Министерство науки и высшего образования

российской федерации

ФЕДЕРАЛЬНОЕ государственное БЮДЖЕТНОЕ образовательное учреждение

высшего образования

«ивановский государственный химико-технологический университет»

**Кафедра информационных технологий и цифровой экономики**

**МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ**

**НА ТЕМУ:**

**«****ПРОГНОЗИРОВАНИЕ РЫНОЧНОЙ СТОИМОСТИ ОБЪЕКТОВ НЕДВИЖИМОСТИ ГОРОДА ИВАНОВО МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ»**

**Автор: \_\_\_\_\_\_\_** И.Е Азорин

**Руководитель: \_\_\_\_\_\_\_** к.х.н., доц. Э. Г. Галиаскаров

**Руководитель магистерской программы: \_\_\_\_\_\_\_** д.т.н., проф. С.П. Бобков

**г. Иваново, 2020 г.**

**АННОТАЦИЯ**

**к выпускной квалификационной работе**

|  |
| --- |
| **Азорин Илья Евгеньевич** |
| **ФИО бакалавра** |

|  |  |
| --- | --- |
| **Тема:** | **Прогнозирование рыночной стоимости объектов недвижимости** |
|  | **города Иваново методами машинного обучения** |

|  |  |
| --- | --- |
| **Актуальность разрабатываемой темы:** |  |
| Данная работа актуальна из-за того, что в настоящий момент не существует | |
| профессиональных инструментов по оценке недвижимости, а существующие в | |
| интернете аналоги не могут удовлетворить потребности оценщиков. | |

|  |  |
| --- | --- |
| **Цель выпускной квалификационной работы:** |  |
| Целью работы является поиск способа проведения оценки недвижимости средствами | |
| машинного обучения на малых объёмах данных. Выбор наиболее оптимального | |
| из методов машинного обучения для использования в оценочной деятельности. | |

|  |  |
| --- | --- |
| **Для достижения цели поставлены следующие задачи:** |  |
| Ознакомление с предметной областью. | |
| Выбор подхода и метода для оценки. | |
| Сбор данных. | |
| Подготовка данных. | |
| Выбор алгоритмов машинного обучения и их обучение. | |
| Тестирование алгоритмов машинного обучения и анализ результатов. | |

|  |  |
| --- | --- |
| **Объект работы:** | В качестве объекта исследования выступает процесс |
| оценки недвижимости оценщиком. | |
|  | |
|  | |

|  |  |
| --- | --- |
| **Предмет работы:** | В качестве предмета исследования выступает процесс |
| оценки жилой недвижимости оценщиком, при помощи сравнительного подхода и | |
| метода сравнения предложений. | |
|  | |

|  |  |
| --- | --- |
| **Структура работ:** | Работа состоит из введения и четырех глав. В введении |
| описывается актуальность темы и ее научная новизна. Глава 1 представляет собой | |
| описание подходов и методик в оценке недвижимости и обосновывает выбор одного из | |
| них. Глава 2 представляет собой обзор процесса сбора, анализа и подготовки данных | |
| перед их использованием в машинном обучении. Глава 3 описывает методы машинного | |
| обучения, подходящие для задачи регрессионного анализа. Глава 4 представляет собой | |
| практическое машинного обучения, а также обзор полученных результатов. | |

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

ИВАНОВСКИЙ ГОСУДАРСТВЕННЫЙ

ХИМИКО-ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ

**Факультет** Техники, управления и цифровой инфраструктуры

**Кафедра** Информационных технологий и цифровой экономики

**Уровень подготовки** Магистратура

**Направление** Информационные системы и технологии

**Магистерская программа** Информационные системы и технологии

УТВЕРЖДАЮ:

**Зав. кафедрой \_\_\_\_\_\_\_\_\_И.А. Астраханцева**

**«03» февраля 2020 г.**

**ЗАДАНИЕ**

**на выпускную квалификационную работу**

**студенту(ке): *Азорин Илья Евгеньевич***

**1.** **Тема**: ***Прогнозирование рыночной стоимости объектов недвижимости города Иваново методами машинного обучения*** утверждена приказом по университету **№ 1122ст-19 от 11.12.2019 г.**

**2. Исходные данные**: законодательные и подзаконные нормативные акты; энциклопедическая и справочная литература; статистические и аналитические материалы; монографии; данные профессиональных периодических изданий; Интернет-ресурсы; самостоятельно собранные первичные материалы.

**3. Содержание работы:**

*Введение*: актуальность темы исследования; цель и задачи работы; предмет и объект исследования; научная новизна исследований с аргументацией и значимость работы.

*Глава 1. Подходы и методы оценки недвижимости.* Анализ подходов и методик в оценке недвижимости, для определения наиболее подходящих для прогнозирования цены объектов недвижимости при помощи использования алгоритмов машинного обучения.

*Глава 2. Сбор, анализ и обработка данных.* Обзор способов получения данных из открытых источников для их оценки; анализ собранных данных, обработка собранных данных и приведение их к виду пригодному для работы алгоритмов машинного обучения.

*Глава 3. Алгоритмы машинного обучения, подходящие для задачи регрессионного анализа*. Обзор методов прогнозирования: метод множественной линейной регрессии, метод на основе деревьев решений, улучшенная версия метода на основе деревьев решений.

*Глава 4. Практическое применение методов машинного обучения*. Практическое применение машинного обучения в рамках текущей задачи; оценка качества работы каждого метода; сравнение результатов.

*Заключение:* основные выводы, к которым пришел автор в ходе исследования, а также предложения и рекомендации.

*Приложения* *к выпускной квалификационной работе*: первичные материалы; громоздкие статистические материалы (таблицы, графики, множественные диаграммы и пр.); исходные тексты программ; копии экранных форм пользовательского интерфейса; формы входной и выходной документации; материалы, иллюстрирующие или детализирующие основные проектные решения.

*Список использованной литературы*.

**4. Вопросы для специальной разработки.** Сравнительный анализ прогнозов, проведенных различными методами прогнозирования.

