

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ НАУК
Кафедра программной и системной инженерии

РЕКОМЕНДОВАНО К ЗАЩИТЕ
В ГЭК И ПРОВЕРЕНО НА ОБЪЕМ
ЗАИМСТВОВАНИЯ

Заведующий кафедрой

А.Г. Ивашко
Д.т.н., профессор

А.Г. Ивашко

2019 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
(магистерская диссертация)

Применение методов машинного обучения в системе RIES для
прогнозирования цен

Прикладная информатика 09.04.03

Магистерская программа: Прикладная информатика в экономике

Выполнил (а) работу
Студент (ка) 2 курса
очной формы обучения

Любякина
(Подпись)

Любякина
Евгения
Владимировна

Научный руководитель
к.т.н., доцент кафедры

Григорьев
(Подпись)

Григорьев
Михаил
Викторович

Рецензент
руководитель технической
поддержки, Е-СОФТ



Семина
Татьяна
Валерьевна

г. Тюмень, 2019

Содержание

Введение	3
Глава 1. Описание исходного набора данных	5
Глава 2. Обзор источников про методы прогнозирования	11
2.1. Методы оценивания стоимости недвижимости	12
2.2. Обзор методов прогнозирования	15
Глава 3. Разработка модели	21
3.1 Предобработка набора данных	21
3.2 Нормализация данных	29
3.3 Корреляция данных.	36
Параметр “Этаж”	41
Параметр “Год постройки”	43
Параметр “Общая площадь”	46
Параметр “Количество собственников”	48
Параметр “Количество фотографий”	50
Параметр “Улица”	51
Параметр “Район”	52
Параметр “Тип дома”	53
Параметр “Этаж”	54
Параметр “Этажность”	55
Параметр “Жилая площадь”	56
Параметр “Площадь кухни”	57
3.4 Выявление значимых параметров	59
Заключение	66
Список литературы.	68
Приложение 1	71
Приложение 2	83

Введение

Улучшить эффективность в сфере недвижимости можно за счет факторов, которые показывают, как настоящее состояние рынка, так и возможное. Цена объектов недвижимости зависит от воздействия множества факторов, например, характеристики самого объекта недвижимости, а также факторы, которые не связаны с объектом: состояние рынка, факторы ликвидности жилья.

Для комплексного исследования оценки недвижимости рассматривают влияние также параметров на уровне регионов, так как в каждом регионе присущи свои характерные факторы влияющие на ценообразование объекта недвижимости. К таким характерным фактором относятся жилищные программы, программы поддержки молодых специалистов, активное развитие ипотечных программ.

Цены, которые формируются на основе среднего значения стоимости объектов первичного и вторичного рынка, являются объектом для исследования рынка недвижимости при помощи оценки и прогнозирования стоимости объектов недвижимости. Данные исследования формируют общую картину для прогноза не только на уровне региона, но и в целом на уровне государства. С помощью этих исследований формируются социальные и ипотечные программы для покупки недвижимости.

Для улучшения исследований на рынке недвижимости необходимо детально изучать методы и модели прогнозирования, выявить все факторы, которые влияют на ценообразования объекта недвижимости.

Целью данной работы является построение модели прогнозирования цены 1 кв.м. вторичной недвижимости в г. Тюмень с точностью не менее 80% от рыночной стоимости.

Для достижения данной цели были поставлены следующие задачи:

1. Проанализировать исходный набор данных.
2. Изучить подходы и модели, которые применяются для прогнозирования цен на недвижимость.
3. Отобрать информативные параметры.
4. Выделить значимые факторы, отражающие процессы ценообразования на рынках жилой недвижимости.
5. Построить и оценить модель прогнозирования цены 1 кв.м вторичной недвижимости в г. Тюмень.

Объектом исследования служит рынок жилой недвижимости города Тюмени.

Глава 1. Описание исходного набора данных

Рынок недвижимости делится на следующие подкатегории:

- жилая недвижимость - недвижимость, которая пригодна для жилья на основе документов;
- коммерческая недвижимость - недвижимость, которая приносит собственнику этой недвижимости дополнительный доход (торговые, складские, офисные помещения и т.д.);
- рынок земельных участков - недвижимость, которая находится за пределами города и не относится к типу жилой недвижимости [22].

Также, выделяют два вида рынка недвижимости: первичный и вторичный. К первичному рынку относят помещения, которые только сданы в эксплуатацию. Сделки по продаже первичного жилья осуществляют застройщики и инвесторы, которые осуществляют строительные работы при помощи выделения денежных средств. Ко вторичному рынку относят объекты недвижимости, которые уже находятся в частной или муниципальной собственности. В рамках данной работы рассмотрена жилая недвижимость класс “Квартиры” вторичного рынка по городу Тюмени.

Перед оформлением сделки по купле-продаже квартиры покупателю, как правило, необходимо изучить все предложения на рынке недвижимости для ознакомления всех стоимостных предложений и сопоставления своих потребностей, желаний для приобретения удобного жилья и со своими финансовыми возможностями. Также, продавец исследует аналогичные предложения на рынке с подобной рыночной стоимостью перед выставлением своего объекта недвижимости

на продажу. Это необходимо так как, недооценка ведет к потере прибыли, а также при завышенной стоимости возможность продать объект значительно уменьшается. “На сегодняшний день доступны четыре способа оценки рыночной стоимости квартиры:

- самостоятельная оценка;
- использование онлайн калькулятора;
- профессиональный оценщик;
- риелтор” [19].

“Переход к рыночной экономике превратил объекты недвижимости в товар, к которому могут быть применены операции купли-продажи, следовательно, возникает необходимость в понятии стоимости. При этом система цен недвижимости имеет свои особенности. Цена объекта складывается из ряда взаимосвязанных факторов, которые учитываются экспертами-оценщиками [17]. К таковым факторам относятся:

- индивидуальные факторы объекта недвижимости;
- особенности внешней среды рынка;
- фактор спроса и предложения;
- факторы пользователя объекта недвижимости;
- фактор целевого назначения объекта.”

К индивидуальным факторам объектам относятся: номер объекта в базе, риелтор, дата создания, дата редактирования, дата приостановления, причина приостановления, время до метро, кадастровый номер, этаж, количество ванных комнат, год постройки, общая площадь, площадь кухни, площадь балкона, ванная комната, комментарии, тип серии, тип стен, балкон, лоджия, высота потолка, количество комнат, площадь комнат, количество парковочных мест.

К особенностям внешней среды относятся: город, район, улица, станция метро, описание территории, описание расположения дома.

Факторы спроса и предложения: количество фотографий, ипотека, сумма ипотеки, обмен, дополнительный комментарий, наличие фотографий, банк ипотеки, обременение, сумма обременения, банк обременения, плита, окно, ремонт, молодая семья, субсидии.

Факторы пользователя объекта недвижимости: цена, торг, ключи в офисе, наличие баннер, количество собственников.

Фактор целевого назначения объекта: класс, тип (продажа/аренда), статус, вторичное жилье, тип.

“Степень влияния каждого из факторов зависит от конкретной ситуации, тем не менее”, при проведении оценки объекта недвижимости необходимо учитывать весь спектр, основываясь на российском стандарте факторов описания квартиры [24].

Характерные черты для рынка жилой недвижимости это высокая стоимость и низкая эластичность по цене, имеющихся вариантов на рынке недвижимости: при увеличении или снижении ценового параметра объекта недвижимости на жилые объекты, объем рынка фактически не изменяется. Такая взаимосвязь: изменение стоимости и объема предложения на рынке, возникает из-за длительного периода на стадии строительства. Поэтому, при росте спроса на жилые объекты недвижимости, рост предложения увеличится в период 1-2 лет, а при падении спроса на жилые объекты, срок строительства не изменяется, поскольку высока вероятность возникновения новых затрат. Таким образом, стоимость является главным компонентом регулирования спроса и предложений на рынке. Поэтому, увеличивается рост ситуаций, когда рыночная стоимость неоправданно

высока стоимости, которая присутствует на реальном рынке недвижимости [21].

Оценка рыночной стоимости квартиры – задача непростая, рынок недвижимости находится в постоянном движении и оценщик должен учитывать все его изменения. Существующие методы, способы, подходы к оценке, которыми пользуются специалисты оценки имеют как свои достоинства, так и недостатки такие как, низкая активность рынка для каждой категории жилья приводит к недопустимости для сравнительного анализа по причине того, что информация устаревает, применяются только при стандартных характеристиках жилья, и не возможны для применения к уникально новопостроенным помещениям. Поэтому человек, который хочет узнать стоимость на свой объект недвижимости, ищет баланс между известными методиками определения стоимости для нахождения оптимального варианта. Так как точность оценки зависит не только от полученного дохода от будущей сделки, но и от степени удовлетворенности участников процесса.

Для исследования данной работы был взят набор данных из риэлторской информационной системы (RIES) по городу Тюмени. Исходный файл выгружен в формате csv. Данные выбраны по вторичному рынку типа недвижимости “Квартира”, так как для исследования рассмотрены характеристики объектов без внешних факторов, которые участвуют в прогнозировании. Набор данных состоит из 801 объекта и 147 характеристик объектов. Типы данных встречаются в данном наборе: текстовый, числовой, логический, дата/время. Полный набор характеристик представлен в приложении 1. Для работы был взят исходный набор данных, состоящий из 801 объектов недвижимости и 147 характеристик объектов (рисунок 1).

id	newhouses_id	class	action	direction	status	user_id	last_editor_id	city_id	district_id	street_id	on_map	photos_cnt	tours_cnt
2066874	0	flats	sale	supply	active	1231213993	1231213993	0	0	0	0	no	11
2080642	0	flats	sale	supply	active	1231213993	1231213993	0	0	0	0	no	19
2080645	0	flats	sale	supply	active	1231213993	1231213993	0	0	0	0	no	27
1009646	84533	flats	sale	supply	active	2117	1010116305	6	1634	35157	no	12	0
86055	11606	flats	sale	supply	active	340	345	23	12440	297	no	10	0
100900	8894	flats	sale	supply	active	788	788	23	117	1205	no	21	1
107187	20776	flats	sale	supply	active	242	242	23	12437	315	no	17	0
117745	12151	flats	sale	supply	active	346	346	23	12440	439	no	0	0
137928	52096	flats	sale	supply	active	301	301	23	12438	297	no	18	0
145357	10535	flats	sale	supply	active	98	98	23	142	288	no	30	0
150128	11687	flats	sale	supply	active	247	247	23	12437	305	no	24	0
150138	10265	flats	sale	supply	active	740	740	23	135	331	no	33	5
150331	11284	flats	sale	supply	active	375	375	23	398	324	no	6	0
151949	12083	flats	sale	supply	active	60730	60730	23	12438	403	no	12	2
154051	16390	flats	sale	supply	active	60537	60537	23	464	283	no	29	4
156936	37080	flats	sale	supply	active	2833	14371	23	12438	256	no	19	0
157289	12323	flats	sale	supply	active	14277	14277	23	120	347	no	17	0
159478	9347	flats	sale	supply	active	60312	1264	23	120	366	no	17	0
162732	10297	flats	sale	supply	active	788	788	23	135	373	no	24	4
165020	12355	flats	sale	supply	active	776	776	23	588	282	no	18	2
165331	12660	flats	sale	supply	active	270	270	23	801	345	no	16	0
166287	8746	flats	sale	supply	active	60579	60579	23	118	316	no	12	0
167488	12689	flats	sale	supply	active	60169	60169	23	808	278	no	19	4
167513	11856	flats	sale	supply	active	2823	2823	23	12437	333	no	23	0
174907	11911	flats	sale	supply	active	486	486	23	12438	342	no	34	3
176581	8772	flats	sale	supply	active	377	377	23	114	389	no	30	0
177199	11978	flats	sale	supply	active	1276	1276	23	126	360	no	30	1
179484	11556	flats	sale	supply	active	1231	1264	23	12437	289	no	48	5
179849	11767	flats	sale	supply	active	770	770	23	12440	329	no	22	0
181324	12221	flats	sale	supply	active	334	334	23	12437	538	no	30	0
183102	17868	flats	sale	supply	active	333	333	23	12440	297	no	10	0
183602	12057	flats	sale	supply	active	60701	471	23	12437	392	no	24	0
187132	10029	flats	sale	supply	active	60101	60101	23	132	291	no	30	4

Рисунок 1 - Исходный набор данных

Вывод

Рассматривая квартиру как товар при операциях купли/продажи в рыночной экономике, необходимо учитывать конкретные факторы, а именно: район, улица, количество фотографий, дата создания, дата редактирования, этаж, этажность, количество комнат, площадь, количество ванных комнат, окно, ремонт, серия дома, балкон, высота потолков. Исходя из имеющихся факторов определяется рыночная цена и рыночная стоимость квартиры. Существует несколько способов для

оценки стоимости квартир. Использование нескольких способов или комбинируя их, можно получить наиболее правильную рыночную стоимость.

Далее работа будет осуществляться с отражением объекта недвижимости в информационном пространстве. Исходный набор данных содержит 147 характеристик. Характеристики объектов представлены с различными типами данных (приложение 1).

Глава 2. Обзор источников про методы прогнозирования

Прогнозирование стоимости объекта недвижимости является заключительным и важным этапом в анализе рынка. Прогноз стоимости строится на основе изменений движения цен объектов недвижимости, а также на основе проведенного анализа рынка.

Более 100 организаций в России, основными представителями которых являются агентства недвижимости, банки, застройщики, используют в своей деятельности инструменты прогнозирования оценки изменения цен объектов недвижимости [16].

Например, сеть агентств недвижимости «ЭКСПЕРТ» в городе Уфа использует уже несколько лет технический анализ тенденций регионального рынка в своей деятельности. В результате использования инструментов прогнозирования агентство недвижимости «ЭКСПЕРТ» стало лидером продаж в Республике Башкортостан.

Так же, инструменты прогнозирования играют важную роль для застройщиков. С их помощью можно предсказать неизвестность в отношении ипотечных программ, от которых зависит сложно прогнозируемое потребительское поведение [26]. Снижение цен может привести к росту неплатежей по банковским кредитам, в результате банки начнут продавать заложенное у них имущество застройщиков.

Кроме организаций, также многим аналитикам интересна тема инструментов прогнозирования цен на недвижимость. Ежегодно публикуются тысячи статей на данную тему. В статье [15] описывается прогноз о взаимосвязи цен на недвижимость и заработную плату жителей Тамбовского района. Согласно основным показателям прогноза, при росте заработной платы Тамбовской области происходит прирост цен на недвижимость.

В данной же статье [13] авторы рассматривают методологии прогнозирования российского рынка недвижимости. “В публикации описываются такие методы как: статистический метод, метод изучения динамики рынка, метод нейронных сетей. Результаты приведенной классификации методов позволяют понять достоинства, недостатки и ограничения применяемых на рынке недвижимости методов и моделей.”

2.1. Методы оценивания стоимости недвижимости

Основными методами для оценивания рыночной стоимости объекта недвижимости являются:

1. Метод сравнительного анализа продаж;
2. Затратный метод;
3. Метод капитализации доходов [5].

“Метод сравнительного анализа продаж применяется в том случае, когда присутствует рынок земли и недвижимости, существуют реальные продажи, когда именно рынок формирует цены, и задача оценщиков заключается в том, чтобы анализировать этот рынок, сравнивать аналогичные продажи и таким образом получать стоимость оцениваемого объекта. Метод построен на сопоставлении предлагаемого для продажи объекта с рыночными аналогами. Метод сравнения продаж применяется при наличии достаточного количества достоверной рыночной информации о сделках купли-продажи объектов, аналогичных оцениваемому” [14].

Основы затратного метода - это изучение финансовых возможностей инвестора для приобретения недвижимости. Анализируется поведение инвестора, где инвестор заплатит наименьшую сумму для покупки объекта, в отличие от затрат если будет возведен аналогичный объект недвижимости. Данный метод может дать положительные

результаты, если корректно будет оценена точность стоимости и исследованы все предложения и спросы относительно рассмотренного объекта недвижимости.

