статистическая метрика, вычисляемая по формуле:

$$Precision = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

- **Recall** также является статистической метрикой:

$$Recall = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

2. Постановка задачи

Целью данной работы является:

- Собрать достаточную для работы базу данных, используя методы интеллектуального анализа:
  - Данных (Data Mining)
  - Веб-ресурсов (Web Mining)
  - Текста (ИАТ)
- Обработать полученные данные, используя алгоритмы:
  - Фильтрации
  - Дополнения
  - Кодирования
- Построить регрессионную модель
- Построить классификационную модель
- Провести апробацию данных моделей на реальных данных
- Провести измерение полученных результатов

3. Обзор

Существует множество работ по предсказанию цены на недвижимость. Однако работ по определению цены на недвижимость меньше. После рассмотрения некоторых из них, были сделаны следующие выводы:

1. Отсутствует универсальный подход для решения задачи
2. Результаты сильно зависят от качества базы данных
3. Данные брались из открытых источников и разрабатываемый алгоритм этим ограничивался
4. Комбинация различных моделей часто улучшает результат

Авторы	Предметная область
Sumit C., Trivikraman T., и др.[3]	Определение цены на недвижимость
Robert F., David M. и др.[4]	Подход к оценке
Weikun Zhao, Cao Sun и др.[14]	Построение модели

Таблица 1: Рассмотренные работы

Анализ данных работ показал, что существуют неочевидные признаки, влияющие на результаты работы алгоритм; комбинации различных моделей показывают разные результаты в зависимости от подхода к их использованию; возникают значительные трудности с интерпретацией различных признаков, а также с выбором метрик для них.

Задача, решаемая в данной работе, является одной из первых, решаемых на момент основания крупнейшего американский портала о недвижимости Zillow[15].

Zillow - это он-лайновый сервис в сфере недвижимости в США, к которому пришли его основатели Рич Бартон и Ллойд Фринк после того, как они провели множество часов занося данные в таблицу, в которой сравнивали привлекательность предложений, чтобы после сложных вычислений определить ценность и преимущества каждого объекта. Основатели пришли к заключению, что миллионы людей сталкиваются с

этой же задачей и терпят те же неудобства при классификации и упорядочивании информации, которую накапливают прежде чем осуществить какую-либо операцию на рынке недвижимости. Так и родилась эта идея, наделить всех желающих той же информацией и правилами определения стоимости недвижимости по определенным параметрам, что используют в своей работе риелторы.

Также в России есть стартап-проект Flatorial[5]. Это сервис, автоматически оценивающий стоимость квартир с помощью алгоритма определяющий стоимость квартиры в Москве и погрешность оценки.

4. Формирование базы данных

4.1. Подбор признаков

Для определения признаков, влияющих на цену недвижимости, пришлось погрузиться в предметную область. В качестве сегмента рынка недвижимости был выбран сегмент строящейся недвижимости. После анализа данной области было принято решение рассматривать такие признаки как например:

- Признаки самого объекта (квартиры/дома):
  - Общая площадь объекта
  - Жилая площадь объекта
  - Нежилая площадь объекта
  - Площадь кухни
  - Количество комнат
  - Качество отделки
  - Высота потолка
  - Тип санузла (совмещённый/раздельный)
  - Наличие балкона
  - Наличие лоджии
  - Сторона(ы) света, на которые приходится вид из окон
  - Площадь балкона
  - Площадь лоджии
  - Этаж
- Признаки жилых комплексов:
  - Близость к метро (в минутах)
  - Этажность
  - Наличие паркинга

- Количество мест в паркинге
- Уровень криминальной активности
- Уровень загрязнённости
- Расстояние до КАД (в км.)
- Надёжность застройщика (рейтинг по международному стандарту)
- Инфраструктура
- Аккредитованность банками
- Наличие разрешение на строительство
- Наличие ипотечного способа оплаты
- Наличия оплаты в рассрочку
- Скидка более 5% при 100%-ой оплате или ипотеке
- Минимальный первый взнос при кредитовании
- Наличие охраны
- Наличие зелёных зон
- Наличие огороженной территории

4.2. Сбор данных

Для сбора информации было выбрано 17 сайтов, посвящённых строящейся недвижимости по городу Санкт-Петербург и Ленинградской области, а также 2 статьи о криминогенной активности и об уровне загрязнения. Основные признаки были получены при извлечении DOM-структур каждого сайта (либо его части) и её последующем анализе и фильтрации. На основе статей составлялись отдельные таблицы, представленные в csv-формате.