**5. Руководитель работы: *к.х.н., доц. Галиаскаров Э.Г.***

**6. Консультанты:**

|  |  |  |  |
| --- | --- | --- | --- |
| Раздел | Консультант | Подпись, дата | |
|  |  | Задание выдал | Задание принял |
|  |  |  |  |

**7. Дата выдачи задания: 03.02.2020** г.

**8. Дата предоставления законченной работы: 08.06.2020 г.**

Руководитель ***Галиаскаров Эдуард Геннадьевич*** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Ф.И.О., подпись)

Студент ***Азорин Илья Евгеньевич*** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Ф.И.О., подпись)

**Календарный план**

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Наименование этапов квалификационной работы | Срок выполнения этапов  работы (проекта) | Примечание |
| 1. | Поиск и исследование исходных данных, проработка практических материалов по теме квалификационной работы | 03.02.2020 - 01.03.2020 |  |
| 2. | Составление плана работы | 01.03.2020 - 09.03.2020 |  |
| 3. | Разработка и согласование с руководителем первой главы квалификационной работы | 10.03-2020 – 27.03.2020 |  |
| 4. | Разработка и согласование с руководителем второй главы квалификационной работы | 06.04.2020 – 26.04.2020 |  |
| 5. | Разработка и согласование с руководителем третьей главы квалификационной работы | 27.04.2020 – 17.05.2020 |  |
| 6. | Согласование с руководителем введения, выводов и предложений | 18.05.2020 – 07.06.2020 |  |
| **7.** | **Представление квалификационной работы на кафедру** | **08.06.2020** |  |
| 8. | Подготовка доклада и графического материала | 08.06.2020 – 15.06.2020  22.06.2020 – 28.06.2020 |  |
| 9. | Проведение предварительной защиты квалификационной работы | 08.06.2020 – 11.06.2020 |  |
| 10. | Защита выпускной квалификационной работы | 29.06.2020-04.07.2020 |  |

Студент \_\_\_\_\_\_\_\_\_\_\_\_ Руководитель работы \_\_\_\_\_\_\_\_\_\_\_\_

РЕФЕРАТ

Пояснительная записка 55 с., 13 рис., 2 таб., 4 частей, 18 источников.

Перечень ключевых слов: машинное обучение, алгоритмы, оценка, жилая недвижимость, данные, анализ.

Целью работы является изучение методов машинного обучения применимых к задаче оценки рыночной стоимости недвижимости в городе Иваново.

В данной работе рассматриваются различные модели машинного обучения, такие как линейная регрессия и алгоритмы на основе решающих деревьев, с целью выбора наиболее точной, которая сможет учитывать специфику данных.

В качестве исходных данных для работы использовались записи о продаже недвижимости, собранные парсером c сайта IvanovoDom.ru.

Введение данной работы описывает причины, из-за которых было начато изучение данной темы, ее цель и актуальность, также оно захватывает описание предметной области, в рамках которой производится данная работа.

В первой главе происходит обзор подходов и методик по оценке недвижимости, описываются их плюсы и минусы и также производится выбор наиболее подходящего подхода и методики для оценки недвижимости при помощи алгоритмов машинного обучения.

Во второй главе описываются источники получения данных, процесс их сбора, анализа и предварительной обработки.

В третьей главе описываются наиболее подходящие для нашей задачи алгоритмы машинного обучения.

Четвертая глава целиком посвящена практическому применению выбранных алгоритмов машинного обучения на имеющемся наборе данных и анализу полученных результатов.

Оглавление

[Определения 3](#_Toc43598931)

[Обозначения и сокращения 4](#_Toc43598932)

[Введение 5](#_Toc43598933)

[1 Подходы и методы оценки недвижимости 7](#_Toc43598934)

[1.1 Обзор подходов в оценке недвижимости 7](#_Toc43598935)

[1.2 Обзор сравнительного подхода 8](#_Toc43598936)

[1.3 Обзор доходного подхода 9](#_Toc43598937)

[1.4 Обзор затратного подхода 10](#_Toc43598938)

[1.5 Выбор подхода и обзор методик 11](#_Toc43598939)

[1.6 Выводы по выбору подхода и метода оценки недвижимости 14](#_Toc43598940)

[2 Сбор, анализ и обработка данных 15](#_Toc43598941)

[2.1 Анализ доступных способов сбора информации 15](#_Toc43598942)

[2.2 Обзор собираемой информации 17](#_Toc43598943)

[2.3 Анализ собираемой информации 18](#_Toc43598944)

[3 Выбор алгоритмов МО 25](#_Toc43598945)

[3.1 Определение задачи регрессии 25](#_Toc43598946)

[3.2 Обзор методов, решающих задачу регрессии 28](#_Toc43598947)

[3.2.1 Обзор линейной регрессии 28](#_Toc43598948)

[3.2.2 Обзор случайного леса 30](#_Toc43598949)

[3.2.3 Метод XGBoost 31](#_Toc43598950)

[4 Результаты машинного обучения 33](#_Toc43598951)

[Заключение 47](#_Toc43598952)

[Список используемых источников 48](#_Toc43598953)

[ПРИЛОЖЕНИЕ А. 50](#_Toc43598954)

[ПРИЛОЖЕНИЕ Б. 51](#_Toc43598955)

[ПРИЛОЖЕНИЕ В 57](#_Toc43598956)

Определения

Объект недвижимости – земельные участки, участки недр и все, что прочно связано с землей, то есть объекты, перемещение которых без несоразмерного ущерба их назначению невозможно, в том числе здания, сооружения, объекты незавершенного строительства.

Датасет – это обработанная и структурированная информация в табличном виде. Строки такой таблицы называются объектами, а столбцы – признаками.

Выборка – часть генеральной совокупности элементов (датасета), которая охватывается экспериментом.

Целевой признак – признак, который вычисляются на основе одного или нескольких независимых признаков.

Независимый признак – признак, содержащий в себе ту или иную информацию о рассматриаемом обьекте.

Корреляция – статистическая взаимосвязь двух или более случайных величин.

Обозначения и сокращения

MO – Машиное обучение

MSE – Mean Squared Error

RMSE – Root Mean Square Error

MAPE – Mean Absolute Percentage Error

MAE – Mean Absolute Error

LR – Линейная Hегрессия (Linear Regression)

RF – Random Forest

XGBoost – EXtreme Gradient Boosting

Введение

Рынок недвижимости является огромным сектором мировой экономики. Данный рынок может разделяться на несколько уровней: мировой, государственный и региональный. Каждый из них имеет свой размер и является полноценной экосистемой. Согласно [1], под рынком недвижимости понимается совокупность участников и сделок, совершаемых между ними на рынке.

Оценка недвижимости производится только специальными людьми, которые имеют лицензию на этот вид деятельности. Под оценкой недвижимости, в [2] понимают процесс определения стоимости объекта или отдельных прав в отношении оцениваемого объекта недвижимости. Оценка стоимости недвижимости включает: определение стоимости права собственности или иных прав, например, права аренды, права пользования и т. д. в отношении различных объектов недвижимости.

В настоящий момент все этапы и способы оценки регламентированы законодательно. Оценщик может выбирать один из доступных ему способов оценки и согласно технологии выбранного метода произвести оценку объекта недвижимости.

Первое, что нужно оценщику при оценке недвижимости согласно стандарту п. 10 ФСО-7 «Оценка недвижимости» — это анализ рынка жилой недвижимости. Для выполнения этой задачи оценщик должен проанализировать сегмент рынка, к которому относится оцениваемый объект. После чего найти максимальную и минимальную границы диапазона цен в сегменте, к которому принадлежит объект оценки. Данная задач достаточно трудоемка, требует большого количества времени и повторения множества однотипных операций, таких как сверка параметров жилой недвижимости. Для упрощения процесса поиска аналогов, их фильтрации и получения наиболее схожих объектов жилой недвижимости, в данной работе предложено использовать методы машинного обучения, основанные на обучении с учителем. Принцип работы данных методов схож по своей сути с основным подходом в оценке жилой недвижимости, но поскольку поиском аналогов и расчётом конечной стоимости будет заниматься программа, то время, затраченное оценщиком на однотипную работу сократится во много раз.