“Метод капитализации доходов основывается на определении стоимости объекта оценки на основе текущей стоимости ожидаемых доходов от владения этим объектом. Основными методами доходного подхода являются: метод прямой капитализации дохода и метод дисконтирования денежного потока. Метод прямой капитализации применяется в случаях, когда прогнозируемый годовой чистый операционный доход является постоянным и не имеет четко выраженной тенденции к изменению, а период его получения не ограничен во времени. Метод дисконтирования денежного потока (непрямой капитализации) применяется в тех случаях, когда прогнозируемые денежные потоки от использования объекта оценки являются не одинаковыми по величине и непостоянными на протяжении выбранного периода прогнозирования. Методы доходного подхода целесообразно применять при рыночных базах оценки стоимости объекта” [2].

Для разработки инструмента прогнозирования стоимости реальных сделок между объектами на рынке продаж недвижимости будем рассматривать метод сравнительного анализа продаж.

Данный метод выбран для разработки, так как он основан на анализе объектов с аналогичными характеристиками.

Так как сравнительный метод рассматривается на анализе рыночных отношений с использованием установленной объективной мерой, ценой и качеством, его можно отнести к рыночному методу. Для сравнений принимают выбор покупателей среди выставленных в рекламу подобных предложений.

Для единицы сравнения при использовании сравнительного анализа применяются:

- стоимость площади участка;
- стоимость площади объектов недвижимости типа сделки “Аренда”;
- стоимость всей площади объекта;
- стоимость за одну комнату;
- стоимость за одну квартиру;
- стоимость недвижимости категории “Коммерческая недвижимость”.

Характеристики объекта недвижимости, которые влекут за собой изменение стоимости на недвижимость относят к элементам сравнения:

- условия продажи (в собственности менее 3 лет, наличие доли несовершеннолетних, ипотека, обременение, общая долевая собственность);
- месторасположение (город, район, улица, номер дома);
- физические характеристики (этаж, этажность дома, число комнат, тип планировки, количество спален, площадь, состояние квартиры), ;
- компоненты стоимости, не связанные с недвижимостью (парковка, двор, детская площадка, детские сады, школы) .

К числу достоинств данного метода относят:

- Большой объем вариантов применения, которые допускают нюансы целевого назначения расчётов.
- Доступность применения, которые допускают применять элементы подходов на простом уровне без учета нюансов.

- Для придания расчетам достоверность, информационная обоснованность опирается на статистику.
- Допускаются корректирующие расчеты при участии лиц, которые осведомлены.
- В других подходах разрешается использовать основы сравнительного анализа.

К недостаткам сравнительного метода относят:

- Попадание методик в зависимость от активности рынка продаж, откуда берется информация для использования.
- Информация устаревает из-за низкой активности того или иного типа жилья, что приводит к невозможности сравнительного анализа рынка недвижимости.
- Невозможно применить к уникальным объектам недвижимости, применяется только при стандартных характеристик объекта недвижимости.
- Только экспертам в области оценке недвижимости доступно выявлять особенности объекта и его нюансов, которые влияющих на образование стоимости.
- Необходимо вносить поправки, которые доступны только при наличии соответствующей квалификации.

2.2. Обзор методов прогнозирования

Многие базовые методы предсказаний относят скорее к отдельным приемам и процедурам прогнозирования, где другие представляют набор

отдельных приемов, которые отличаются от базовых или друг от друга количеством частных приемов и последовательностью их применения.

“Метод регрессионного анализа при проведении фундаментального анализа позволяет выявить важнейшие факторы, оказывающие влияние на изменение цены на основе эмпирико-индуктивных математических методов. Из-за включения в модель множества факторов, оказывающих воздействие в одно и то же время, такой подход также называется многофакторной моделью. Веса для оценки влияния различных факторов определяются на основе анализа массивов фактической информации о прошедших периодах. Всесторонние эконометрические расчеты на основе многих баз данных дают возможность получить в результате регрессионную модель, которая может быть использована для прогнозирования будущих изменений цен” [11].

“На рынке недвижимости России подобные исследования проводились профессором кафедры «Управление проектами и программами» Российского экономического университета им. Г.В.Плеханова, Г.М. Стерником на примере Москвы, Санкт-Петербурга и городов Московской области. Выявлено, что динамические закономерности рынка недвижимости в развивающихся экономиках существенно отличаются от классических законов функционирования рыночной экономики вследствие специфических факторов и условий в них. При этом учитываются и общее, присущее всем рынкам городов России, и отличия, несинхронность их поведения вследствие региональных различий” [3].

Одним из основных методов прогнозирования является метод нейронных сетей, который адаптирован для анализа рынка недвижимости

[6,7,12]. Из многофакторных аналитических методов моделирования построена нейронная сеть.

Нейронная сеть - самообучающейся система, в основе которой лежит действие нейронов человеческого мозга. Если не известна взаимосвязь между входными и выходными данными необходимо использовать нейронную сеть, данная особенность является преимуществом нейронных сетей. В таких случаях аналитику не требуется высказывать свое мнение о характере взаимосвязи между входными и выходными параметрами. В процессе обучения определяется вид связи и происходит настройка весовых коэффициентов. Также преимуществом нейронной сети является способность нахождения оптимальное предсказывание для данного ряда, путем нахождения для индикаторов для данного инструмента. Стратегии являются адаптивными, и могут меняться вместе с рынком, что очень важно для анализа активно развивающегося рынка недвижимости. Недостаток метода, – необходимо включать в расчет предварительный прогноз по каждому фактору.

Как показано в работе [6], возможно этот метод применять с другими, например с методом дискретного пространственно-параметрического моделирования.

В [12] приводится пример “нейросетевого моделирования рынков недвижимости г. Москвы. В рассматриваемом примере выходами сети являлись ставки аренды торговой и офисной недвижимости, а также цены на квартиры г. Москвы.

На вход подавались статистические данные о:

1. валовом внутреннем продукте;
2. объемах промышленного производства;
3. объемах продукции сельского хозяйства;

4. инвестициях в основной капитал;
5. индексе потребительских цен на товары и услуги;
6. общем уровне безработицы;
7. организованных сбережениях населения;
8. количества денег на руках у населения;
9. объемах экспорта;
10. объемах импорта;
11. курсе национальной валюты;
12. среднемесячной заработной плате;
13. уровне инфляции;
14. наличии или отсутствии парламентских (президентских) выборов;
15. политической стабильности.”

При моделировании в [12] использовалась так называемая нейронная сеть с одним скрытым слоем нейронов. “Сигналы, поступающие на входы X_1, \dots, X_n , умножаются на коэффициенты (веса, синапсы) соответствующие каждому входу и определяют уровень возбуждения нейрона. Выходной сигнал получается пропусканием суммарного сигнала возбужденных нейронов скрытого слоя через нелинейную функцию. В качестве нелинейной функции использовалась сигмоидная функция, являющаяся непрерывной и сколько угодно раз дифференцируемой.”

“Фундаментальным продвижением методологии моделирования и прогнозирования рынка недвижимости является работа [4], в которой сформированы уравнения спроса и предложения на рынке жилья.”

“Рассматриваемые модели спроса и предложения учитывают следующие факторы рынка недвижимости: темп роста цен на жилье;

ожидаемая инфляция будущего периода; ожидаемая доходность от альтернативных инвестиций; степень доступности кредитных средств; предложение жилья на первичном рынке; прирост по отношению к прошлому году индекса РТС; реальный доход домохозяйства; суммарный запас жилищного фонда на душу населения; издержки строительства; доля инвестиций в нежилое строительство; стоимость строительно-монтажных работ. При всей ценности данной работы в теоретическом отношении, в прагматическом аспекте необходимо отметить, что использование полученных авторами уравнений спроса и предложения для прогнозирования требует, как и в случае обычных регрессионных многофакторных моделей и нейросетевых моделей, предварительного прогноза динамики каждого из рассматриваемых факторов, что зачастую нереализуемо” [13].

Проведенные исследования в работе [25] показывают, что при наличии достоинствах нейронной сети, которые очевидны, присутствует и ряд недостатков, к ним относятся склонность сети на срабатывание по отношению к ложным реакциям и низкая эффективность при необходимости доучивании, когда требуется добавить новые данные, не изменяю информацию, которая была запомнена ранее.

Также недостатком нейронной сети с точки зрения прогнозирования является то, что сначала необходимо сформулировать эталон распознаваемых образов. Но эталоны должны формироваться на основе анализа в самой сети [25].

Вывод

В современных условиях рынка существует множество различных методов и приемов как для оценки реальной рыночной стоимости недвижимости, так и для прогнозирования изменения цены на

недвижимость. Невозможно выделить или унифицировать какой-то один метод оценки рыночной стоимости недвижимости, по причине того, что у каждого способа существуют свои особенности. Таким образом, в данной работе будут рассмотрены и применены методы машинного обучения для выявления значимых параметров для прогнозирования цены на недвижимость такие как: метод случайного леса, метод ближайших соседей, метод градиентного бустинга.

Глава 3. Разработка модели

3.1 Предобработка набора данных

Для работы был взят исходный набор данных, состоящий из 801 объекта недвижимости и 147 характеристик объектов вторичной недвижимости типа “Квартира” по городу Тюмени. Данный тип недвижимости выбран по причине того, что выборка проданных объектов в риэлторской информационной системе составляет 60 процентов от всех проданных объектов. Кроме того, при прогнозировании объектов типа “Квартира” рассматриваются характеристики самого объекта без внешних факторов.

Выборка данных по объектам предоставляет довольно большой объем параметров. Не все параметры необходимы для построения модели прогнозирования, существует минимальный набор параметров для оценки недвижимости на основе Российского стандарта описания квартиры [24]. В связи с этим необходимо внести изменения в структуру набора данных и исключить не информативные параметры.

Во всех случаях рассмотрены объекты, сделки по которым совершались в целом, а не по долям. При анализе данные примеры не рассматривались потому что, объекты по долям являются выбросами, так как набор параметров для данных объектов не будут соответствовать ценовой категории объектов, у которых собственник недвижимости представлен одним лицом.

Дата сделки - это дата подписания договора купли продажи объекта недвижимости. Дата сделки с формата ММ.ГГ изменена на количество месяцев, начиная с января 2006 года, так как это первая дата,

встречающаяся в используемых данных. При построения модели прогнозирования можно использовать только числовые значения.

Поскольку использовать адрес в математической модели в строковом формате не представляется возможным, то необходимо выбрать другой способ представления адреса. В связи с тем, что адрес в нашей выборке отображен только улицей заменить текст будем частотой встречаемости необходимой улицы в представленных данных.

Были удалены столбцы, по мнению эксперта, работающего в агентстве недвижимости по городу Тюмени, не носящие ценной информации для прогноза: класс, тип (продажа/аренда), статус, код агента, город, количество фото, наличие фотографий, наличие баннера, комментарии, дополнительный комментарий, количество собственников, ключи в офисе, дата приостановления, причина приостановления. На пример, на рисунке 2 отображена взаимосвязь между ценой за 1 кв.м и параметром “Наличие баннера”.

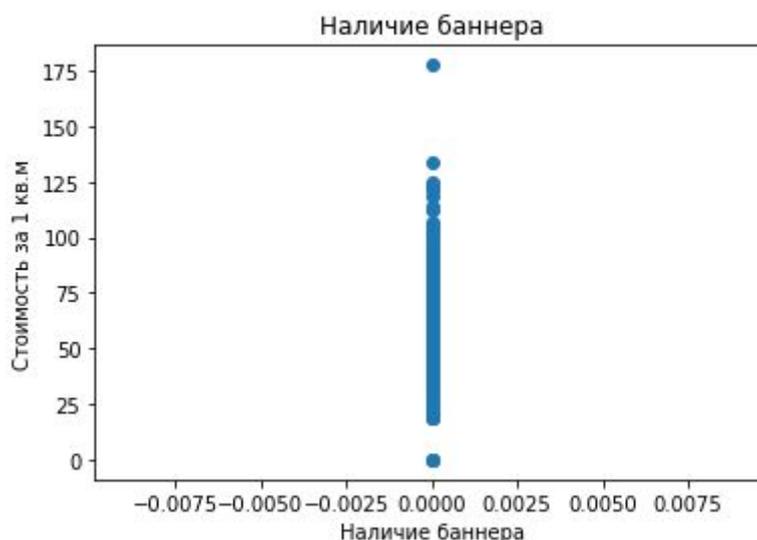


Рисунок 2 - Зависимость стоимости 1 кв.м от параметра “Наличие баннера”

Такие признаки как ванная комната и ванная комнаты, количество ванных комнат и количество сан.узлов, ипотека и ипотека, сумма ипотеки и сумма одобренной ипотеки, описание расположения дома и расположение дома удаляем, так как данные признаки дублируются в структуре. Признаки метро и время до метро убираем из структуры, потому что метро отсутствует в городе Тюмени. Признаки: ипотека, сумма ипотеки, банк ипотеки удаляем, данные признаки взаимосвязаны с признаками: обременения, сумма обременения, банк обременения. Признак кадастрового номера необходимо исключить – данный признак для каждого объекта уникальный.

Признаки, используемые для обучения модели, оказывают большое влияние на качество результатов. Неинформативные или слабо информативные признаки могут существенно понизить эффективность модели.

Отбор признаков – это процесс выбора признаков, имеющих наиболее тесные взаимосвязи с целевой переменной.

Присутствие в данных неинформативных признаков приводит к снижению точности многих моделей, особенно линейных, таких как линейная и логистическая регрессия.

Отбор признаков перед моделированием обеспечивает три следующих преимущества:

- Уменьшение переобучения. Чем меньше избыточных данных, тем меньше возможностей для модели принимать решения на основе «шума».

- Повышение точности. Чем меньше противоречивых данных, тем выше точность.

- Сокращение времени обучения. Чем меньше данных, тем быстрее обучается модель.

Рекурсивное исключение признаков

Метод рекурсивного исключения признаков (recursive feature elimination, RFE) реализует алгоритм:

- модель обучается на исходном наборе признаков;
- оценивает их значимость;
- исключает один или несколько наименее значимых признаков;
- модель обучается на оставшихся признаках, и так далее, пока не останется заданное количество лучших признаков.

Учитывая внешнюю оценку, которая присваивает весовые характеристики (например, коэффициенты линейной модели), целью рекурсивного устранения элемента (RFE) является выбор функций путем рекурсивного рассмотрения меньших и меньших наборов признаков. Во-первых, оценщик обучается по первому набору признаков, и значимость каждой функции получается либо через coef_атрибут, либо через feature_importances_ атрибут. Затем наименее значимые признаки обрезаются из текущего набора функций. Результат вычислений данного метода представлен в таблице 5 в столбце “Результат 1”.

Метод главных компонент

Метод главных компонент (principal component analysis, PCA) позволяет уменьшить размерность данных с помощью преобразования на основе линейной алгебры. Пользователь может задать требуемое количество измерений (главных компонент) в результирующих данных.

Уменьшение линейной размерности с использованием сингулярного значения Разложение данных для проецирования его в более низкое пространство. Результат вычислений данного метода представлен в таблице 5 в столбце “Результат 2”.

Отбор на основе важности признаков

Ансамблевые алгоритмы на основе деревьев решений, такие как случайный лес (random forest), позволяют оценить важность признаков.

Классификатор дополнительных деревьев реализует мета-оценку, которая подходит для ряда рандомизированных деревьев решений (aka extra-trees) для различных подэмпллов набора данных и использует усреднение для повышения точности прогнозирования и контроля над переустановкой. Чем больше значение оценки, тем важнее признак.

Результат вычислений данного метода представлен в таблице 5 в столбце “Результат 3”.

На вход поступают:

X – пространство объектов (множество всех возможных объектов, для которых может потребоваться предсказание);

Y – пространство ответов (множество всех возможных ответов).

Результаты отбора по каждой функции отображено в таблице 1.