Для получения ряда признаков необходимо было воспользоваться информационным анализом текста. Например, для нахождения расстояния до КАД нужно было проанализировать описание объекта и найти соответствия сигнальным словам, которые определяли, какую часть

информации следует извлечь, и, затем представить её в виде действительного числа.

Более сложные признаки получались посредством объединения парсеров различных веб-ресурсов и операциям с данными, полученными в процессе их работы. Так, например, чтобы определить среднюю ликвидность недвижимости, понадобилось сначала определить среднюю годовую арендную ставку в радиусе километра от объекта, затем вычислить среднюю стоимость объектов в рассматриваемом доме, корпусе или участке данного типа, и в итоге найти частное, выраженное в долях.

4.3. Инструменты

Для реализации парсеров рассматривались такие языки программирования как Python, PHP и Ruby. Было принято решение в пользу языка Python, из-за его удобства и наличия библиотек с надлежащим качеством документаций и множеством полезных методов. Пример таких библиотек: Lxml[9] , BeautifulSoup[1] , Grab[7] .

4.4. Подготовка данных

Необходимо было отфильтровать признаки для лучшего качества выборки, а соответственно и самих алгоритмов. При этом несколько из них было решено в дальнейшем не учитывать, так как информации для интерпретации было недостаточно. Было создано три группы признаков:

- Признаки малых объектов недвижимости (квартиры)
- Признаки средних объектов недвижимости (частные дома, коттеджы)
- Признаки больших объектов недвижимости (жилые комплексы)

Данное разбиение помогло точнее определить структуру конечных объектов, а также проследить влияние различных факторов на конечный

результат.

После фильтрации было принято решение ввести новые признаки, такие как:

- Среднее количество квартир на лестничной площадке
- Средняя рентабельность квартиры (в %)
- Доля квартир-студий на лестничной площадке (в %)

Один из самых трудоёмких этапов стал этап кодирования признаков. Необходимо было правильно выбрать метрики признаков для верной интерпретации данных. Этот шаг был важен, поскольку от этого зависела точность построенных моделей. Был закодирован ряд признаков, например:

- Качество отделки (0 - без отделки, 1 - предчистовая, 2 - чистовая, 3 - стандартная, 4 - эконом, 5 - люкс)
- Сторона(ы) света, на которые приходится вид из окон (0 - юг, 1 - север, 2 - восток, 3 - запад, 4-7 - сз, св, сз, юв соответственно)
- Уровень криминальной активности (количество преступлений на 10000 жителей (по районам))
- Уровень загрязнённости (в тысячах кубических метров выброса всех загрязняющих веществ)
- Аккредитованность банками (0 - совокупный рейтинг банков ниже 6, 1- от 6 до 10, 2 - от 10 до 14, 3 - выше 14)
- Инфраструктура (количество объектов социального назначения (школы, детские сады, поликлиники, торговые центры и др.) в радиусе двух квадратных километров)

После этого, данные были нормированы для более качественной работы алгоритмов и наглядного представления при выявлении закономерностей и дополнительного их анализа.

Выборка получилась размером в 140000 объектов, каждый из которых имел 43 признака.

На этапе обработки данных было достигнуто качество базы данных, позволявшее работать с ней далее, применяя различные алгоритмы машинного обучения и математической статистики.