Целью данной работы является проверка возможности использования методов машинного обучения в оценке жилой недвижимости в городе Иваново.

Для достиэения данной цели требуется выполнить такие задачи как: сбор данных, их анализ и обработка, создание и тренеровка алгоритомов машинного обучения, тестирование и анализ полученных результатов.

В качестве объекта исследования выступает процесс оценки недвижимости оценщиком.

В качестве предмета исследования выступает процесс оценки жилой недвижимости оценщиком, при помощи сравнительного подхода и метода сравнения предложений.

В качестве новизны, которую предлагает данная работа, выступают анализ датасета по продажам жилой недвижимости в городе Иваново с выбором подходящих признаков, а так же исследование применимости методов машинного обучения на малых обьемах данных.

Данная тема имеет актуальность поскольку во время оценки, оценщик должен указывать в отчете факторы по которым производит оценку, но поскольку для разных объектов имеется свой ограниченный объём информации, то оценщик не может использовать одни и те же факторы постоянно. Из-за этого возникает проблема, поскольку на данный момент нет доступных оценщику инструментов, которые позволили бы ему провести оценку «автоматезированно». Доступные в открытом доступе сервисы, например сайты по продаже недвижимости, предлагают спрогнозировать стоимость недвижимости только по определенным категориям объектов и только на текущий момент времени, что не подходит оценщику в его работе.

1. Подходы и методы оценки недвижимости
   1. Обзор подходов в оценке недвижимости

Деятельность по оценке недвижимости регулируется законодательно. Основные моменты, касающиеся оценки объектов недвижимости описаны в федеральных стандартах [3] и [4]. В стандарте [3] дается определение объекту недвижимости, т.е. объекту для которого производится оценка. В нем перечисляются основные обязанности оценщика и последовательность его действий. Также в нем перечислены основные подходы, которыми может пользоваться оценщик во время процесса оценки.

В настоящий момент в данных стандартах выделяют три подхода к оценке недвижимости, а именно:

* Сравнительный подход;
* Доходный подход;
* Затратный подход.

Все три подхода позволяют произвести оценку недвижимости при помощи разных параметров и разной исходной информации. Для сравнительного подхода достаточно определить критерии оценки и найти аналогичные объекты, при помощи которых можно оценить исследуемый объект. Для доходного подхода производится оценка объекта недвижимости при помощи анализа ожидаемых доходов, которые владелец может получить с этого объекта. Затратный же подход включает в себя расчет стоимости недвижимости через стоимость ее покупки или воссоздания с нуля (с учетом всего её износа).

Как видно из выше сказанного все три подхода уникальны, соответственно выбор того или иного подхода производит оценщик на основе доступных ему данных.

Далее будут рассмотрены все три подхода и выбран наиболее подходящий для применения в машинном обучении.

* 1. Обзор сравнительного подхода

Согласно п. 22 ФСО-7 «Оценка недвижимости» сравнительный подход применяется для оценки недвижимости, когда можно подобрать достаточное для оценки количество объектов-аналогов с известными ценами сделок и (или) предложений. Сравнительный подход основывается на прямом сравнении оцениваемого объекта с другими объектами недвижимости, которые были проданы или включены в реестр на продажу. Рыночная стоимость недвижимости определяется ценой, которую заплатит типичный покупатель за аналогичный по качеству и полезности объект. При использовании сравнительного подхода необходимо выполнить следующие шаги:

1. Сбор данных, изучение рынка недвижимости, отбор аналогов из числа сделок купли – продажи и предложений на продажу (публичных оферт).

2. Проверка информации по каждому отобранному аналогу о цене продажи и запрашиваемой цене, оплате сделки, физических характеристиках, местонахождении и условиях продажи.

3. Анализ и сравнение каждого аналога с объектом оценки по времени продажи (выставлению оферты), местоположению, физическим характеристикам и условиям продажи.

4. Корректировка цен продаж или запрашиваемых цен по каждому аналогу в соответствии с имеющимися различиями между ними и объектом оценки.

5. Согласование скорректированных цен аналогов и вывод показателя оценки стоимости объекта.

Сравнительный подход в основном применяется при условии развитого рынка недвижимости. Поскольку чем больше развит рынок, тем больше недвижимости аналогичной оцениваемому объекту можно найти. В следствии этого данный подход не следует использовать для узкоспециализированной недвижимости, такой как водонапорные станции для многоквартирных домов. Поскольку предложений об их продаже мало или практически нет, то очень сложно найти аналоги, а следовательно и оценить такой объект при помощи сравнительного подхода невозможно.

* 1. Обзор доходного подхода

Доходный подход рекомендуется применять, когда существует достоверная информация, позволяющая прогнозировать будущие доходы, которые исследуемые нежилые здания способны приносить, а также известны связанные с объектом оценки расходы. Согласно п. 23 ФСО-7 «Оценка недвижимости», при применении доходного подхода эксперт учитывает следующие положения:

* доходный подход применяется для оценки недвижимости, генерирующей или способной генерировать потоки доходов;
* в рамках доходного подхода стоимость недвижимости может определяться методом прямой капитализации, методом дисконтирования денежных потоков или методом капитализации по расчетным моделям;
* метод прямой капитализации применяется для оценки объектов недвижимости, не требующих значительных капитальных вложений в их ремонт или реконструкцию, фактическое использование которых соответствует их наиболее эффективному использованию. Определение стоимости объектов недвижимости с использованием данного метода выполняется путем деления соответствующего рынку годового дохода от объекта на общую ставку капитализации, которая при этом определяется на основе анализа рыночных данных о соотношениях доходов и цен объектов недвижимости, аналогичных оцениваемому объекту;
* метод дисконтирования денежных потоков применяется для оценки недвижимости, генерирующей или способной генерировать потоки доходов с произвольной динамикой их изменения во времени путем дисконтирования их по ставке, соответствующей доходности инвестиций в аналогичную недвижимость;
* метод капитализации по расчетным моделям применяется для оценки недвижимости, генерирующей регулярные потоки доходов с ожидаемой динамикой их изменения. Капитализация таких доходов проводится по общей ставке капитализации, конструируемой на основе ставки дисконтирования, принимаемой в расчет модели возврата капитала, способов и условий финансирования, а также ожидаемых изменений доходов и стоимости недвижимости в будущем;
* структура (учет налогов, возврата капитала, темпов изменения доходов и стоимости актива) используемых ставок дисконтирования и (или) капитализации должна соответствовать структуре дисконтируемого (капитализируемого) дохода;
* для недвижимости, которую можно сдавать в аренду, в качестве источника доходов следует рассматривать арендные платежи;
* оценка недвижимости, предназначенной для ведения определенного вида бизнеса (например, гостиницы, рестораны, автозаправочные станции), может проводиться на основании информации об операционной деятельности этого бизнеса путем выделения из его стоимости составляющих, не относящихся к оцениваемой недвижимости.