Таблица 1 Результаты отбора с помощью методов

Признак	Результат 1	Результат 2	Результат 3
District_id	False	0.0263158	0.996371
Street_id	False	0	0.00306976
Photos_cnt	True	0.0473684	0.000264741

Date_create	True	0.0315789	0.000110487
Date_update	True	0.0421053	8.17027e-05
Trade_in	True	0.0105263	6.63961e-05
Auction	True	0.0105263	2.37044e-05
Cadaster_number	True	0.0210526	4.74424e-06
Floor_num	True	0.0526316	3.22975e-06
Floors_cnt	True	0.0631579	2.0341e-06
Rooms_cnt	True	0.0684211	8.58137e-07
Bedrooms_cnt	True	0.0210526	3.55135e-07
Building_year	False	0.0736842	2.07637e-07
Area_total	True	0.0526316	1.1785e-07
Area_live	True	0.0105263	9.42959e-08
Area_kitchen	True	0.0421053	8.58345e-08
Area_balcony	False	0	3.14888e-08
Secondary	False	0	1.30621e-08
Type	True	0.0578947	6.78682e-09
Bathroom	True	0.0157895	8.12415e-34
Bathrooms_cnt	True	0.0157895	0.996371
Plate	True	0.0421053	0.00306976
Window	True	0.0157895	0.000264741
Keep	False	0.0526316	0.000110487
Komunal_cost	False	0.0157895	8.17027e-05
Series_id	True	0.0368421	6.63961e-05
Wall_id	True	0.0157895	2.37044e-05
Young_family	True	0.0210526	4.74424e-06
Subsidy	True	0.0105263	3.22975e-06
Balcon	False	0.0842105	2.0341e-06

Loggia	False	0	8.58137e-07
Ceiling_height	True	0.0421053	3.55135e-07
Parking_cnt	False	0.0263158	2.07637e-07

Сравнив полученные результаты всех методов, значимыми признаками являются: район, улица, количество фотографий, дата создания, дата редактирования, этаж, этажность, количество комнат, площадь, количество ванных комнат, окно, ремонт, серия дома, балкон, высота потолков.

Таким образом, в полученном наборе данных было оставлено 800 объектов недвижимости и 109 характеристик объектов.

Рассмотрели 3 метода прогнозирования:

- метод случайного леса [23],
- градиентный бустинг [18],
- метод ближайших соседей [20].

Модель для прогнозирования - RandomForestRegressor. Осуществлялось случайное деление на тестовую и обучающую выборку в процентном соотношении 40:60 соответственно при помощи метода `train_test_split()`. Данное деление необходимо для того, чтобы избежать переобучение модели. На обучающей выборке происходит обучение модели. На тестовой выборке осуществляем проверку построенной модели. Данные выборки не должны пересекаться. В результате вычисления точности на данных параметров оценка составила - 0.06988374984208612 в процентном соотношении 6,98% при дельта равное 0. Количество, правильно определенных объектов, составило 23.

```

x_train, x_test, y_train, y_test = model_selection.train_test_split(x, y,
test_size=0.4,random_state=1)

model1=ensemble.RandomForestRegressor()

model1 = model1.fit(x_train, y_train)

label_predict1 = model1.predict(x_test)

l2_model1 = model1.score(x_test, y_test)

```

Листинг 1. Построение модели RandomForestRegressor

Модель для прогнозирования - KNeighborsRegressor. Случайное деление на тестовую и обучающую выборку в процентном соотношении 40:60 соответственно при помощи метода train_test_split() бралось предыдущее. В результате вычисления точности на данных параметров оценка составила - 0.00474027924675724 в процентном соотношении 0,47% при дельта равное 0. Количество, правильно определенных объектов, составило 2.

```

model1=KNeighborsRegressor(n_neighbors=6)

model1 = model1.fit(x_train, y_train)

label_predict1 = model1.predict(x_test)

l2_model1 = model1.score(x_test, y_test)

```

Листинг 2. Построение модели KNeighborsRegressor

Модель для прогноирования - GradientBoostingRegressor. Случайное деление на тестовую и обучающую выборку в процентном соотношении 40:60 соответственно при помощи метода train_test_split() бралось предыдущее. В результате вычисления точности на данных параметров оценка составила - 0.08318414306602084 в процентном соотношении 8,31% при дельта равное 0. Количество, правильно определенных объектов, составило 27.

```

clf=ensemble.GradientBoostingRegressor()

```

```
clf = clf.fit(X_train,y_train)

clf_predict1 = clf.predict(X_test)

l2_clf = clf.score(X_test, y_test)
```

Листинг 3. Построение модели GradientBoostingRegressor

Данная оценка очень мала по причине того, что в исходном наборе данных не осуществлена предобработка данных и не проведена нормализация, но она выше из всех рассмотренных методов ранее.

3.2 Нормализация данных

Для произвольного объекта недвижимости требуется по общей информации об объекте недвижимости и данных о его параметрах за предыдущий период научиться предсказывать стоимость объекта недвижимости в данный момент времени.

“Обучение с учителем (Supervised learning) — один из разделов машинного обучения, посвященный решению задачи, у которой имеется множество объектов (ситуаций) и множество возможных ответов (откликов, реакций). Существует некоторая зависимость между ответами и объектами, но она неизвестна. Известна только конечная совокупность прецедентов — пар «объект, ответ», называемая обучающей выборкой. На основе этих данных требуется восстановить зависимость, то есть построить алгоритм, способный для любого объекта выдать достаточно точный ответ. Для измерения точности ответов определенным образом вводится функционал качества.

Под учителем понимается либо сама обучающая выборка, либо тот, кто указал на заданных объектах правильные ответы” [17].

Так как в работе для построения классификатора предполагается использование алгоритмов машинного обучения с учителем, одним из

необходимых этапов обработки данных является разделение параметров на два класса - числовые и категориальные (листинг 1).

```
categorical_columns = [c for c in data.columns if data[c].dtype.name == 'object']
numerical_columns = [c for c in data.columns if data[c].dtype.name != 'object']
print categorical_columns
print numerical_columns
```

Листинг 1. Числовые и категориальные признаки

Для каждого класса можно получить некоторую общую информацию. Для каждого категориального параметра приведено общее число заполненных ячеек (`count`), количество значений, которые принимает данный признак (`unique`), часто встречающееся значение этого признака (`top`) и количество объектов, в которых встречаются самое частое значение данного признака (`freq`) (листинг 2).

```
count      800
unique      9
top    yes_balcon
freq       291
Name: balcon, dtype: object
```

Листинг 2. Общая информация по параметру

Алгоритмы машинного обучения не работают напрямую с категориальными признаками и данными, в которых имеются пропущенные значения. Для начала необходимо подготовить данные.

Простая стратегия - заполнение пропущенных значений самым популярным в столбце. Прежде чем подавать данные на вход алгоритмов машинного обучения преобразуем категориальные признаки в количественные. Категориальные признаки, которые могут принимать два значения и принимать большее количество значений будем обрабатывать по разному. Необходимо выделить бинарные и не бинарные признаки (листинг 3).

```
binary_columns = [c for c in categorical_columns if data.describe[c]['unique'] == 2]
nonbinary_columns = [c for c in categorical_columns if data.describe[c]['unique'] > 2]
print binary_columns, nonbinary_columns
```

Листинг 3. Определение бинарных и не бинарных признаков

Для бинарных признаков будет просто замена на 0 и 1 (листинг 4).

```
for c in binary_columns[1:]:
    top = data_describe[c]['top']
    top_items = data[c] == top
    data.loc[top_items, c] = 0
    data.loc[~top_items, c] = 1
```

Листинг 4. Цикл замены бинарных признаков на 0 и 1

Для не бинарных признаков каждому уникальному значению в столбце присваиваем целое число. По данному параметру заменяем строковые значения на присвоенное число (листинг 5).

```
df.at[df['encumbrance_bank'] == 'Сбербанк', 'encumbrance_bank'] = 1
df.at[df['encumbrance_bank'] == 'АИЖК', 'encumbrance_bank'] = 2
df.at[df['encumbrance_bank'] == 'Абсолют банк', 'encumbrance_bank'] = 3
df.at[df['encumbrance_bank'] == 'ВТБ ', 'encumbrance_bank'] = 4
df.at[df['encumbrance_bank'] == 'ВТБ 24', 'encumbrance_bank'] = 5
df.at[df['encumbrance_bank'] == 'Газпромбанк', 'encumbrance_bank'] = 6
df.at[df['encumbrance_bank'] == 'Запсибкомбанк', 'encumbrance_bank'] = 7
df.at[df['encumbrance_bank'] == 'СКБ-банк', 'encumbrance_bank'] = 8
```

Листинг 5. Замена не бинарных признаков

Числовые признаки будем обрабатывать все сразу. Необходимо узнать количество заполненных элементов, для этого воспользуемся методом count. Параметр axis = 0 указывает, что мы движемся по размерности 0 (сверху вниз), нас интересует количество заполненных элементов в каждом столбце (Листинг 6).

```
id          801
newhouses_id  801
class       801
action      801
direction   801
status      801
user_id     801
last_editor_id  801
city_id     801
district_id  801
street_id   801
on_map      801
photos_cnt  801
tours_cnt   801
attach_cnt  801
comments_cnt 801
date_create 801
date_update 801
date_rise   801
```

```

rise            801
date_locked    138
mortgage       596
mortgage_summ  0
trade_in       798
contract_num   399
contract_type  396
contract_cost  797
contract_date  396
contract_raschets 0
express_contract 801
....
has2           797
has3           44
has4           1
komunal_cost   52
notes          801
series_id      800
wall_id        800
young_family   800
subsidy        800
more_than_3    800
allowed        0
people_count   0
beds_count     0
balcon         800
loggia         800
apartment_letter 25
ceiling_height 800
set_ceiling_height 799
total_rooms_cnt 799
sold_rooms_cnt 799
closed_yard    0
area_rooms     95
studio_type    728
parking_cnt    796
house_location 66
rooms_comfort  66
pond_type      0
pond_distance  0
layout_type    0
in_city        653

```

Листинг 6. Общая информация о параметрах.

Если данные содержат пропущенные значения, то можно:

- удалить столбцы с пропущенными значениями,
- удалить строки с пропущенными значениями.

Данные варианты могут привести к потере данных, поэтому рассмотрим простой способ заполнения пропущенных значений. Заполнить пропущенные значения можно с помощью метода fillna, медианными значениями. (листинг 7)

```
id          801
newhouses_id  801
direction    801
district_id  801
street_id    801
on_map       801
photos_cnt   801
tours_cnt    801
attach_cnt   801
comments_cnt 801
rise         801
mortgage     801
mortgage_summ 801
trade_in     801
contract_num  801
contract_type 801
contract_cost 801
contract_date 801
contract_raschets 801
express_contract 801
...
has2         801
has3         801
has4         801
komunal_cost 801
notes        801
series_id    801
wall_id      801
young_family 801
subsidy      801
more_than_3  801
allowed      801
people_count 801
beds_count   801
balcon       801
loggia       801
apartment_letter 801
ceiling_height 801
set_ceiling_height 801
total_rooms_cnt 801
sold_rooms_cnt 801
closed_yard  801
area_rooms   801
studio_type  801
parking_cnt  801
```

```
house_location    801
rooms_comfort     801
pond_type         801
pond_distance     801
layout_type       801
in_city           801
```

Листинг 7. Заполненные пропущенные значения.

Так как алгоритмы машинного обучения чувствительны к масштабированию данных, все признаки необходимо нормализовать. Это можно сделать путем приведения признака к нулевому среднему и единичному среднеквадратичному отклонению (рисунок 3).

Index	newhouses_id	district_id	street_id	on_map	photos_cnt	tours_cnt	attach_cnt
0	-0.932837	-0.605222	-0.251872	-0.229122	-1.03533	-0.877671	-0.05
1	-0.932837	-0.605222	-0.251872	-0.229122	-0.0703273	-0.877671	-0.05
2	-0.932837	-0.605222	-0.251872	-0.229122	0.894678	-0.877671	-0.05
3	2.84876	-0.319163	5.21951	-0.229122	-0.914707	-0.877671	-0.05
4	-0.413641	1.5726	-0.205651	-0.229122	-1.15596	-0.877671	-0.05
5	-0.534963	-0.584739	-0.0643414	-0.229122	0.170924	-0.446638	-0.05
6	-0.00342048	1.57208	-0.202849	-0.229122	-0.311579	-0.877671	-0.05
7	-0.389261	1.57208	-0.183552	-0.229122	-2.36222	-0.877671	-0.05
8	1.39768	1.57225	-0.205651	-0.229122	-0.190953	-0.877671	-0.05
9	-0.461553	-0.580363	-0.210164	-0.229122	1.25656	-0.877671	-0.05
10	-0.410018	1.57208	-0.204406	-0.229122	0.532801	-0.877671	-0.05
11	-0.473631	-0.581588	-0.200359	-0.229122	1.61843	1.27749	-0.05
12	-0.428046	-0.535546	-0.201449	-0.229122	-1.63846	-0.877671	-0.05
13	-0.392303	1.57225	-0.189154	-0.229122	-0.914707	-0.0156054	-0.05

Рисунок 3. Нормализация путем приведения признака

Нормализация предполагает замену признаков так, чтобы каждый из них лежал в диапазоне от 0 до 1. В Scikit-Learn уже есть готовая для этого функция `normalize` (рисунок 4).

	0	1	2	3	4	5	6
0	0	0	0	0	8.93427e-09	0	0
1	0	0	0	0	1.54319e-08	0	0
2	0	0	0	0	2.19296e-08	0	0
3	0.922374	0.0178292	0.383612	0	0.000130937	0	0
4	0.652792	0.699701	0.0167051	0	0.000562461	0	0
5	0.914797	0.0120341	0.123941	0	0.00215996	0.000102855	0
6	0.752224	0.450299	0.011405	0	0.000615509	0	0
7	0.193654	0.19826	0.00699646	0	0	0	0
8	0.960225	0.229255	0.00547426	0	0.000331773	0	0
9	0.790695	0.0106577	0.0201145	0	0.00225162	0	0
10	0.41273	0.439216	0.0107712	0	0.000847567	0	0
11	0.908228	0.0119446	0.0292863	0	0.00291978	0.000442391	0
12	0.949293	0.0334827	0.0272573	0	0.000504764	0	0

Рисунок 4. Нормализация с помощью встроенной функции normalize

После этапа нормализации данных остается 76 признаков и один целевой.

Модель для прогнозирования - RandomForestRegressor. Осуществлялось случайное деление на тестовую и обучающую выборку в процентном соотношении 40:60 соответственно при помощи метода `train_test_split()`. В результате вычисления точности для 76 признаков оценка составила - 0.8627139175090569 в процентном соотношении 86,2% при дельта равное 0. Количество, правильно определенных объектов, составило 277.

Модель для прогнозирования - KNeighborsRegressor. Осуществлялось случайное деление на тестовую и обучающую выборку в процентном соотношении 40:60 соответственно при помощи метода `train_test_split()`. В результате вычисления точности для 76 признаков оценка составила - 0.47263770511597486 в процентном соотношении

47,2% при дельта равное 0. Количество, правильно определенных объектов, составило 152.

Модель для прогнозирования - GradientBoostingRegressor. Осуществлялось случайное деление на тестовую и обучающую выборку в процентном соотношении 40:60 соответственно при помощи метода `train_test_split()`. В результате вычисления точности для 76 признаков оценка составила - 0.5094947572884383 в процентном соотношении 50,9% при дельта равное 0. Количество, правильно определенных объектов, составило 164.

Данные оценки получены на данных, в которых не определены параметры, которые влияют на стоимость одного квадратного метра.

3.3 Корреляция данных.

На этапе обработки данных была рассмотрена возможность наличия линейной зависимости между независимыми переменными и зависимой. С этой целью был вычислен коэффициент ранговой корреляции Спирмена для каждого признака и зависимой переменной. Для вычисления коэффициента корреляции необходимо все категориальные признаки преобразовать в количественный тип. Данное преобразование осуществляем при помощи замены частоты встречаемости категорий в выборке данных для признаков: район, улица, банк обременения, застройщик, тип (квартира/общежитие/комната), плита (электрическая/газовая), окна, ремонт, серия дома, тип стен.