5. Регрессионная модель

Общее назначение множественной регрессии состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной. При анализе каждого объекта, нужно понять связаны ли и каким образом признаки этого объекта с ценой. Таким образом, при фильтрации была удалена из рассмотрения лишь часть признаков, вторая же часть была отсеяна в процессе обучения алгоритма. Также были обнаружены "выбросы", т.е. объекты, которые имеют слишком низкую или слишком высокую цену, учитывая их расположение и характеристики. В процессе построения различных моделей, и их ансамблей, подобные проблемы были разрешены путём анализа остатков (разностей между действительным результатом и предсказанным), а также основываясь на визуальном анализе закономерностей. Для определения конкретного значения цены были отобраны 4 регрессионные модели:

- Метод наименьших квадратов (МНК)
- Стохастический градиентный бустинг (СГБ)
- Гребневая регрессия (ГР)
- Локально взвешенное сглаживание (ЛВС)

При этом дополнительно были рассмотрены следующие работы:

Авторы	Алгоритм
Miller, Steven J [10]	МНК
Friedman, Jerome H [6]	СГБ
Hoerl, Arthur E and Kennard, Robert W [8]	ГР
Cohen, Robert A [2]	ЛВС

Таблица 2: Рассмотренные работы

При применении ансамблей алгоритмов в разных сочетаниях были получены следующие результаты:

Ансамбль	RAE	COD
МНК + ЛВС	0.43	0.61
ГР + СГБ	0.52	0.47
ГР + МНК	0.37	0.78
ЛВС + СГБ	0.32	0.81

Таблица 3: Результаты работы ансамблей регрессионных моделей

При анализе полученных результатов следует учитывать коэффициент смешанной корреляции (COD), так как он наиболее точно отображает точность алгоритма (более 0.75 алгоритм считается приемлемо точным, более 0.85 очень точным).

На основании вычислений точности рассмотренных моделей можно сделать вывод, что ансамбль состоящий из взвешенной суммы метода наименьших квадратов и стохастического градиентного бустинга, наиболее точно позволяет определять цену на недвижимость.

6. Классификация

Решая задачу классификации необходимо знать точное количество классов и их границы, так как выбор неверного разбиения может ухудшить показатели модели.

На основании того, что 95% объектов выборки имеет цену от 1 до 17 млн. рублей, было принято решение разбить стоимость на 40 классов по 400 тыс. руб. (от 1 млн. до 17 млн.) и 2 класса (от 0 до 1 млн. и от 17 млн.).

- k Ближайших Соседей (kБС)
- Стохастический градиентный бустинг (СГБ) — веса объектов вместо градиента
- Наивный байесовский классификатор (НБК)
- Случайный лес (СЛ)

При применении ансамблей алгоритмов в разных сочетаниях были получены следующие результаты:

Ансамбль	Recall	Precision	Accuracy	F
kБС + НБК	0.70	0.75	0.87	0,72
kБС + СГБ	0.62	0.81	0.85	0,70
kБС + 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	0.90	0,77

Таблица 4: Результаты работы ансамблей классификационных моделей

В данном случае на каждом шаге выбираются 2 класса, один - целевой (он же положительный), а другой - объединение всех остальных (отрицательный).

Здесь значения Recall и Precision — среднее арифметическое Recall и Precision по каждому классу.

Очевидно, что чем выше точность и полнота, тем лучше модель. Однако, в реальной жизни максимальная точность и полнота не достижимы одновременно и приходится искать некий баланс. Для того, чтобы объединить информацию о точности и полноте алгоритма была применена метрика F (F-мера).

F-мера представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремится к нулю.

$$F = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Делая вывод о лучшем ансамбле моделей, можно заметить, что, несмотря на более высокое соответствие модели ансамбля наивного байесовского классификатора и случайного леса (0.92) F-мера у него меньше, чем у ансамбля наивного байесовского классификатора и стохастического градиентного бустинга ($0.76 < 0.77$). Значит, последняя модель наиболее выгодна для рассмотрения.