Данный подход в основном используется для оценки жилой недвижимости, для которой сформирован большой и активный рынок, а параметры для сравнения давно определены и универсальны для большинства случаев оценки. Примером такой недвижимости могут послужить рестораны, складские помещения и подобная недвижимость.

* 1. Обзор затратного подхода

Согласно п. 24 ФСО-7 «Оценка недвижимости» затратный подход рекомендуется применять для оценки объектов недвижимости - земельных участков, застроенных объектами капитального строительства, или объектов капитального строительства.

Затратный подход применяется, когда существует возможность заменить исследуемое жилое помещение другим объектом, который либо является точной копией объекта оценки, либо имеет аналогичные полезные свойства. Если объекту оценки свойственно уменьшение стоимости в связи с физическим состоянием, функциональным или экономическим устареванием, при применении затратного подхода необходимо учитывать износ и все виды устареваний. При расчете затратным необходимым условием является достаточно детальная оценка затрат на строительство идентичного (аналогичного) объекта с последующим учетом износа оцениваемого объекта. Основные этапы применения затратного подхода при оценке недвижимости:

* оценка затрат на воспроизводство (замещение) оцениваемого объекта;
* оценка величины предпринимательской прибыли (прибыли инвестора);
* расчет выявленных видов износа;
* оценка рыночной стоимости земельного участка;
* расчет итоговой стоимости объекта оценки путем корректировки затрат на воспроизводство (замещение) на износ с последующим увеличением полученной величины на стоимость земельного участка.

Затратный подход применяют в основном когда невозможно применить сравнительный или доходный подходы. В основном он применяется при технико-экономическом обосновании нового строительства, при оценке рыночной стоимости объектов незавершенного строительства и реконструируемых объектов. В ряде случаев затратный подход позволяет оценить эффективность инвестиционного проекта.

* 1. Выбор подхода и обзор методик

Из вышеперечисленных подходов больше всего приемлем сравнительный поход из-за того, что имеющиеся общедоступные средства массовой информации, содержащие объявления с предложениями на продажу объектов жилой недвижимости, позволяют провести такую оценку. Так же в самом стандарте [4], говорится что во время экспертизы при использовании сравнительного подхода, оценщик решает задачу регрессионного анализа, которая является одной из основных задач в машинном обучении. В сравнительном подходе, как и в машинном обучении уделяется большое внимание данным. Используемая в экспертизе информация об объектах-аналогах должна являться:

* достаточной – поскольку не противоречит данным анализа рынка и не ведет к существенному изменению характеристик и итоговой величины стоимости;
* достоверной – поскольку соответствует действительности (что установлено экспертом в результате телефонных переговоров с сотрудниками организаций, предлагающих на продажу объекты) и позволяет пользователю экспертизы об оценке делать правильные выводы о характеристиках, исследовавшихся экспертом при проведении оценки и определении итоговой величины стоимости исследуемого жилого помещения, и принимать базирующиеся на этих выводах обоснованные решения.

В рамках сравнительного подхода к оценке жилого помещения могут быть использованы следующие методы:

* Метод валового рентного множителя;
* Метод сравнения продаж/предложений.

В методе валового рентного множителя используется отношение стоимости имущества к валовому доходу, приносимому за счет использования этого имущества. Валовый рентный множитель интегрально отражает соотношение стоимости и доходности для каждого конкретного вида использования имущества. Данный метод применим, в частности, к недвижимому имуществу с различным назначением.

Недостаток метода использования валового рентного множителя заключается в том, что он не учитывает операционные расходы, то есть расходы на обеспечение функционирования объекта, возможные потери в сборе поступающих платежей (арендной платы) вследствие различных причин, а также различие в рисках вложения капитала в сопоставимые объекты.

Преимущества метода заключаются в следующем: использование этого метода позволяет оценщику достаточно оперативно получить приближенную оценку стоимости объекта имущества, а также спрогнозировать среднюю доходность имущества того или иного вида; значение валового рентного множителя уже включает в себя различия сопоставимых объектов (для имущества потребительская ценность которого меньше, меньше также и его рыночная стоимость и поступающие платежи), поэтому его величину не нужно корректировать в зависимости от различий оцениваемого и сопоставимых объектов, например жилой недвижимости.

Метод сравнения продаж/предложений является основным методом в рамках сравнительного подхода. Метод основан на прямом сравнении оцениваемого объекта с другими подобными ему объектами (объектами-аналогами), которые были недавно проданы/выставлены на продажу.

Преимущества метода таковы: он является наиболее простым методом; статистически обоснованным; допускается применение методов корректировки величин стоимости объектов-аналогов; обеспечивает получение данных для применения в других подходах к оценке имущества.

Недостатки метода: требует активного развитого сегмента рынка имущества данного вида (сегмента, к которому относится объект оценки); сравнительные данные не всегда имеются; требует внесения поправок (корректировок), большое количество которых может оказывать влияние на достоверность результатов. Например при сравнении двух почти одинаковых объектов жилой недвижимости (квартира), оба объекта будут иметь похожую площадь, этажность и находиться в соседних домах, но здания в которых они находятся построены из разного материала, что требует внесения корректировок в оценку стоимости материалов стен и перекрытий. Большое количество таких несоответствий при сравнении усложняет процесс оценки и уменьшает достоверность её итогового результата.

* 1. Выводы по выбору подхода и метода оценки недвижимости

В контексте текущей работы, когда рассматривается использование методов машинного обучения для оценки недвижимости, должен быть выбран метод который работает с наиболее простыми и доступными данными. Сама суть метода не должна быть слишком сложной, а алгоритм его работы мог бы быть смоделирован методами машинного обучения.

Как было сказано в предыдущем пункте, наилучшим подходом для решения данной задачи является сравнительный (рыночный). Суть этого подхода очень близка к задаче регрессии из машинного обучения, а алгоритм его работы возможно смоделировать при помощи методов машинного обучения из категории обучения с учителем, где модель сначала обучается на доступных данных и потом делает прогноз для объекта с заданными параметрами.

В рамках сравнительного подхода наиболее близкой методикой расчета стоимости недвижимости является именно метод сравнения продаж/предложений, поскольку именно этот метод имеет наибольшее сходства с алгоритмами основанными на обучении с учителем.

В расчете стоимости недвижимости «вручную» или при помощи машинного обучения в рамках сравнительного подхода большое значение будут иметь данные на основе которых эти расчеты будут производиться, именно сбору, анализу и обработке данных посвящена следующая глава.

1. Сбор, анализ и обработка данных
   1. Анализ доступных способов сбора информации

В настоящее время существует много способов сбора информации, которые варьируются от коммерческих систем, специализированных на сборе текстовой информации из указанных ресурсов, до инструментов предоставляемых самими хозяевами ресурсов.

К первому типу инструментов для сбора информации можно отнести автоматизированные системы по сбору текстовых данных, например, “RoadRunner” [5]. Такие системы специализированы на сборе текстовых данных с сайтов с большим объёмом информации. Они разделены на те, которые могут собирать связанный и бессвязный текст. В основном, такие системы узкоспециализированны под конкретные задачи или ресурсы, поэтому их использование не будет применяться в данной работе.