Результаты вычислений приведены в таблице 2.

Таблица 2 – Значения коэффициента корреляции

Параметр	Коэффициент корреляции
----------	------------------------

District_id	0,206455
Street_id	0
Photos_cnt	0,213479
Date_create	0,02429658
Date_update	-0,311662199
Trade_in	-0,27288
Auction	-0,06165
Cadaster_number	0
Floor_num	0,215459
Floors_cnt	0,116035
Bedrooms_cnt	0,629391
Building_year	0,325199
Area_total	0,952669
Area_live	0,031676
Area_kitchen	0,530456
Area_balcony	0
Secondary	0
Type	0,139054
Bathroom	0,002387
Bathrooms_cnt	0,904627
Plate	0,086175

Windows	0,221851
Keep	-0,35734
Komunal_cost	0
Series_id	0,318044
Wall_id	0,365028
Young_family	0,140178
Subsidy	0,140178
Balcon	0,110512367
Loggia	0
Ceiling_height	-0,01851
Rooms_cnt	0,725609
Parking_cnt	0

По результатам расчетов на рисунке 3 построена диаграмма абсолютных значений коэффициента корреляции.

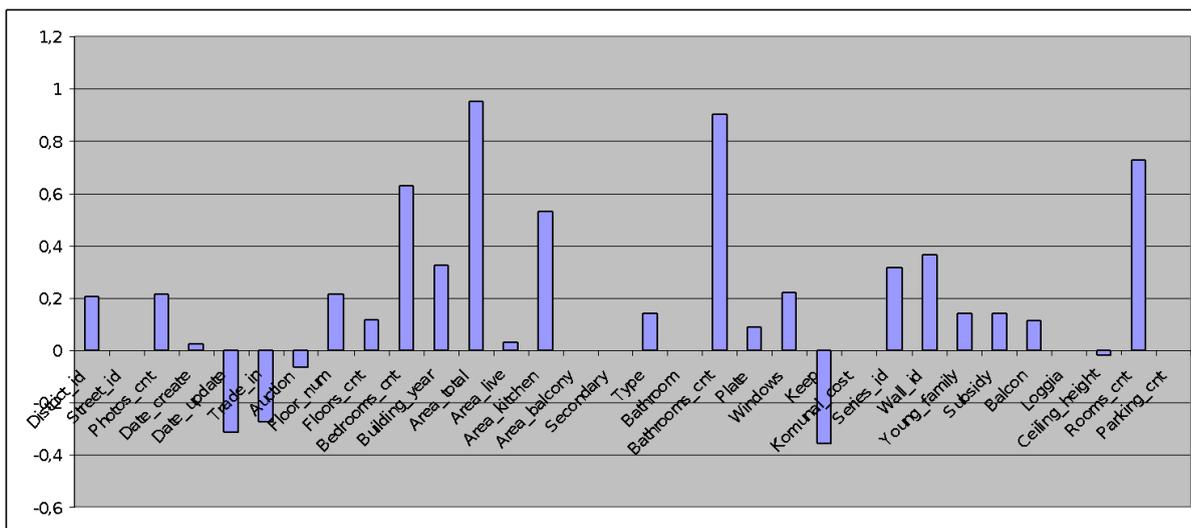


Рисунок 3 – Коэффициенты корреляции

Из диаграммы видно, что между признаками улица, дата изменения, обмен, торг, площадь балкон, вторичное жилье, ремонт, стоимость коммунальных расходов, лоджия, высота потолков, количество парковочных мест и целевой функцией price отсутствует линейная зависимость. Между признаками количество ванных комнат, общая площадь, площадь кухни, количество комнат и целевой функцией price присутствует линейная зависимость, что соответствует общим представлениям:

- с увеличением площади, увеличивается стоимость;
- с увеличением количеством комнат, стоимость увеличивается;
- с наличием раздельного санузла, увеличивается.

Очень слабая связь между признаками район, количество фотографий, дата создания, этаж, этажность, год постройки, жилая площадь, тип, плита, окно, тип серий, тип стен, молодая семья, субсидии, балкон и цена.

В математической модели в качестве зависимой переменной, для которой будет проводиться прогнозирование, используется поле

стоимость (Price). Остальные поля выступают как независимые переменные, признаки.

На этапе исследования корреляции было выявлено, что зависимость между параметрами присутствует. Для уточнения какая зависимость и между какими параметрами присутствует была построена матрица корреляции представлена на рисунке 5.

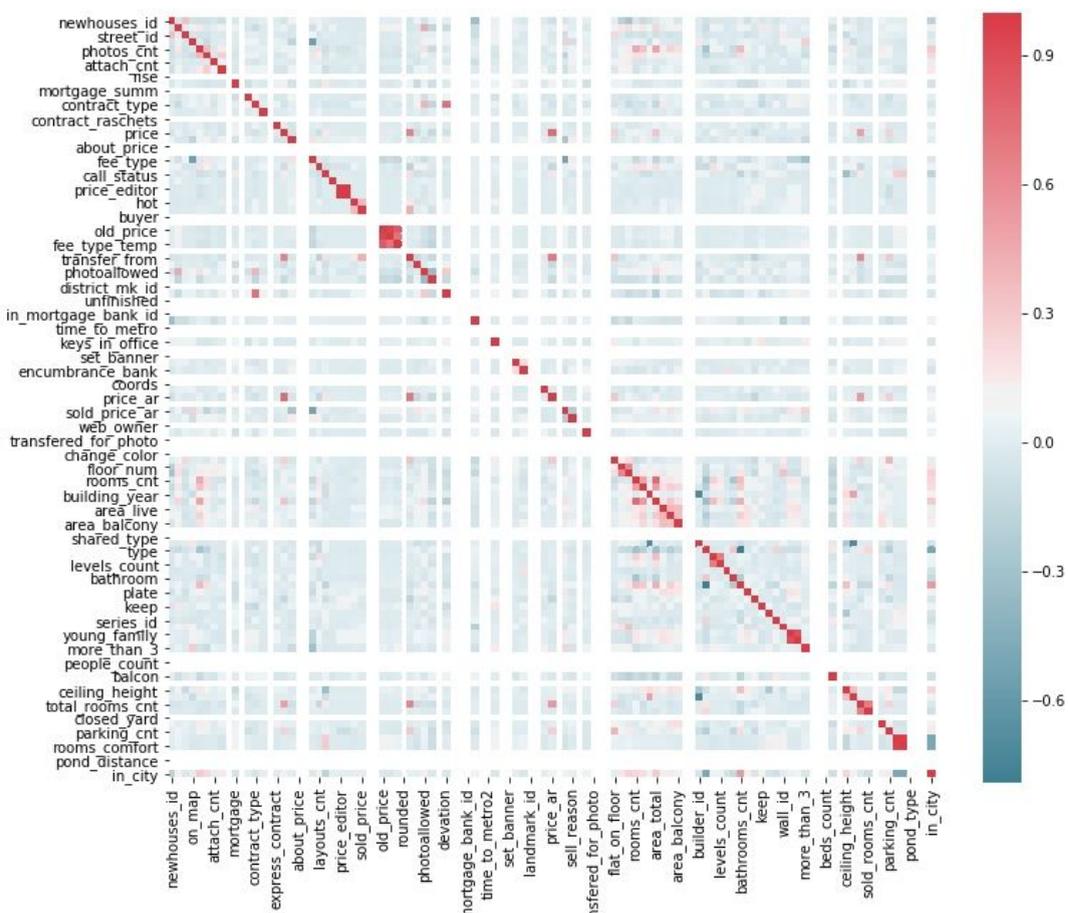


Рисунок 5 – Матрица корреляции для 76 признаков

Из рисунка матрицы корреляции видно, что присутствует взаимосвязь между параметрами. Для определения веса каждого параметра в модели прогнозирования проанализируем поведение параметров в уравнение регрессии.

Уравнение регрессии представлено в формуле 1.

$$y_x = r_{xy} * \frac{x - \bar{x}}{S(x)} * S(y) + \bar{y} \quad (1)$$

где r_{xy} - коэффициент корреляции, \bar{x} , \bar{y} - выборочные средние, $S(x)$, $S(y)$ - среднеквадратические отклонения.

Параметр “Этаж”

Уровень здания, где располагается объект недвижимости.

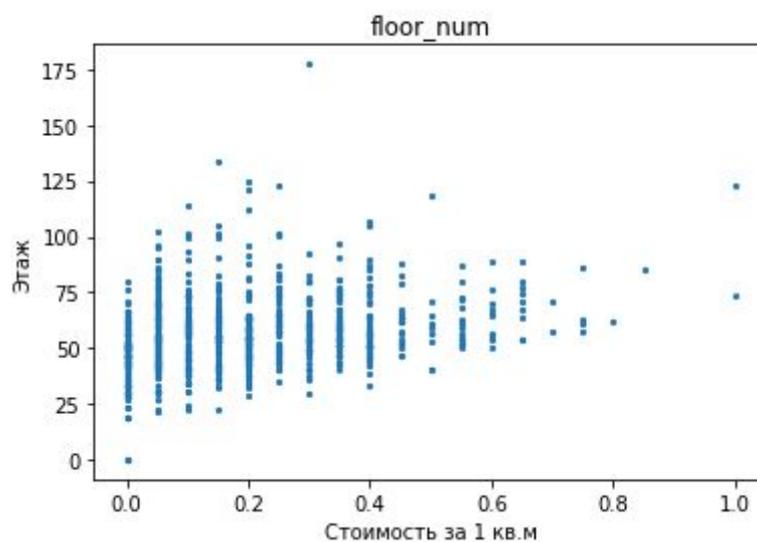


Рисунок 6. Параметр “Этаж”

Необходимо рассмотреть параметр “Этаж” на наличие выбросов. Используя метод LocalOutlierFactor, рассмотрим данный параметр без выбросов, найдем коэффициент уравнения и свободный член уравнения.

На ось абсцисс взяты значения параметра “Этаж”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

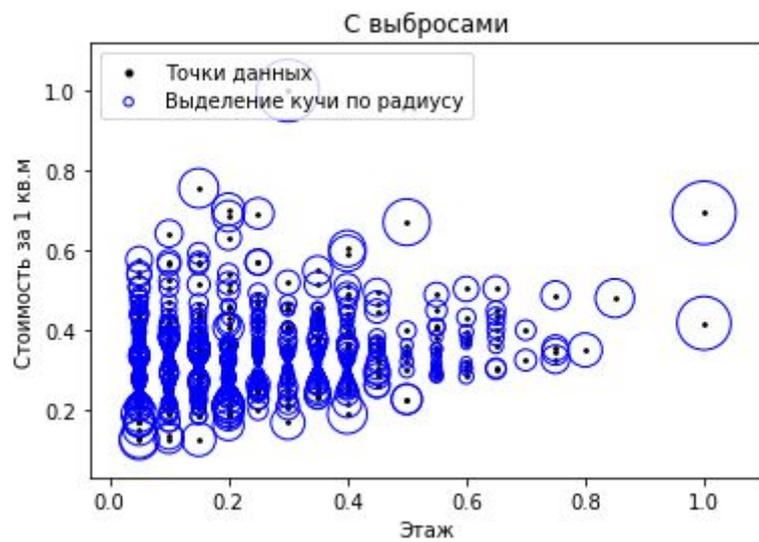


Рисунок 6. Параметр “Этаж” с выбросами

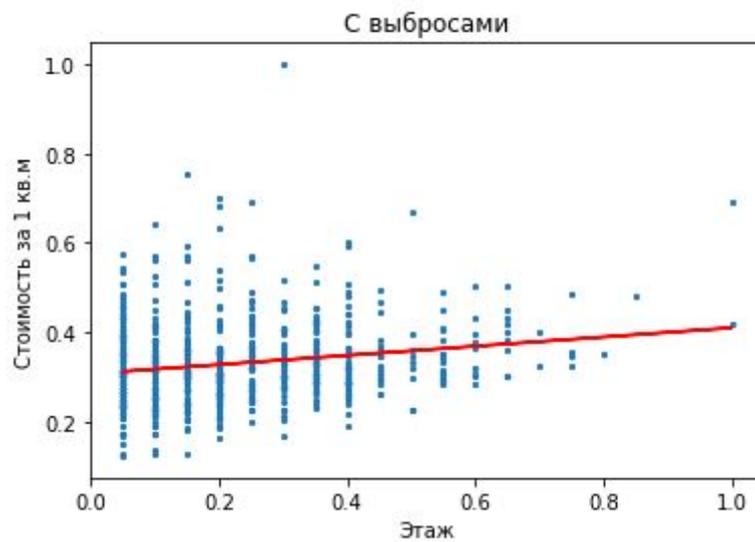


Рисунок 7. Параметр “Этаж” с выбросами

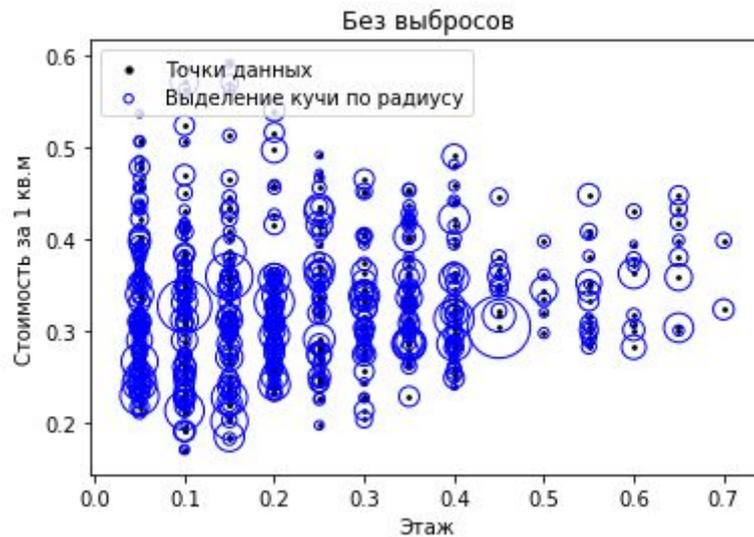


Рисунок 8. Параметр “Этаж” без выбросов

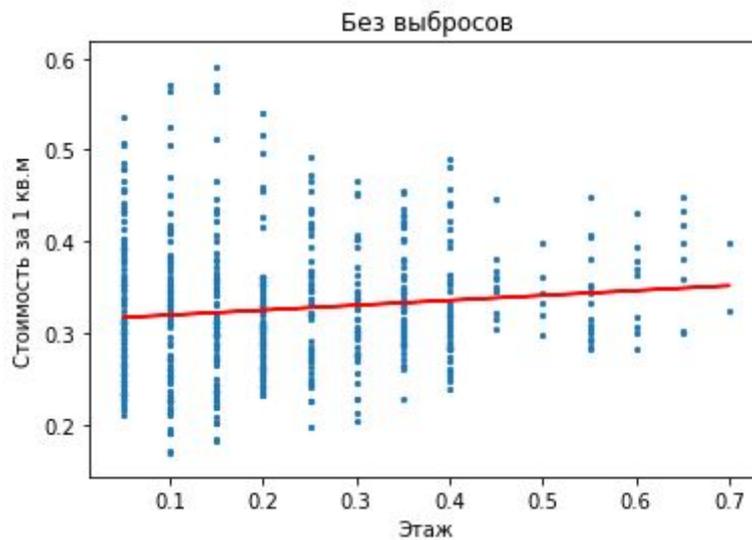


Рисунок 9. Параметр “Этаж” без выбросов

Параметр “Год постройки”

Год завершения строительства объекта недвижимости. Год постройки необходимо уточнять из документа кадастрового паспорта или кадастровой выписки.