7. Измерение результатов

7.1. Описание методов

Измерения качества алгоритмов проводилось с использованием кросс-валидации. Это техника оценки качества предсказательного алгоритма, при котором данные разбиваются на k частей, на k–1 части производится обучение, а на одной — проверка. . В данной работе данные разбиваются на две части, соответственно половина используется как тренировочная, половина — как тестовая выборка.

В качестве метрик используются precision и recall[13]. Precision, в данном случае, — отношение верно определённого диапазона (класса) цены ко всем определённым. Recall — отношение верно определённого диапазона (класса) цены ко всем верным.

Регрессионная модель показала лучший результат на ансамбле алгоритмов стохастического градиентного бустинга и метода наименьших квадратов с коэффициентом смешанной корреляции 0.81, что является достаточно высоким показателем оптимальности.

Классификационная модель показала лучший результат на ансамбле алгоритмов наивного байесовского классификатора и стохастического градиентного бустинга с коэффициентом точности 0.90 и коэффициентом оптимальности 0.77, что также является высоким показателем оптимальности.

7.2. Инструменты

Реализация алгоритмов и их тестирование производилась на языке Python, так как он удобен для машинного обучения и научных расчетов из-за существования таких библиотек, как numpy[16] , scipy[12] , scikit-learn[11] .

Заключение

В рамках данной работы был разработан алгоритм определения цены на недвижимость на основе регрессионной и классификационной моделей с высокими показателями точности. В частности, были решены следующие задачи:

- Собрана база данных размером 140000 объектов на 43 признака, необходимая для анализа и построения моделей.
- Полученные данные обработаны, используя методы фильтрации, дополнения и кодирования
- Построена регрессионная модель с коэффициентом эффективности 0.81
- Построена классификационная модель с коэффициентом эффективности 0.77 и точностью 0.90
- Проведена апробация этих моделей на реальных данных
- Проведены измерения полученных результатов

Список литературы

[1] BeautifulSoup. — 2016. — May. — URL: https://www.crummy.com/software/BeautifulSoup/.

[2] Cohen Robert A. An introduction to PROC LOESS for local regression // Proceedings of the 24th SAS users group international conference, Paper / Citeseer. — Vol. 273. — 1999.

[3] Discovering the hidden structure of house prices with a non-parametric latent manifold model / Sumit Chopra, Trivikraman Thampy, John Leahy et al. // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007. — 2007. — P. 173–182. — URL: http://doi.acm.org/10.1145/1281192.1281214.

[4] Engle Robert F, Lilien David M, Watson Mark. A dymimic model of housing price determination // Journal of Econometrics. — 1985. — Vol. 28, no. 3. — P. 307–326.

[5] Flatorial. — 2016. — May. — URL: http://flatorial.ru/.

[6] Friedman Jerome H. Stochastic gradient boosting // Computational Statistics & Data Analysis. — 2002. — Vol. 38, no. 4. — P. 367–378.

[7] Grab. — 2016. — May. — URL: http://grablib.org/ru/.

[8] Hoerl Arthur E, Kennard Robert W. Ridge regression: Biased estimation for nonorthogonal problems // Technometrics. — 1970. — Vol. 12, no. 1. — P. 55–67.

[9] Lxml. — 2016. — May. — URL: http://lxml.de/.

[10] Miller Steven J. The method of least squares // Mathematics Department Brown University. — 2006. — P. 1–7.

[11] Scikit-learn. — 2016. — May. — URL: http://scikit-learn.org/stable/.

[12] Scipy. — 2016. — May. — URL: https://www.scipy.org/.

[13] Wikipedia. Precision and recall // Wikipedia, the free encyclopedia. — 2016. — URL: https://en.wikipedia.org/wiki/Precision_and_recall.

[14] Zhao Weikun, Sun Cao, Wang Ji. The Research on Price Prediction of Second-hand houses based on KNN and Stimulated Annealing Algorithm // International Journal of Smart Home. — 2014. — Vol. 8, no. 2. — P. 191–200.

[15] Zillow. — 2016. — May. — URL: http://www.zillow.com/.

[16] numpy. — 2016. — May. — URL: http://www.numpy.org/.