Ко второму способу сбора данных относятся инструменты по выгрузке данных предоставленные самими разработчиками источников информации. Такой способ наиболее удобен из-за того, что такие инструменты интегрированы в сам ресурс и работают быстро и качественно. Однако в использовании такого способа есть множество проблем:

* Не все разработчики предоставляют такие инструменты;
* Содержание выгружаемой информации определяется самим хозяином ресурса, и изменение ее объема может быть одним из коммерческих ходов хозяина ресурса.
* Зачастую сам инструмент предоставляется на платной основе.

В силу вышеперечисленных причин второй способ тоже не будет использоваться.

Другим популярным способом сбора информация является использование API предоставляемого разработчиками. Данный способ позволяет быстро и удобно запрашивать нужную информацию из источника. API можно разделить на 2 группы: предоставляемые разработчиками ресурса и предоставляемые коммерческими ресурсами.

К первой группе относят наиболее популярные сайты, хозяева которых имеют время и деньги на создание подобного интерфейса, в большинстве случаев данный интерфейс предоставляется бесплатно с платным расширенным функционалом.

Если ресурс относительно небольшой, но популярный у пользователей, то есть возможность найти коммерческую платформу по предоставлению API, которая предоставляет пользователю удобный интерфейс работы, а сама при помощи алгоритмов собирает информацию с указанного сайта. Такая платформа в большинстве случаев платна, результаты работы могут отличаться от ожидаемых, поскольку не пользователь решает как собирается информация, а скорость ее работы сравнима с обычным парсером, если запрашиваемая информация уже не хранится на платформе, а собирается с нуля.

Последний вариант это собственноручная обработка интернет-ресурса. Поскольку на каждом сайте счет объявлений о продажах квартир идет на тысячи, то человек не сможет самостоятельно перенести все данные с сайта в собственное хранилище данных. В этом случае принято использовать парсер сайта, который является частным случаем поискового робота “web crawler” поскольку предназначается исключительно для обработки лишь одного сайта.

В настоящее время задача извлечения информации из интернет-источников относительно сложна поскольку не существует единого стандарта построения сайтов. Каждый сайт строится в зависимости от своей тематики, требований заказчика, опыта разработчиков и фантазии дизайнеров, поэтому к каждому сайту нужен индивидуальный подход по сбору с него информации. Единственное, что объединяет различные сайты, это язык HTML, который определяет внешний вид Интернет-ресурсов, но не может описать его содержание и интерфейс DOM, который позволяет получить доступ к информации, содержащейся на странице. Для данной работы используется простой парсер, который собирает информацию из нескольких источников и объединяет ее, после чего сохраняет для последующей обработки.

* 1. Обзор собираемой информации

В данной работе информация собирается при помощи созданного парсера для сайтов http://ivanovodom.ru, откуда берется информация о продаваемых квартирах и для сайта http://dom.mingkh.ru с которого собирается информация о строениях в которых располагаются продаваемые квартиры.

Ниже представлена информация собираемая о квартире:

* Цена [1];
* Общая площадь [2];
* Жилая площадь [3];
* Площадь кухни [4];
* Количество комнат [5];
* Этаж [6];
* Район [7];
* Улица [8];
* Номер дома [9];

И информация, которая собирается о доме:

* Наличие мусоропровода [10];
* Наличие во дворе детской площадки [11];
* Наличие во дворе спортивной площадки [12];
* Широта [13];
* Долгота [14];
* Количество этажей в доме [15];
* Год постройки здания [16];
* Тип материала стен [17];
* Тип материала перекрытий [18];
* Тип крыши [19];
* Тип холодного водоснабжения [20];
* Тип горячего водоснабжения [21];
* Тип газоснабжения [22];
* Класс энергетической эффективности здания (если установлен) [23];
* Тип противопожарной системы (если установлена) [24];
* Тип отопления [25];
* Тип водоснабжения [26];
* Тип электроснабжения [27].
  1. Анализ собираемой информации

Первой задачей после сбора данных является их очистка и приведение к удобному для работы методов машинного обучения виду.

В собранных данных присутствуют следующие виды признаков:

* Категориальные – признаки, выраженные словесно;
* Количественные – признаки, отдельные варианты которых имеют числовое выражение; отражают размеры, масштабы изучаемого объекта или явления;
* Бинарные – признаки, имеющие только два варианта значений.
* Под придачей данным вида удобного для анализа и машинного обучения подразумевается их кодирование, т.е. перевод булевых и строковых значений в числовые. Сначала больше всего интересны категориальные признаки. Категориальные признаки называют по-разному: факторными, номинальными. Их значения определяют факт принадлежности к какой-то категории. Примеры таких признаков: пол, страна проживания, номер группы, категория товаров и т.п. Ясно, что для компьютерной обработки вместо «понятного для человека» значения (в случае района — “Ленинский”, “Фрунзенский”, “Октябрьский”, и т.п.) хранят числа.

Все категориальные признаки проходят кодировку при помощи самого простого метода, кодирование через уникальные значения, такой метод кодирует каждую новую категорию у признака следующим по порядку числом начиная с ноля. В ходе работы были использованы и другие способы кодирования, такие как кодирование категорий через среднее значение целевой переменной или через количество вхождений объектов с данной категорией в признаке или же кодирование категории через специальные коэффициенты (например, для каждого материала стен присваивается свой числовой коэффициент). Данные способы никак не повлияли на дальнейшую работу алгоритма, при этом сильно усложняли понимание самих категорий.

При помощи кодирования были изменены значения признаков 7, 8, 9, 17-27.

После кодирования данных произведём отбор признаков по принципу пропущенных значений. Для этого посчитаем количество пропущенных значений для каждого признака.

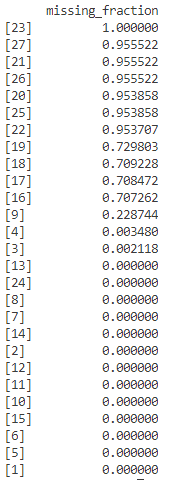


Рисунок 1 – Потери значений у незакодированной выборки

Как видно из рисунка 1, большой показатель пропусков имеют 11 признаков, все они несут какую-либо информацию о доме, а не о квартире, которая в нем находится. Это показывает, что в основном данные недополучены с сайта ЖКХ.

Только у признаков, для которых значение на рисунке 1 больше 0,6, хранится какое-то значение, что составляет только 40% или меньше от общего количества. Поскольку у 60% или более объектов эти признаки ничего не значат, то от таких признаков совершенно не будет толку во время обучения или они даже могут ухудшить точность модели. Поэтому данные признаки подлежат удалению из выборки.

Далее проверим корреляцию между признаками, поскольку, чем больше связь между независимыми признаками, тем хуже будут работать алгоритмы машинного обучения.