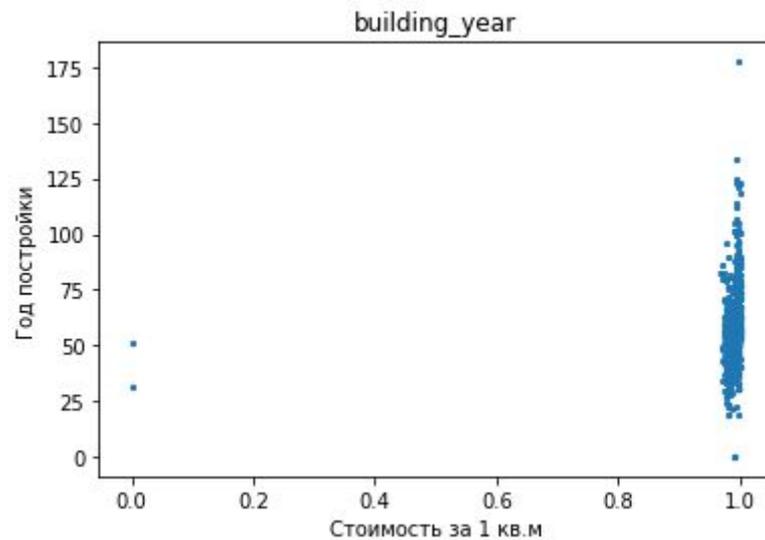


Рисунок 10. Параметр “Год постройки”

Из рисунка 6 видно, что необходимо рассматривать данные без нулевых значений, так как нулевые значения уменьшают точность вычислений. Используя метод LocalOutlierFactor рассмотрим параметр “Год постройки” и целевую функцию. Для них найдем значение коэффициента уравнения и свободный член уравнения.

На ось абсцисс взяты значения параметра “Год постройки”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

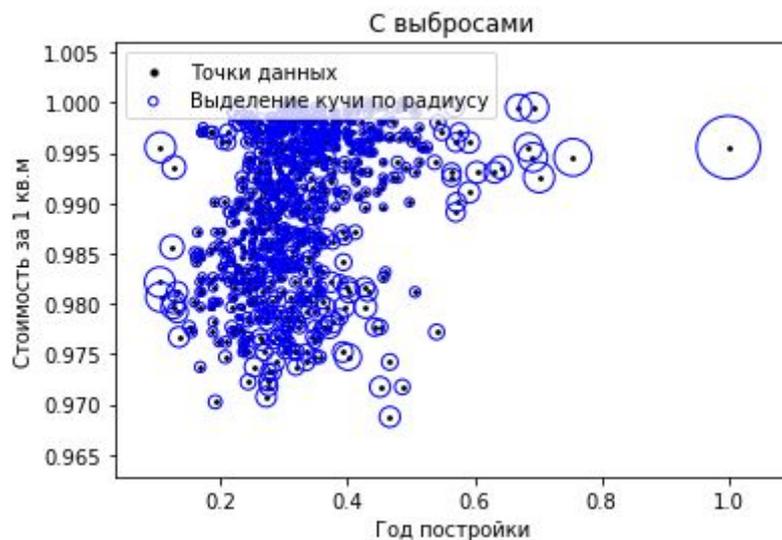


Рисунок 11. Параметр “Год постройки” с выбросами

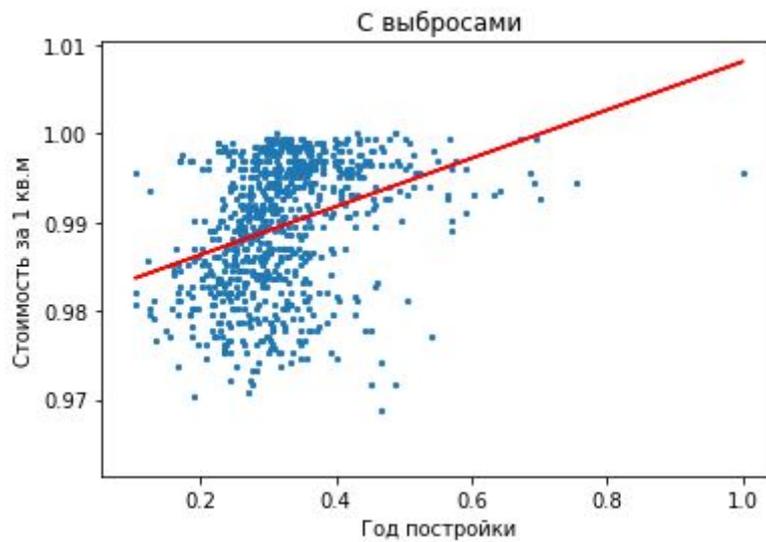


Рисунок 12. Параметр “Год постройки” с выбросами

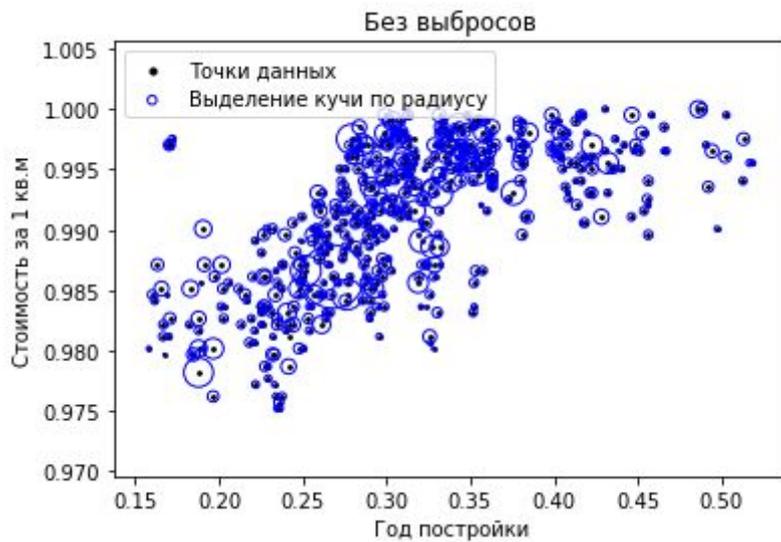


Рисунок 13. Параметр “Год постройки” без выбросов

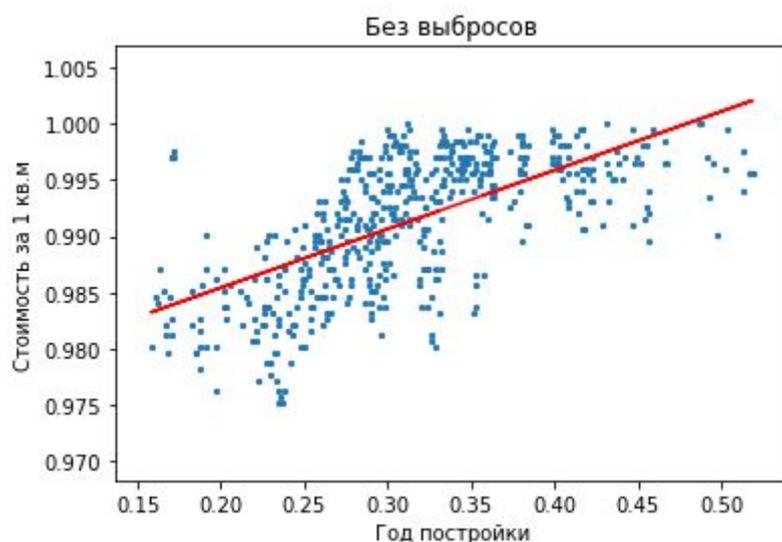


Рисунок 14. Параметр “Год постройки” без выбросов

Параметр “Общая площадь”

Общая площадь жилого помещения состоит из суммы площади всех частей такого помещения, включая площадь помещений вспомогательного использования, предназначенных для удовлетворения гражданами бытовых и иных нужд, связанных с их проживанием в жилом помещении, за исключением балконов, лоджий, веранд и террас.

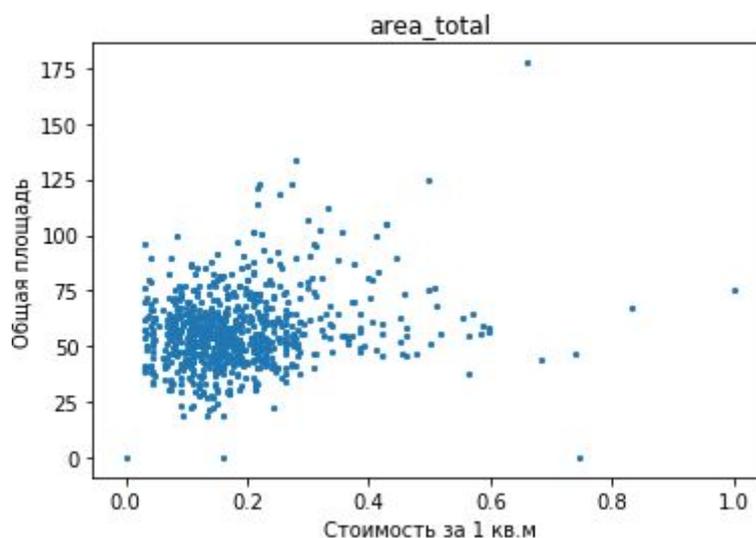


Рисунок 15. Параметр “Общая площадь”

Проанализировав рисунок 11, находим коэффициент уравнения и свободный член уравнения, убирая выбросы из набора данных для параметра “Общая площадь” и целевая функция. Количество выбросов составило 40.

На ось абсцисс взяты значения параметра “Общая площадь”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

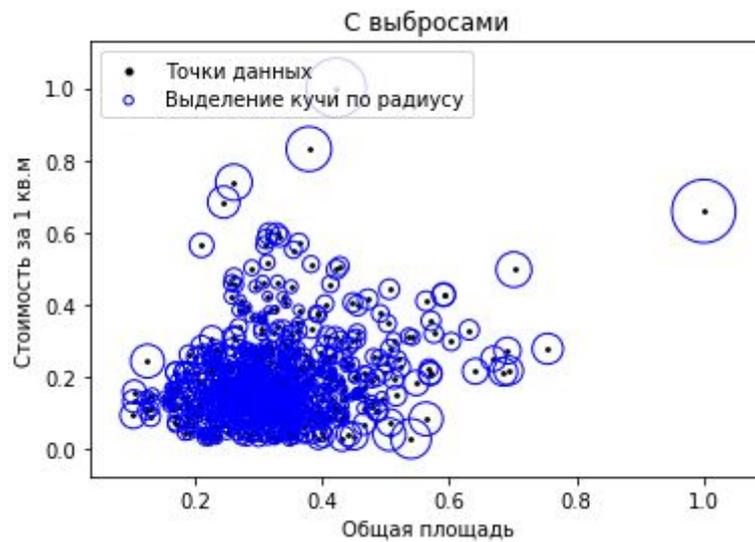


Рисунок 16. Параметр “Общая площадь” с выбросами

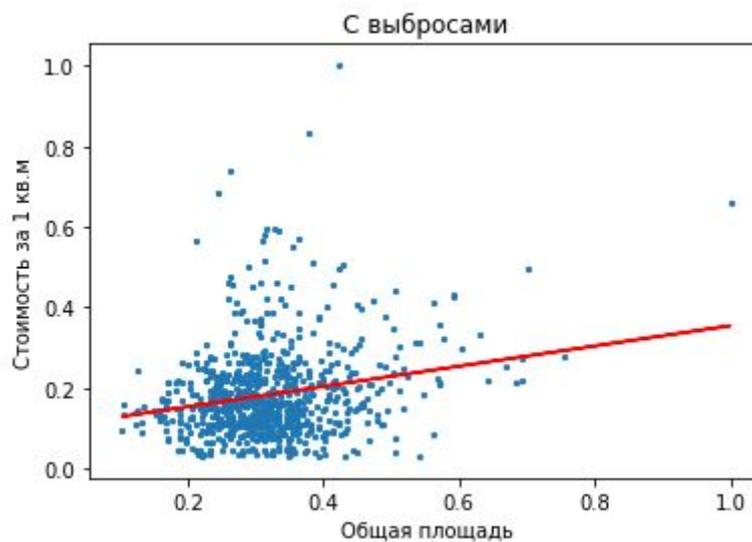


Рисунок 17. Параметр “Общая площадь” с выбросами

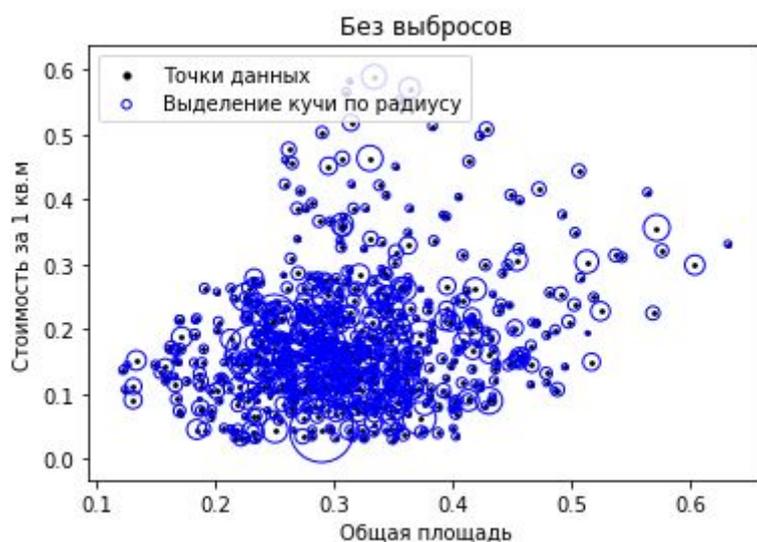


Рисунок 18. Параметр “Общая площадь” без выбросов

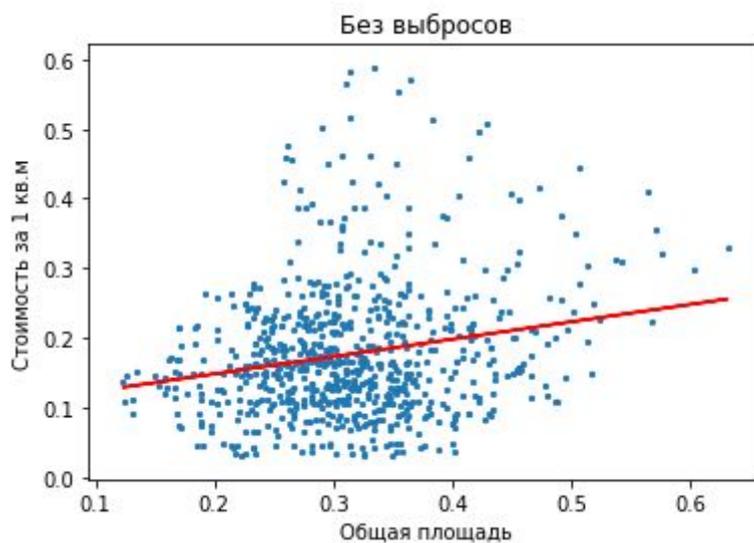


Рисунок 19. Параметр “Общая площадь” без выбросов

Параметр “Количество собственников”

Количество собственников объекта недвижимости.

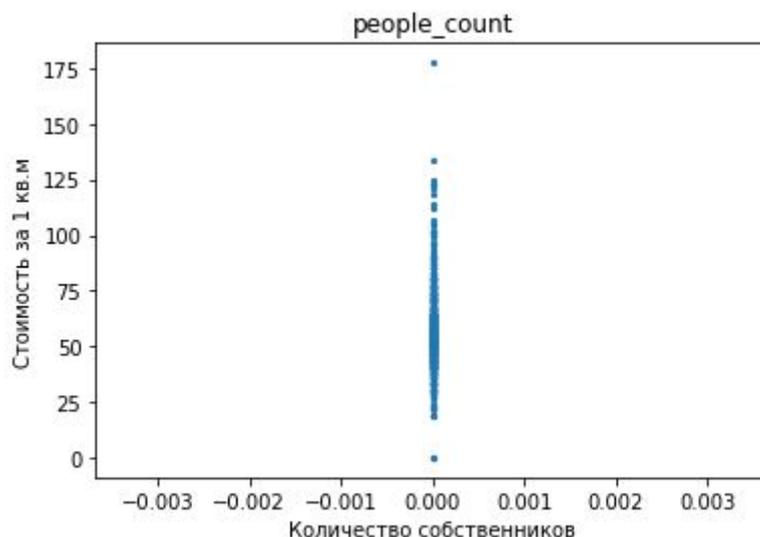


Рисунок 20. Параметр “Количество собственников”

На ось абсцисс взяты значения параметра “Количество собственников”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

Данный параметр необходимо исключить из модели прогнозирования в виду отсутствия значимости данного параметра, потому что на графике видно, что взаимосвязь между параметром “Количество собственников” и целевой функцией отсутствует.