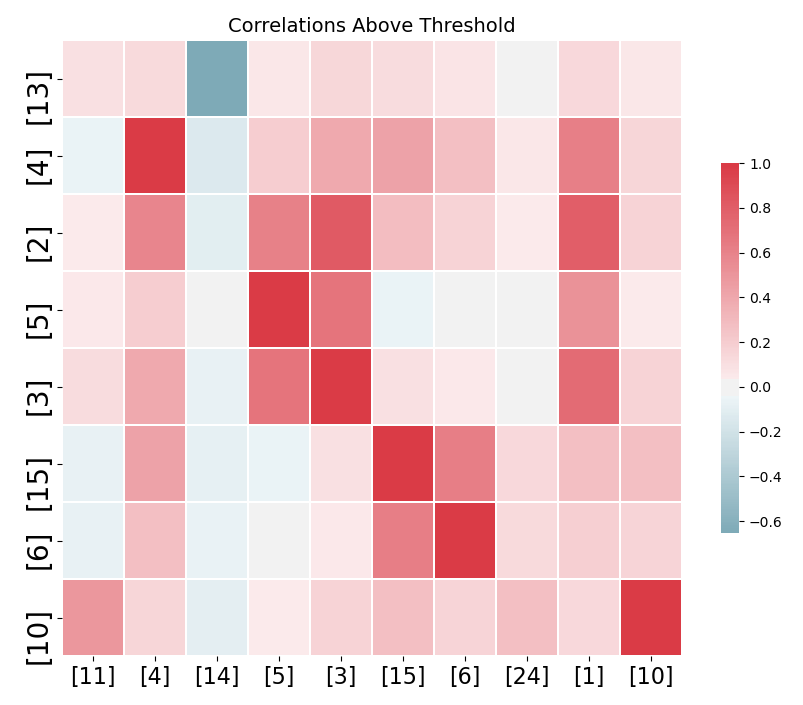


Рисунок 2 – Тепловая карта корреляции признаков

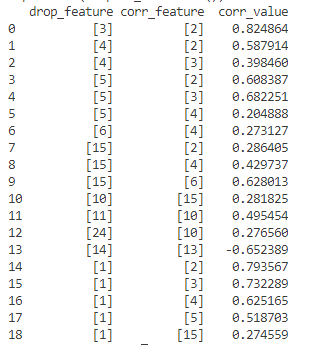


Рисунок 3 – Коэффициенты корреляции признаков

Как видно из рис. 2 и рис. 3, корреляция присутствует между закономерными величинами. Присутствует высокая связь между общей (признак 2), жилой (признак 3) и кухонной (признак 4) площадью, что очевидно поскольку общая площадь квартиры содержит в себе как жилую площадь, так и площадь кухни.

Также велика корреляция между общей площадью, жилой площадью и количеством комнат (признак 5), корреляция между ними около 60-70 процентов.

На рис. 3 видно, что признак «стоимость квартиры» имеет корреляцию с достаточно большим количеством признаков, в данном случает большее значение как раз имеет положительный эффект, так как, чем больше корреляция между целевым и независимыми признаками, тем ценнее эти признаки будут для алгоритмов машинного обучения.

Далее преобразуем данные по следующим правилам:

* 1. Поскольку в оценке принято измерять стоимость недвижимости не целиком, а за квадратный метр, то исходя из этого уберем из датасета лишние площади и стоимость квартиры и заменим их на стоимость за квадратный метр и ее общую площадь.
  2. Поскольку нам известны координаты для каждой квартиры, которые так же отражают и адрес, то это позволяет нам избавиться от признаков содержащих районы, улицы и номера домов.
  3. Сами по себе координаты ничего не значат, интерпретировать их сложно, а в рис. 3 видно, что они никак не коррелируют с целевым параметром, для исправления этого вместо координат рассчитаем удаленность каждой квартиры от центра города (поскольку принято считать что центр города это наиболее развитый и дорогой участок и по мере удаления от него стоимость квадратного метра жилья будет уменьшаться)
  4. Так же координаты могут позволить определить направление расположения квартиры относительно центра, что позволит определить перекос стоимости квадратного метра жилья в зависимости от стороны света.
  5. Так же из рис. 3 мы видим что цена имеет корреляцию с количеством этажей в здании (строка 18), хоть корреляция и невелика, но показывает о взаимосвязи между целевой и независимой переменными, поэтому оставим данный признак в выборке.
  6. Поскольку в выборке остается количество этажей в здании, то логично оставить и признак показывающий этаж на котором находится квартира. В оценочной деятельности из этих признаков можно выявить факт расположения квартиры на первом или последнем этажах, что сильно влияет на ее стоимость.

После всех преобразований у нас получился значительно сокращенный датасет, состоящий из следующих параметров:

Ниже представлена информация, собираемая о квартире:

* Цена за квадратный метр [1];
* Удаленность от центра [2];
* Направление относительно центра (азимут) [3];
* Этаж [4];
* Этажность здания [5];
* Общая площадь здания [6]

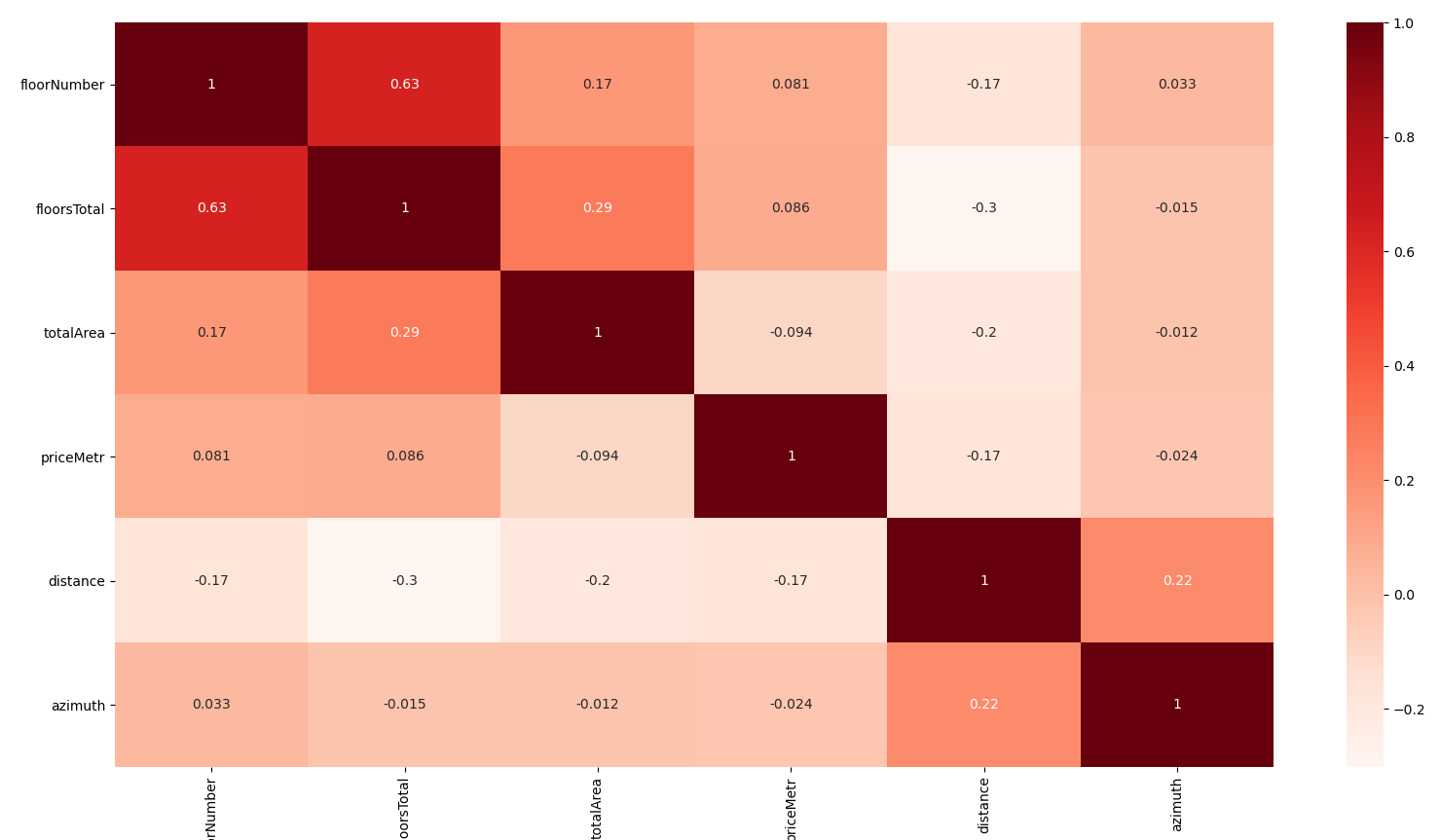


Рисунок 4 – Коэффициенты корреляции оставшихся признаков

Из рис. 4 видно, что высокая корреляция только между этажом, на котором находится квартира, и общим количеством этажей в доме, что может быть связано с тем, что в большем количестве объектов квартира находится на первом или на последнем этажах. В остальном же другие признаки слабо коррелируют между собой, и при этом имеют слабую корреляцию со стоимостью за квадратный метр.

Так же из рисунка видно, что бывшая целевая перменная «цена» квартиры, а в данный момент «стоимость квадратного метра», имеет слабую кореляцию со всеми остальными признаками. Если раньше кореляция между целевым признаком и общей площадью составляла 0.79, то сейчас она упала до -0.094, что сильно сказалось на значимости этого признака. На текущий момент наибольшую корреляцию с целевым признаком имеет «удаленность» от центра города, равную 0.17. Из-за такой низкой значимости между независимыми признаками и целевой перменной, можно предсказать низкую результативность работы линейной регрессии, так как она основана именно на взаимодействии целевого и независимых признаков.

1. Выбор алгоритмов МО
   1. Определение задачи регрессии

Согласно [4], задача регрессии отличается от других задач машинного обучения тем, что допустимым ответом для нее является действительное число или числовой вектор, то есть оценка или прогноз.

Основными целями регрессии являются:

* Построение прогноза на основе имеющихся данных;
* Оценка влияния параметров модели на целевую функцию;
* Оценка значимости построенной модели и ее адекватности относительно реальных данных.

Первая цель отражает само решение задачи регрессии, поскольку является результатом работы алгоритмов, решающих эту задачу. Вторая цель неотделима от первой, так как ее достижение позволяет определить наиболее значимые признаки и улучшить точность работы разрабатываемой модели. Третья же цель позволяет оценить качество уже самой построенной модели или то, как точно она отражает реальные данные

Достижение любой из этих целей необходимо чтобы построить адекватную модель для оценки или прогнозирования данных. Любой алгоритм, будь то линейная регрессия, ступенчатая регрессия, логическая регрессия, бустинг или случайный лес, строят свою работу на подборе оптимального уравнения, которое отражает предоставляемые данные. В качестве искомой величины которого в нашем случае выступает цена объекта недвижимости, а переменными признаки это объекта.

Оценка качества модели производится при помощи общего для этого класса задач, коэффициента детерминации [6], который показывает насколько похож полученный результат с истинным, а также при помощи средней абсолютной ошибки, о чем рассказано в [7]. Основной их обзор и применимость можно увидеть в статье [17].

Коэффициент детерминации в основном всегда используется для задач регрессионного анализа, поскольку является показательным и легко интерпретируемым. , является еще одним показателем, который можно использовать для оценки модели, и он тесно связан с MSE, но имеет преимущество в том, что являются ли выходные значения очень большими или очень маленькими, R² всегда будет между -∞ и 1.

|  |  |
| --- | --- |
|  | (1) |
|  |  |

Когда R² отрицательно, это означает, что модель хуже, чем предсказание среднего значения.

MSE модели рассчитываются, как указано в 2, в то время как MSE базовой линии определяется как в формуле 3:

|  |  |
| --- | --- |
|  | (2) |
|  | (3) |

где,

выборочное среднее

 соответственно, фактические и расчетные значения объясняемой переменной,.

Значение, близкое к 1, указывает на модель с ошибкой, близкой к нулю, а значение, близкое к нулю, указывает на модель, очень близкую к базовой линии.

В заключение, R² - это соотношение между тем, насколько хороша наша модель, и тем, насколько хороша модель наивного среднего.

Среди всех встречающихся оценок ошибки прогнозирования стоит отметить две, которые в настоящее время, являются наиболее популярными: MAE и MAPE.

MAPE – средняя абсолютная ошибка в процентах, формула 4:

|  |  |
| --- | --- |
|  | (4) |

MAE – средняя абсолютная ошибка, формула 5:

|  |  |
| --- | --- |
|  | (5) |

MAE широко используется в финансах, где ошибка в 10 долларов обычно в два раза хуже, чем ошибка в 5 долларов. С другой стороны, метрика MSE считает, что ошибка в 10 долларов в четыре раза хуже, чем ошибка в 5 долларов. MAE легче обосновать, чем RMSE.

MAPE является лишь более легкой для восприятия версией MAE, поскольку переводится в проценты.

Кроме указанных иногда используют другие оценки ошибки, менее популярные, но также применимые.

MSE – среднеквадратичная ошибка, описанная выше формулой 2.

RMSE – квадратный корень из среднеквадратичной ошибки, описан формулой 6

|  |  |
| --- | --- |
|  | (6) |

MSE это наиболее простой и распространенный показатель оценки регрессии, но и наименее полезный. Основным его недостатком является то, что если произойдет один очень плохой прогноз, то возведение в квадрат сделает ошибку еще хуже, и это может исказить метрику в сторону переоценки модели в худшую сторону. Это будет сильно проявляться при зашумленных данных, из-за этого даже у практически идеальных данных может быть высокий MSE, что усложняет его интерпретацию.

RMSE - это просто квадратный корень из MSE. Квадратный корень введен, чтобы масштаб ошибок был таким же, как масштаб целей. Но поскольку этот показатель рассчитывается на основе MSE, то мы все равно можем сравнивать наши модели, используя MSE, поскольку MSE упорядочит модели так же, как RMSE.