Параметры, которые подобны рассмотренным ранее параметрам “Год постройки”, “Общая площадь”, “Этаж”, будут анализироваться методом кластеризации для упорядочивания данных в сравнительно однородные группы.

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны. Главное отличие кластеризации

от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Применение кластерного анализа в общем виде сводится к следующим этапам:

1. Отбор выборки объектов для кластеризации.
2. Определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных.
3. Вычисление значений меры сходства между объектами.
4. Применение метода кластерного анализа для создания групп сходных объектов (кластеров).
5. Представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Параметр “Количество фотографий”

Количественное значение качественных фотографий, имеющихся в описании и максимально полно визуально представляющих объект недвижимости. Предполагается, что на одно помещение в идеале должно быть 3 фотографии минимум.

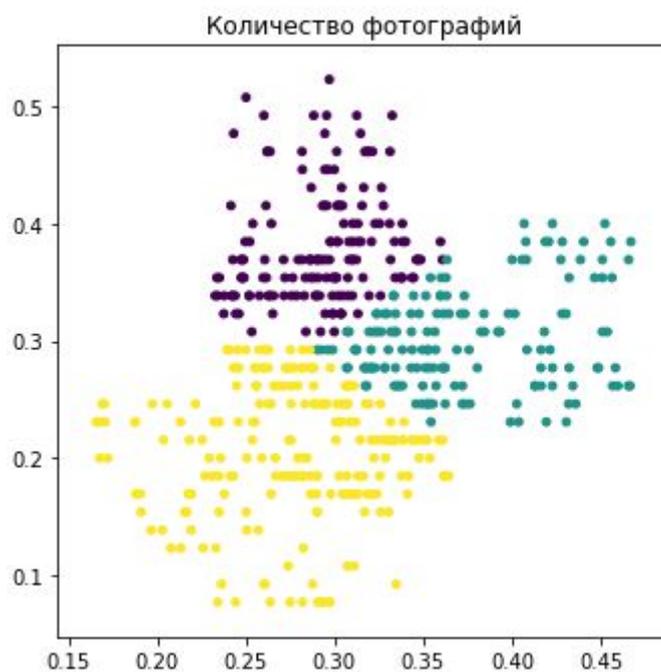


Рисунок 21. Параметр “Количество фотографий”

На ось абсцисс взяты значения параметра “Количество фотографий”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

Параметр “Улица”

Название элемента городской инфраструктуры, к которому прикреплен объект недвижимости согласно кадастровым документам.

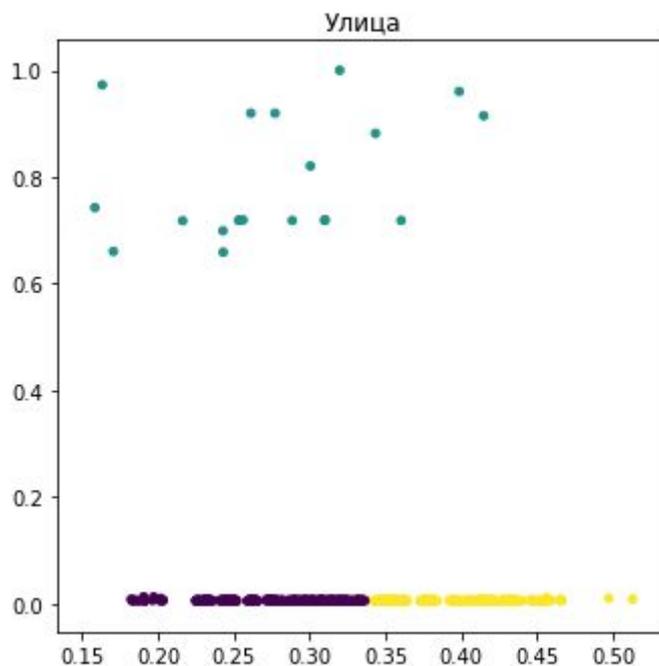


Рисунок 22. Параметр “Улица”

На ось абсцисс взяты значения параметра “Улица”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

Параметр “Район”

Район - городская территориальная единица, в которой находится объект недвижимости согласно официальным географическим данным.

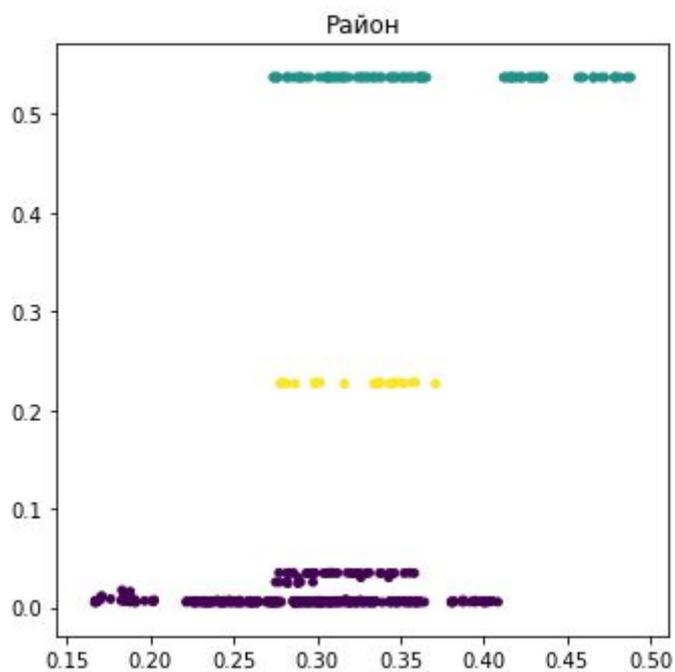


Рисунок 23. Параметр “Район”

На ось абсцисс взяты значения параметра “Район”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

Параметр “Тип дома”

Жилые здания, построенные по объединенной в серию группе типовых проектов.

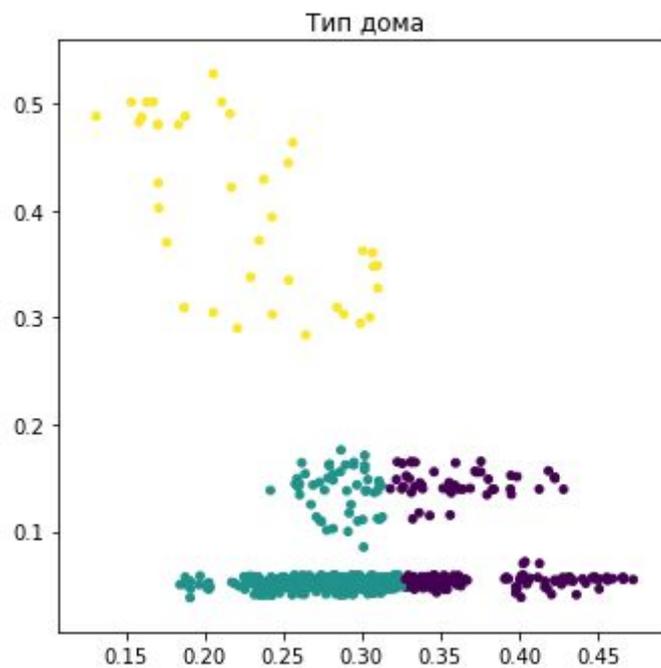


Рисунок 24. Параметр “Тип дома”

На ось абсцисс взяты значения параметра “Тип дома”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

Параметр “Этаж”

Уровень здания над (или под) уровнем земли, на котором находится объект недвижимости.

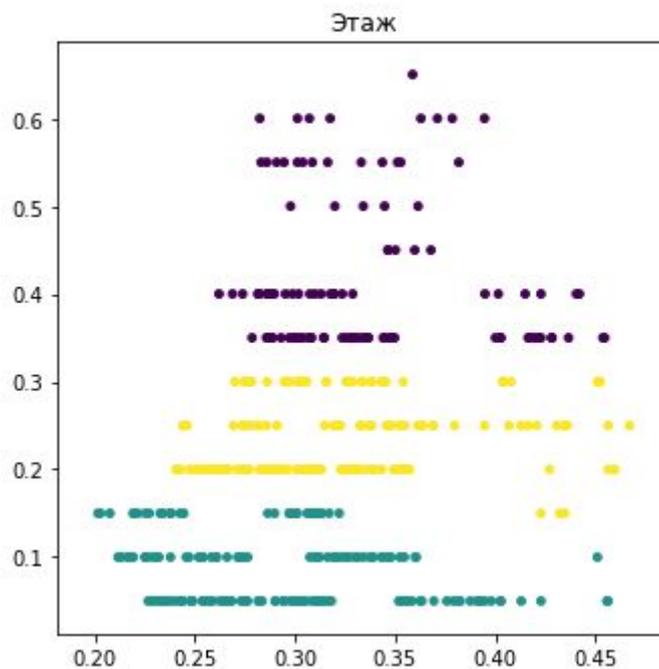


Рисунок 25. Параметр “Этаж”

На ось абсцисс взяты значения параметра “Этаж”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

Параметр “Этажность”

Количество надземных этажей, в том числе технического этажа, мансардного, а также цокольного этажа, если верх его перекрытия находится выше средней планировочной отметки земли не менее чем на 2 м.

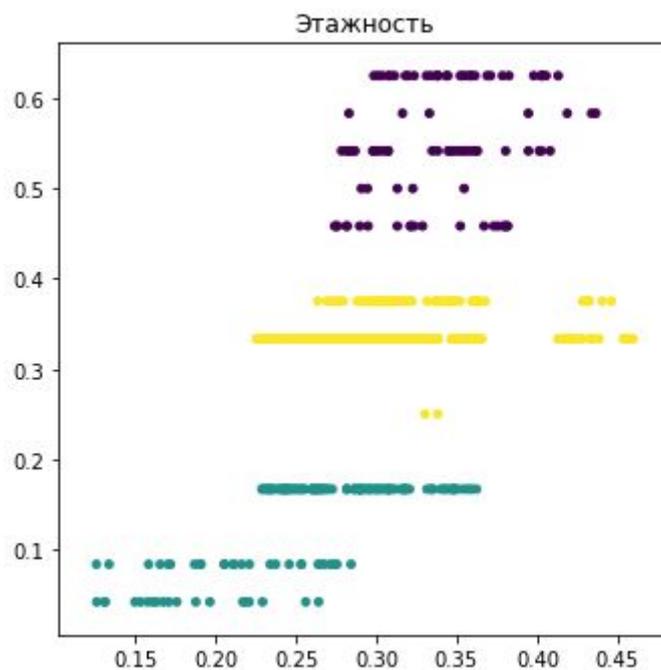


Рисунок 26. Параметр “Этажность”

На ось абсцисс взяты значения параметра “Этажность”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

Параметр “Жилая площадь”

Сумма площадей жилых комнат, не включающая в себя площадь подсобных помещений (кухонь, коридоров, ванных, санузлов, встроенных шкафов, кладовых, а также площадь, занятая внутриквартирной лестницей).

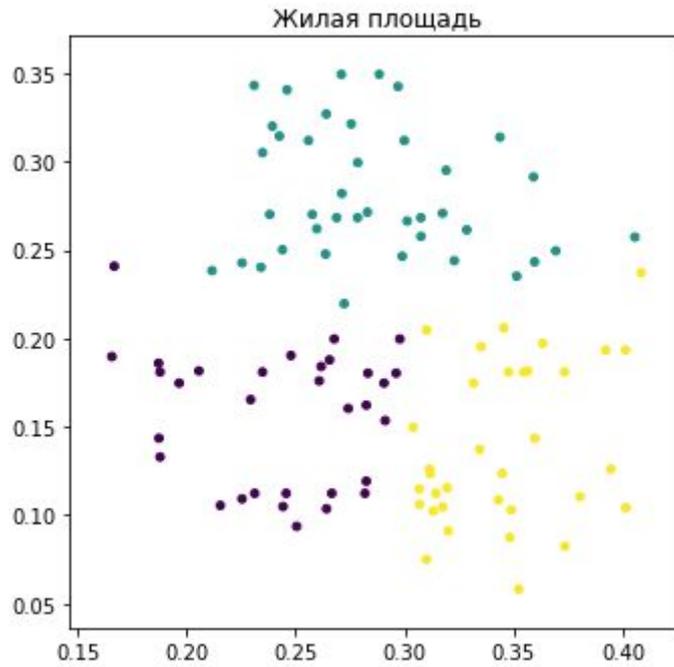


Рисунок 27. Параметр “Жилая площадь”

На ось абсцисс взяты значения параметра “Жилая площадь”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

Параметр “Площадь кухни”

Площадь кухонного помещения.

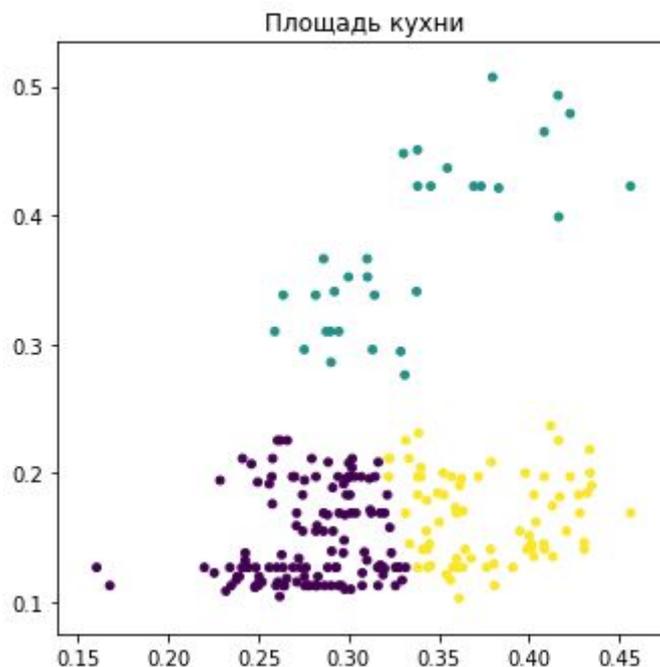


Рисунок 28. Параметр “Площадь кухни”

На ось абсцисс взяты значения параметра “Площадь кухни”.

На ось ординат взяты значения целевой функции “Цена за 1 кв.м.”.

Модель для прогнозирования - RandomForestRegressor. Осуществлялось случайное деление на тестовую и обучающую выборку в процентном соотношении 40:60 соответственно при помощи метода `train_test_split()`. В результате вычисления точности для 61 признака оценка составила - 0.9557631711896246 в процентном соотношении 95,5% при дельта равное 0. Количество, правильно определенных объектов, составило 307.

Модель для прогнозирования - KNeighborsRegressor. Осуществлялось случайное деление на тестовую и обучающую выборку в процентном соотношении 40:60 соответственно при помощи метода `train_test_split()`. В результате вычисления точности для 61 признака оценка составила - 0.4746412552949123 в процентном соотношении 47,4%

при дельта равное 0. Количество, правильно определенных объектов, составило 153.

Модель для прогнозирования - GradientBoostingRegressor. Осуществлялось случайное деление на тестовую и обучающую выборку в процентном соотношении 40:60 соответственно при помощи метода `train_test_split()`. В результате вычисления точности для 61 признака оценка составила - 0.5096146117169429 в процентном соотношении 50,9% при дельта равное 0. Количество, правильно определенных объектов, составило 164.

3.4 Выявление значимых параметров

Рассчитаем оценки точности моделей, используя все параметры после нормализации в количестве 61 параметров.

Оценка точности моделей равна:

Случайный лес - 0.9557631711896246,

Ближайших соседей - 0.4746412552949123,

Градиентный бустинг - 0.5096146117169429.

Рассчитаем точность моделей без таких параметров как: жилая площадь, площадь кухни, площадь балкона. Данные параметры были объединены в группу так как эти параметры содержат информацию о площади.

Оценка точности моделей равна:

Случайный лес - 0.9268580905442678,

Ближайших соседей - 0.47111289503248466,

Градиентный бустинг - 0.5095218251374142.

Так как точность моделей уменьшается, при удалении данных параметров, параметры являются значимыми.

Рассчитаем точность моделей без таких параметров как: количество квартир на лестничной площадке, описания расположение дома объекта недвижимости. Данные параметры были объединены в группу так как эти параметры содержат информацию об этажности.

Оценка точности моделей равна:

Случайный лес - 0.9970357247419017,

Ближайших соседей - 0.47111289503248466,

Градиентный бустинг - 0.5095040342769709.

Так как точность моделей увеличивается, при удалении данных параметров, параметры являются не значимыми.

Рассчитаем точность моделей без таких параметров как: требуется ли фотографирование, медиафайл для отображения на сайте, количество планировок. Данные параметры были объединены в группу так как эти параметры содержат информацию об визуализации объекта.

Оценка точности моделей равна:

Случайный лес - 0.9679623418467267,

Ближайших соседей - 0.47111289503248466,

Градиентный бустинг - 0.5095691575831118.

Так как точность моделей увеличивается, при удалении данных параметров, параметры являются не значимыми.

Рассчитаем точность моделей без таких параметров как: отображение на карте, в черте города. Данные параметры были объединены в группу так как эти параметры содержат информацию об расположении объекта.

Оценка точности моделей равна:

Случайный лес - 0.9840322185761271,

Ближайших соседей - 0.47111289503248466,

Градиентный бустинг - 0.5095594870303479.

Так как точность моделей увеличивается, при удалении данных параметров, параметры являются не значимыми.

Рассчитаем точность моделей без таких параметров как: тип стен, тип дома, год постройки, планировка студии, количество уровней объекта недвижимости, многоуровневый ли объект, износ дома, застройщик дома. Данные параметры были объединены в группу так как эти параметры содержат информацию об индивидуальных характеристиках дома.

Оценка точности моделей равна:

Случайный лес - 0.8602975000373134,

Ближайших соседей - 0.47257493735752715,

Градиентный бустинг - 0.5094353193015106.

Так как точность моделей уменьшается, при удалении данных параметров, параметры являются значимыми.

Рассчитаем точность моделей без таких параметров как: тип комфорта объекта недвижимости, количество проданных комнат в объекте, количество спален, общее количество комнат. Данные параметры были объединены в группу так как эти параметры содержат информацию о количестве комнат в объекте недвижимости.

Оценка точности моделей равна:

Случайный лес - 0.9830355172738731,

Ближайших соседей - 0.47111289503248466,

Градиентный бустинг - 0.5096038723554324.

Так как точность моделей увеличивается, при удалении данных параметров, параметры являются не значимыми.

Рассчитаем точность моделей без таких параметров как: сумма услуг по договору, цена продажи объекте недвижимости, последний

редактор стоимости объекта недвижимости, ликвидность цены, тип предложения, предыдущее значение стоимости, цена за кв.м. проданного объекта. Данные параметры были объединены в группу так как эти параметры содержат информацию о изменении стоимости объекта недвижимости.

Оценка точности моделей равна:

Случайный лес - 0.9967177596817098,

Ближайших соседей - 0.46286153599265817,

Градиентный бустинг - 0.7453973086466483.

Так как точность моделей увеличивается, при удалении данных параметров, параметры являются не значимыми.

Рассчитаем точность моделей без таких параметров как: тип плиты (электрическая или газовая), тип санузла, тип недвижимости (квартира/общежитие/комната), количество парковочных мест. Данные параметры были объединены в группу так как эти параметры содержат информацию о характеристиках самого объекта недвижимости.

Оценка точности моделей равна:

Случайный лес - 0.9276657917011472,

Ближайших соседей - 0.47111289503248466,

Градиентный бустинг - 0.5092868458668924.

Так как точность моделей уменьшается, при удалении данных параметров, параметры являются значимыми.

Рассчитаем точность моделей без таких параметров как: программа “Молодая семья”; субсидии; причина продажи; причина изменения статуса объекта в риэлторской информационной системе; банк, наложивший обременение; срочная продажа; наличие встречного варианта; вероятность торга; тип договора, заключенного между

риэлтором и клиентом; требуется ли быстрая продажа; находится ли объект в ипотеке. Данные параметры были объединены в группу так как эти параметры содержат информацию о характеристиках сделки с объектами недвижимости.

Оценка точности моделей равна:

Случайный лес - 0.9967053275089517,

Ближайших соседей - 0.471260687166843,

Градиентный бустинг - 0.5095987645359934.

Так как точность моделей увеличивается, при удалении данных параметров, параметры являются не значимыми.

Рассчитаем точность моделей без таких параметров как: сдается ли объект недвижимости квартирантам, более 3 лет в собственности, количество комментариев, наличие почтового адреса клиента, статус прозвона объекта недвижимости, код сотрудника, осуществляющий прозвон объекта. Данные параметры были объединены в группу так как эти параметры содержат информацию для риэлторов.

Оценка точности моделей равна:

Случайный лес - 0.9970544782074844,

Ближайших соседей - 0.5061086491444773,

Градиентный бустинг - 0.5096398392889846.

Так как точность моделей увеличивается, при удалении данных параметров, параметры являются не значимыми.

Опытным путем был определен список параметров, которые являются значимыми при прогнозировании 1 кв.м. Также значимыми являются параметры, которые были определены ранее: район, улица, количество фотографий, дата создания, дата редактирования, этаж, этажность, количество комнат, площадь, количество ванных комнат, окно,

ремонт, серия дома, балкон, высота потолков. Рассчитаем точность моделей для значимых параметров: район, улица, количество фотографий, дата создания, дата редактирования, этаж, этажность, количество комнат, площадь, количество ванных комнат, окно, ремонт, серия дома, балкон, высота потолков, жилая площадь, площадь кухни, площадь балкона, тип стен, тип дома, год постройки, планировка студии, количество уровней объекта недвижимости, многоуровневый ли объект, износ дома, застройщик дома, тип плиты (электрическая или газовая), тип санузла, тип недвижимости (квартира/общежитие/комната), количество парковочных мест.

Для точных оценок модели необходимо рассчитать абсолютную (2) и относительную (3) погрешность.

$$\Delta X = |X_{meas} - X_{true}| \quad (2),$$

где X_{meas} - полученные значения при прогнозировании стоимости 1 кв.м,

X_{true} - значения стоимости 1 кв.м. из исходного набора данных.

$$\delta_x = \frac{\Delta X}{X_{true}} * 100\% \quad (3),$$

где ΔX - абсолютная погрешность,

X_{true} - значения стоимости 1 кв.м. из исходного набора данных.

Оценка точности моделей равна:

Случайный лес - 0.9798353592900931 в процентном соотношении 97,7%. Количество, правильно определенной стоимости 1 кв.м. для объектов недвижимости, составило 314. Абсолютная погрешность для случайного леса - 0,9. Относительная погрешность для случайного леса - 5,9%.

Ближайших соседей - 0.2396388518571313 в процентном соотношении 23,9%. Количество, правильно определенной стоимости 1

кв.м. для объектов недвижимости, составило 77. Абсолютная погрешность для случайного леса - 0,6. Относительная погрешность для случайного леса - 13,9%.

Градиентный бустинг - 0.8625631913074149 в процентном соотношении 86,2%. Количество, правильно определенной стоимости 1 кв.м. для объектов недвижимости, составило 277. Абсолютная погрешность для случайного леса - 1,3. Относительная погрешность для случайного леса - 5%.

Вывод: Выборка данных по объектам представляла довольно большой объем параметров. Были внесены изменения в структуру набора данных. На этапе предобработки были удалены параметры, которые не информативны. Так как алгоритмы машинного обучения чувствительны к масштабированию данных, все признаки необходимо было нормализовать. После обработки параметров, опытным путем была выявлена лучшая модель прогнозирования: модель случайного леса. Оценка точности модели равна - 0.9798353592900931 в процентном соотношении 97,7%. Количество, правильно определенных объектов, составило 314. Абсолютная погрешность для случайного леса - 0,9. Относительная погрешность для случайного леса - 5,9%.

Например, объект 222509 - это трехкомнатная квартира, площадью 112 квадратных метров в центре города, находящаяся на улице Минская на 5 этаже девятиэтажного дома. Стоимость данной квартиры из исходного набора данных составляет 64.285 тысяч рублей за 1 кв.м. Модель случайного леса предсказала стоимость для данной квартиры равную 64.257 тысяч рублей за 1 кв.м.

Заключение

Рассматривая квартиру как товар при операциях купли/продажи в рыночной экономике, необходимо учитывать конкретные факторы, а именно: район, улица, количество фотографий, дата создания, дата редактирования, этаж, этажность, количество комнат, площадь, количество ванных комнат, окно, ремонт, серия дома, балкон, высота потолков. Исходя из имеющихся факторов определяется рыночная цена и рыночная стоимость квартиры. Существует несколько способов для оценки стоимости квартир. Использование нескольких способов или комбинируя их, можно получить наиболее правильную рыночную стоимость.

Работа осуществлялась с отражением объекта недвижимости в информационном пространстве. Исходный набор данных содержит 147 характеристик. Характеристики объектов представлены с различными типами данных (приложение 1).

В современных условиях рынка существует множество различных методов и приемов как для оценки реальной рыночной стоимости недвижимости, так и для прогнозирования изменения цены на недвижимость. Невозможно выделить или унифицировать какой-то один метод оценки рыночной стоимости недвижимости, по причине того, что у каждого способа существуют свои особенности. Таким образом, в данной работе будут рассмотрены и применены методы машинного обучения для выявления значимых параметров для прогнозирования цены на недвижимость такие как: метод случайного леса, метод ближайших соседей, метод градиентного бустинга.

Выборка данных по объектам представляла довольно большой объем параметров. Были внесены изменения в структуру набора данных.

На этапе предобработки были удалены параметры, которые не информативны. Так как алгоритмы машинного обучения чувствительны к масштабированию данных, все признаки необходимо было нормализовать. После обработки параметров, опытным путем была выявлена лучшая модель прогнозирования: модель случайного леса. Оценка точности модели равна - 0.9798353592900931 в процентном соотношении 97,7%. Количество, правильно определенных объектов, составило 314. Абсолютная погрешность для случайного леса - 0,9. Относительная погрешность для случайного леса - 5,9%.

Например, объект 222509 - это трехкомнатная квартира, площадью 112 квадратных метров в центре города, находящаяся на улице Минская на 5 этаже девятиэтажного дома. Стоимость данной квартиры из исходного набора данных составляет 64.285 тысяч рублей за 1 кв.м. Модель случайного леса предсказала стоимость для данной квартиры равную 64.257 тысяч рублей за 1 кв.м.

Список литературы.

1. Асаул А.Н., Карасев А.В. Экономика недвижимости. М. ; СПб., 2004
2. Бирман Э.Г. Сравнительный анализ методов прогнозирования //НТИ. Сер.2 – 1986. – №1. – С. 11–16.
3. Бодянский Е.В., Кучеренко Е.И. Диагностика и прогнозирование временных рядов многослойной радиально-базисной нейронной сети //Труды VIII Всероссийской конференции «Нейрокомпьютеры и их применение»: Сб. докл., 2002. – С. 69–72.
4. Власов А.Д., Понько В.А., Хизаметдинов С.В. Прогноз рыночной цены квартир многоэтажной жилой застройки методом геокосмических аналогий. Сибирский научный центр «Экопрогноз». Новосибирск, 2007.
5. Гэри У. Элдред. Недвижимость 101.Баланс Бизнес Букс, 2008.
6. Евстафьев А.И., Гордиенко В.А. Прогнозирование индикаторов рынка недвижимости путем двумерного разложения дискретной пространственно-параметрической модели на основе применения нейросетей // Материалы XI Национального конгресса по недвижимости. М., 2007
7. Комаров С.И. Прогнозирование стоимости земельных участков для индивидуального жилищного строительства (на примере Московской области). – Автореф. дисс.канд.эконом.наук. М., 2007.
8. Оценка недвижимости: Учебник / под ред. А.Г. Грязновой, М.А. Федотовой. – М., “Финансы и статистика”, 2002.
9. Оценка рыночной стоимости недвижимости, под общей редакцией Зарубина В.Н и Рутгайзера В.М., М.: Дело, 1998.
- 10.Печенкина А.В. Применение метода регрессионного анализа при прогнозировании цен на рынке жилья // Современный финансовый

рынок РФ: материалы междунар. науч.-практ. конф. (25 апреля 2008 г., г. Пермь) / Перм. гос. ун-т. Пермь, 2008. 382 с.

11. Рубинштейн Е.Д. Анализ рынка недвижимости и его прогнозирование. Теория и практика общественного развития. 2015 №12
12. Снитюк В.Е. Применение метода стохастической релаксации для прогнозирования рынка недвижимости // Материалы Международной научной конференции «Нейросетевые технологии и их применение. Краматорск, 2003.
13. Стерник Г.М., Стерник С.Г., Свиридов А.В. Методология прогнозирования российского рынка недвижимости Часть 3. Эволюция методов прогнозирования рынка жилья // Механизация строительства. — 2014. — № 2. — с. 60-64.
14. Стерник Г.М., Стерник С.Г. Анализ рынка недвижимости для профессионалов. М.: Экономика, 2009. 606 с.
15. Сутягин В.Ю., Черкашнев Р.Ю. Прогнозирование цен на рынке жилой недвижимости на Тамбовском региональном рынке. Социально-экономические явления и процессы, 2012.
16. Хаефельдер М., Лозовская Е., Хануш Е. Фундаментальный и технический анализ рынка ценных бумаг. СПб.: Питер, 2005. 352 с.
17. Виноградов Д.В. Факторы, влияющие на стоимость объектов недвижимости. Оценка недвижимости в соответствии с международными стандартами. Экономика недвижимости: учебное пособие.
<http://subschet.ru/subschet.nsf/docs/88FBB50288389CE7C325723100651161.html>

18. Градиентный бустинг — просто о сложном
<https://neurohive.io/ru/osnovy-data-science/gradientyj-busting/>
19. Зачем нужен сервис онлайн-оценки недвижимости?
<http://www.estimatica.info/assessment/real-estate/126-servis-onlajn-otsenki-nedvizhimosti>
20. Метод ближайших соседей <http://www.machinelearning.ru>
21. Особенности функционирования региональных рынков жилой недвижимости. Значимые факторы, отражающие процессы ценообразования на рынках жилой недвижимости в регионах. Построение многофакторных регрессионных моделей стоимости жилья в некоторых областях РФ. <http://www.allbest.ru/>
22. Сегментация рынков недвижимости
<http://market-pages.ru/ocenkanedvij/8.html>
23. Случайный лес (Random Forest)
<https://dyakonov.org/2016/11/14/%D1%81%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D1%8B%D0%B9-%D0%BB%D0%B5%D1%81-random-forest/>
24. Стерник Г.М. Методические рекомендации по анализу рынка недвижимости. <https://studfiles.net/preview/3852679>
25. Прогнозирование на основе аппарата нейронных сетей.
<https://studport.ru/referaty/tehnika/6620-prognozirovanie-na-osnove-apparata-nejronnyh-setej>
26. Топ - 100 застройщиков в России.
<http://www.vestnikstroy.ru/ratings/building/#!/11858>

Приложение 1

Таблица 1 Характеристики объекта

Атрибут	Описание	Тип
Id	Номер объекта в риэлторской информационной системе	Числовой
Newhouses_id	Тип дома в риэлторской информационной системе	Числовой
Class	Класс недвижимости в риэлторской информационной системе. К классам относятся: квартиры, загородная, коммерческая, гаражи.	Текстовый
Action	Тип сделки(продажа/аренда)	Текстовый
Direction	Предложение	Текстовый
Status	Статус объекта недвижимости в риэлторской информационной системе	Текстовый
User_id	Код риэлтора ответственного за продажу объекта недвижимости	Текстовый
last_editor_id	Код последнего редактора информации об объекте недвижимости	Текстовый
City_id	Город, где находится объект недвижимости продажи	Текстовый

District_id	Район, где находится объект недвижимости	Текстовый
Street_id	Улица, где находится объект недвижимости	Текстовый
On_map	Отображение на карте объекта недвижимости	Бинарный
Photos_cnt	Количество фотографий загруженных для объекта недвижимости в риэлторской информационной системе	Числовой
comments_cnt	Количество комментариев, оставленных по объекту недвижимости риэлторской информационной системе	Текстовый
date_create	Дата создания объекта недвижимости в риэлторской информационной системе	Дата
date_update	Дата обновления объекта недвижимости в риэлторской информационной системе	Дата
date_rise	Дата поднятия объекта в списке объектов в риэлторской информационной системе	Дата
rise	Поднятие объекта в риэлторской информационной системе	Бинарный

mortgage	Находится ли объект недвижимости в ипотеке	Бинарный
mortgage_sum m	Сумма действующей ипотеки, которую необходимо заплатить перед продажей объекта недвижимости	Числовой
trade_in	Имеется ли у продавцов уже встречный вариант объекта недвижимости	Бинарный
contract_num	Номер договора, заключенный между продавцом и риэлтором	Числовой
contract_type	Тип договора, заключенный между продавцом и риэлтором	Текстовый
contract_cost	Сумма по договору, заключенному между продавцом и риэлтором	Числовой
contract_date	Дата заключения договора между продавцом и риэлтором	Дата
contract_raschet s	Дата расчета по договору заключенному между продавцом и риэлтором	Дата
express_contract	Требуется ли быстрая продажа объекта недвижимости	Бинарный
price	Цена объекта (тыс.рублей)	Числовой
auction	Вероятность торга	Бинарный

about_price	Информация о цене объекта недвижимости в риэлторской информационной системе	Текстовый
fee_type	Тип предложения, к которому относится объект недвижимости (квартира или комната или пансионат и т.д)	Текстовый
layouts_cnt	Количество планировок загруженных в объект недвижимости в риэлторской информационной системе	Числовой
call_status	Статус прозвона объекта недвижимости службой качества риэлторского агенства	Текстовый
callstat_agents_id	Код сотрудника, который осуществлял прозвон объекта недвижимости на актуальность продажи.	Числовой
suspected	Является ли объект подозрительным. Данное подозрение устанавливает служба качества, ответственная за прозвон объектов недвижимости.	Бинарный
price_changed	Причина изменения цены объекта недвижимости.	Текстовый

price_editor	Код редактора, который вносил изменения по цене объекта недвижимости.	Числовой
probable_price	Ликвидная цена объекта недвижимости, рассчитанная риэлторской информационной системой.	Числовой
hot	Требуется ли срочная продажа объекта недвижимости.	Бинарный
sold_date	Дата заключения сделки продажи по объекту недвижимости.	Дата
sold_price	Цена продажи, указанная в договоре купли-продажи.	Числовой
buyer	Код риэлтора покупателя, с кем осуществляется сделка купли-продажи.	Числовой
buyer_ticket	Код заявки покупателя в риэлторской информационной системе.	Числовой
suspend_date	Дата приостановления продажи объекта недвижимости в риэлторской информационной системе.	Дата
suspend_reason	Причина приостановления продажи объекта недвижимости в риэлторской информационной системе.	Текстовый

old_price	Предыдущее значение стоимости объекта недвижимости до изменения объекта недвижимости.	Числовой
own_note	Дополнительный комментарий для риэлтора, ответственного за продажу объекта недвижимости	Текстовый
visual	Какой медиафайл необходимо отображать на сайте в первую очередь	Текстовый
common_note	Дополнительный комментарий для коллег, скрытый от клиентов агентства недвижимости.	Текстовый
photoallowed	Требуется ли фотографирование объекта недвижимости.	Бинарный
timephotoallowed	Комментарий для фотографа (о месте, дате и времени для осуществления фотографирования объекта недвижимости)	Текстовый
photo_ticket_params	Параметры заявки на фотографирования объекта недвижимости (для отдела фотографирования)	Текстовый
status_mail	Наличие почтового адреса клиента в риэлторской информационной системе.	Бинарный

time_photo	Время затраченное на фотографирование объекта недвижимости фотографом.	Числовой
unifinished	Закрыта ли заявка на фотографировани фотографом в риэлторской информационной системе.	Бинарный
in_mortgage	Находиться ли объект недвижимости в ипотеке	Бинарный
in_mortgage_summ	Сумма ипотечных средств, которая будет указана в договоре купли-продажи.	Числовой
in_mortgage_bank_id	Банк, выдавший ипотеку на приобретение объекта недвижимости.	Текстовый
metro_station_id	Станция метро, которая ближе всего к объекту недвижимости.	Бинарный
time_to_metro	Время до метро (в минутах) пешком	Числовой
time_to_metro2	Время до метро (в минутах) на транспорте	Числовой
sold_without_us	Продажа осуществляется по доверенности, которая оформлена на ответственного риэлтора.	Бинарный

cadaster_number	Кадастровый номер объекта недвижимости.	Числовой
keys_in_office	Ключи от объекта недвижимости находятся в офисе агентства.	Бинарный
deposit_source	Задаток от компании	Числовой
set_banner	Наличие баннера для рекламы объекта недвижимости.	Бинарный
encumbrance	Имеется ли обременение на объекте недвижимости.	Бинарный
encumbrance_summ	Сумма обременения на объект недвижимости.	Числовой
encumbrance_bank	Банк, который наложил обременение на объект недвижимости.	Текстовый
tenants	Сдается ли объект недвижимости квартирантам.	Бинарный
tenants_info	Контактная информация квартирантов	Текстовый
price_ar	Стоимость 1 кв.м. объекта недвижимости	Числовой
agency_id	Агентство покупателя, участвующее в сделке купли-продажи.	Числовой
sold_price_ar	Цена продажи объекта недвижимости.	Числовой

sell_reason	Причина продажи объекта недвижимости	Текстовый
web_owner		Числовой
reason_change_status	Причина изменения статуса объекта недвижимости в риэлторской информационной системе	Текстовый
transferred_for_photo	Количество переданных фотографий	Числовой
banner_presence	Наличия баннера для рекламы объекта недвижимости	Числовой
flat_on_floor	Количество квартир на лестничной площадке	Числовой
floor_num	Этаж, на котором находится объект недвижимости	Числовой
floors_cnt	Этажность дома, в котором находится объект недвижимости	Числовой
rooms_cnt	Число комнат объекта недвижимости	Числовой
bedrooms_cnt	Количество спален объекта недвижимости	Числовой
building_year	Год постройки объекта недвижимости	Числовой
area_total	Общая площадь объекта недвижимости	Числовой

area_live	Жилая площадь объекта недвижимости	Числовой
area_kitchen	Площадь кухни объекта недвижимости	Числовой
area_balcony	Площадь балкона объекта недвижимости	Числовой
secondary	Тип объекта недвижимости: вторичное или доленое	Бинарный
shared_type	Тип новостройки (переуступка, доленое строительство)	Текстовый
date_out	Дата сдачи новостройки	Дата
builder_id	Застройщик новостройки	Числовой
type	Тип (квартира/общежитие/комната)	Текстовый
two_levels	Многоуровневый объект недвижимости	Бинарный
levels_count	Количество уровней объекта недвижимости	Числовой
usage_percent	Износ дома, в котором располагается объект недвижимости	Числовой
bathroom	Вид санузла объекта недвижимости	Бинарный
bathrooms_cnt	Количество санузлов объекта недвижимости	Числовой

plate	Плита (электрическая/ газовая) объекта недвижимости	Текстовый
windows	Материал окон объекта недвижимости	Текстовый
territory	Описание территории объекта недвижимости	Текстовый
keep	Состояние ремонта объекта недвижимости	Текстовый
has	Техника, которая остается в объекте недвижимости после совершения сделки купли-продажи	Текстовый
has2	Мебель, которая остается в объекте недвижимости после совершения сделки купли-продажи	Текстовый
has3	Свободный текст для описания того, что остается в объекте недвижимости после совершения сделки купи-продажи	Текстовый
komunal_cost	Стоимость коммунальных расходов (в среднем) по объекту недвижимости	Числовой
notes	Дополнительное описание объекта недвижимости	Текстовый

series_id	Серия дома, в котором находится объект недвижимости	Текстовый
wall_id	Тип стен объекта недвижимости	Текстовый
young_family	Согласие продажи объекта недвижимости по государственному сертификату молодая семья	Бинарный
subsidy	Согласие продажи объекта недвижимости по субсидиям	Бинарный
more_than_3	Находиться ли объект недвижимости в собственности более 3 лет	Бинарный
people_count	Количество собственников объекта недвижимости	Числовой
beds_count	Количество спальных мест в объекте недвижимости	Числовой
balcon	Тип балкона в объекте недвижимости	Текстовый
loggia	Наличие лоджии в объекте недвижимости	Бинарный
apartment_letter	Буквенное обозначение номера объекта недвижимости	Текстовый
ceiling_height	Высота потолков в объекте недвижимости	Числовой
total_rooms_cnt	Общее количество комнат в объекте недвижимости	Числовой

sold_rooms_cnt	Количество проданных комнат в объекте недвижимости	Числовой
closed_yard	Тип закрытого двора в объекте недвижимости	Текстовый
area_rooms	Площадь комнат объекта недвижимости	Числовой
studio_type	Планировка студии объекта недвижимости	Текстовый
parking_cnt	Количество парковок при продаже объекта недвижимости	Числовой
house_location	Описание расположения дома объекта недвижимости	Текстовый
rooms_comfort	Тип комфорта объекта недвижимости	Текстовый
layout_type	Тип планировки объекта недвижимости	Текстовый
in_city	В черте города находится объект недвижимости или нет	Бинарный

Приложение 2

Параметр	Перечень значений
Id	2066874 2080642 1009546

	...
Newhouses_id	84533 11606 8894 ...
Class	flats
Action	sale
Direction	supp,y
Status	active
User_id	2117 788 60226 ...
last_editor_id	345 788 60579 ...
City_id	23
District_id	12438 46310 135 ...
Street_id	297 538 329 ...

On_map	no yes
Photos_cnt	27 17 0 ...
comments_cnt	0 3 2 ...
date_create	16.03.18 12:52 25.03.18 13:43 10.07.12 11:23 ..
date_update	16.03.18 12:52 25.05.18 15:03 10.07.18 13:33 ...
date_rise	16.03.18 12:52 25.05.18 16:04 10.07.18 14:01 ...
rise	no yes
mortgage	yes no
mortgage_summ	\N

trade_in	no yes
contract_num	no 1921733 ...
contract_type	vozmezdnyy exclusive
contract_cost	0 \N
contract_date	2016-11-10 2011-10-17 ...
contract_raschets	\N
express_contract	no yes
price	3100 2060 1230 ...
auction	no yes
about_price	0
fee_type	ruble
layouts_cnt	0 1 2 ...

call_status	none unanswered
callstat_agents_id	\N
suspected	no
price_changed	no
price_editor	\N 0
probable_price	\N 0
hot	no yes
sold_date	\N
sold_price	\N 0
buyer	
buyer_ticket	0
suspend_date	\N 0 2018-02-19 ...
suspend_reason	ДО ВЫЯСНЕНИЯ пока не продается. не устраивает цена ...
old_price	\N 3500 2000

	...
own_note	квартиру пока не продают живет отец Опека - 1 ребёнок Нина Александровна по этому телефону. ...
visual	photo \N
common_note	Готовы обмен на квартиру в Москве. Готовность объекта 70% Квартира свободна ...
photoallowed	no yes
timephotoallowed	0 Да в любое время можно Жанна ВИКТОРОВНА ВСЕГДА ДОМА днём на работе ...
photo_ticket_params	a:2:{s:19:"photo_type_checkbox";s:4:"free"; s:20:"photo_call_date_time";s:17:"2018-02-0 7 12:00 "}; a:2:{s:19:"photo_type_checkbox";s:4:"free"; s:20:"photo_call_date_time";s:0:""}; a:2:{s:19:"photo_type_checkbox";s:4:"free"; s:20:"photo_call_date_time";s:0:""}; ...
status_mail	0 1 ...

time_photo	0 2011-12-30 2012-03-27 ...
unifinished	0
in_mortgage	no yes
in_mortgage_summ	0
in_mortgage_bank_id	0
metro_station_id	0
time_to_metro	0
time_to_metro2	0
sold_without_us	no yes
cadaster_number	72:23:0430005:2474 72:23:0218006:23301 0 ...
keys_in_office	no yes
deposit_source	0
set_banner	0 1
encumbrance	no yes

encumbrance_summ	0
encumbrance_bank	0
tenants	no yes
tenants_info	0
price_ar	43.1034 50 65.1163 ...
agency_id	0
sold_price_ar	0 52.8736 48.6486 ...
sell_reason	0 family need ...
web_owner	-1
reason_change_status	48 36 14 ...
transferred_for_photo	0
banner_presence	0

flat_on_floor	0 8 3 ...
floor_num	9 11 12 ...
floors_cnt	17 16 24 ...
rooms_cnt	2 1 3 ...
bedrooms_cnt	2 1 3 ...
building_year	2015 2016 2000 ...
area_total	78.8 37 48.5 ...

area_live	0 14.6 36.3 ...
area_kitchen	0 12 10.37 ...
area_balcony	3.5 2.75 4.9 ...
secondary	yes
shared_type	0
date_out	2015-10-01 2013-10-01 0 ...
builder_id	0 6 ...
type	flat apartmen ...
two_levels	no yes

levels_count	1 2
usage_percent	33 46 0 ...
bathroom	separate unite unite.bath ...
bathrooms_cnt	0 1 2 ...
plate	electric no_plate gas ...
windows	plastic wood plastic_wood ...
territory	child_place.closed.parking.elevator.cargo_elevator parking.yard_opened.elevator.cargo_elevator kolyasochnaya.parking.guest_parking.closed.elevator ...

keep	cosmetic good design ...
has	internet.cable_tv.domofon.iron_door phone.internet.cable_tv.domofon.iron_door water_counters.hot_counters.iron_door.consj erg ...
has2	nothing kitchen.shkaf.sofa kitchen ...
has3	0 спальный гарнитур по договоренности ...
komunal_cost	0 1500 8000 ...
notes	Продаётся 1 комнатная квартира. 35.8 кв. м. на 2 этаже 9 этажного кирпичного дома напротив Ленты в Южном микрорайоне. Дом расположен на второй линии. во дворе. Вас не будет беспокоить шум дороги. Двор оборудован детской площадкой и просторной парковкой. В шаговой доступности находятся торговые центры "Метро". "Лента". "Ашан". гимназия и

	детский сад. При покупке за наличные разумный торг. ...
series_id	133 46 44 ...
wall_id	48 11 12 ...
young_family	no yes
subsidy	no yes
more_than_3	no yes
people_count	0
beds_count	0
balcon	yes_balcon balcon loggia ...
loggia	0
apartment_letter	0 КОМН.8 ...

ceiling_height	3.3 3 2.5
total_rooms_cnt	5 3 1 ...
sold_rooms_cnt	2 0 1 ...
closed_yard	0
area_rooms	18.1 30.1 0 27 19 ...
studio_type	euro no studio ...
parking_cnt	0 1 2 ...
house_location	удачное расположение дома обычное расположение дома 0 ...

rooms_comfort	особо удобная для проживания квартира типовая квартира без особенностей 0 ...
layout_type	yes no
in_city	yes no