В MAE ошибка рассчитывается как среднее абсолютных разностей между известными значениями и прогнозами. MAE - это линейная оценка, которая означает, что все индивидуальные различия взвешены в среднем одинаково. Например, разница между 10 и 0 будет вдвое больше разницы между 5 и 0. Однако то же самое не будет верным для RMSE.

Что важно в этой метрике, так это то, что она наказывает большие ошибки, что в итоге выдает не такой плохой результат, как MSE. Таким образом, данная метрика не так чувствительна к выбросам, как среднеквадратическая ошибка.

* 1. Обзор методов, решающих задачу регрессии
     1. Обзор линейной регрессии

Регрессионный анализ является важным статистическим методом для анализа различных данных. Он позволяет идентифицировать и охарактеризовать взаимосвязи между несколькими факторами. Он также позволяет идентифицировать важные для оценки признаки и рассчитать значение для предоставляемого объекта недвижимости.

Согласно [8], линейная регрессия использует модель, которая описывает отношения между зависимыми переменными (признаками) и независимыми переменными (целевой признак).

Как указано в [9], общая модель линейной регрессии может выглядеть как показано в формуле 7.

|  |  |
| --- | --- |
| , | (7) |

где b — это параметры модели,

ɛ— случайная ошибка модели.

Данная модель будет называться линейной если функция регрессии f(x,b) имеет вид как в формуле 8.

|  |  |
| --- | --- |
|  | (8) |

где b — это параметры (коэффициенты) регрессии,

x — регрессоры (признаки модели),

k — количество признаков модели.

Коэффициенты линейной регрессии показывают скорость изменения зависимой переменной по данному фактору, при фиксированных остальных факторах (в линейной модели эта скорость постоянна).

Как видно из рисунка, данная функция не содержит нелинейных коэффициентов и, таким образом, подходит только для моделирования линейных сепарабельных данных. При работе алгоритма производится взвешивание значений каждой переменной x\_n с помощью весового коэффициента b\_n. Данные весовые коэффициенты a\_n.

Как указано в [10], регрессионные модели и линейная регрессия в частности имеют свои достоинства и недостатки.

К достоинствам принято относить:

* Простоту вычислительных алгоритмов;
* Наглядность и интерпретируемость результатов.

К недостаткам:

* Невысокая точность прогноза (в основном - интерполяция данных);
* Субъективный характер выбора вида конкретной зависимости (формальная подгонка модели под эмпирический материал);
* Отсутствие объяснительной функции (невозможность объяснения причинно-следственной связи);
* Высокая чувствительность к исходным данным (к выбросам и размытости данных).
  + 1. Обзор случайного леса

Случайный лес — это алгоритм машинного обучения, суть которого в использовании комитета (ансамбля) решающих деревьев. Алгоритм применяется для задач классификации, регрессии и кластеризации. Основная его идея заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень неточный результат, но за счёт их большого количества результат получается хорошим [11]. Для более полного представления об этом алгоритме следует дать разъяснение термину «решающее дерево».

Решающее дерево (также имеющее названия дерево классификации или регрессионное дерево) – средство поддержки принятия решений, использующееся в машинном обучении, анализе данных и статистике. Структура дерева представляет собой «листья» и «ветки». На «ветвях» дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Подобные деревья решений широко используются в интеллектуальном анализе данных. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе [12]. Стандартная структура дерева решений представлена на рисунке 5.

Наиболее распространённый способ построения случайного леса следующий:

* 1. Генерируется случайная подвыборка с повторениями размером N из обучающей выборки.
  2. Строится решающее дерево, классифицирующее образцы данной подвыборки, при этом в ходе создания нового узла дерева будет выбираться набор признаков, на основе которых производится разбиение.
  3. Дерево продолжает строиться до полного исчерпания подвыборки и не подвергается процедуре отсечения ветвей.

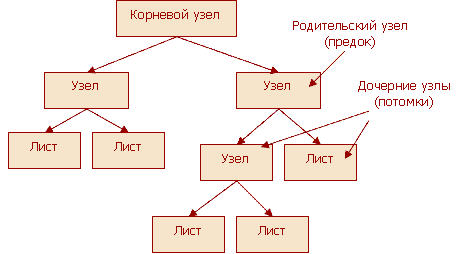


Рисунок 5 – Структура дерева принятия решений

К важным для нас достоинствам данного метода можно отнести:

* Способность обрабатывать данные с большим числом признаков;
* Нечувствительность к масштабированию значений признаков;
* Есть методы оценки значимости признаков в модели;

К недостаткам можно отнести:

* Большой размер итоговых моделей.
  + 1. Метод XGBoost

XGBoost — это алгоритм, который в последнее время доминирует в прикладном машинном обучении и соревнованиях Kaggle для структурированных или табличных данных [13].

Этот метод является реализацией градиентного бустинга деревьев решений, предназначенных для улучшения их скорости и производительности [14].

Градиентный бустинг — это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. Обучение ансамбля проводится последовательно в отличие, например, от бэггинга. На каждой итерации вычисляются отклонения предсказаний уже обученного ансамбля на обучающей выборке. Следующая модель, которая будет добавлена в ансамбль, будет предсказывать эти отклонения. Таким образом, добавив предсказания нового дерева к предсказаниям обученного ансамбля, мы можем уменьшить среднее отклонение модели, которое является целью оптимизационной задачи. Новые деревья добавляются в ансамбль до тех пор, пока ошибка уменьшается, либо пока не выполняется одно из правил "ранней остановки".

Функция градиентного бустинга для деревьев решений выглядит как в формуле 9:

|  |  |
| --- | --- |
|  | (9) |

где, — это функция потерь.

— это значение i-го элемента обучающей выборки и сумма предсказаний первых t деревьев соответственно

— набор признаков i-го элемента обучающей выборки.

— функция (в нашем случае дерево), которую мы хотим обучить на шаге t.

— это регуляризация функции, где t — шаг, на котором происходит обучение.

1. Результаты машинного обучения

В данной главе будут представлены практические результаты применения методов регрессионного анализа для очищенного набора данных из открытых источников. В качестве рабочих алгоритмов будут использоваться методы, входящие в состав математической библиотеки для языка Python под названием Scikit-learn и XGBoost.

* 1. Подготовка выборок

Первым этапом перед использованием любого метода является загрузка данных и первичная их подготовка. На следующем рисунке показан код подготовки данных, а именно их загрузка в программу, расчет удаленности от центра и азимута (направления) на основе координат объекта недвижимости.

Данный код представлен в приложении 2. В коде используется метод для расчета азимута и удаленности, после чего полученный значения округляются и добавляются в изначальный датафрейм.

После этого в датафрейме усредняются значения, что подразумевает отсев всех предложений о продаже квартиры у которых стоимость за квадратный метр больше или меньше определенных рамок. Для имеющейся выборки был рассчитана медиана цены за квадратный метр, которая составила 41292 рублей за квадратный метр, нижняя граница равна 38773 рублей за квадратный метр и верхняя граница составила 45238 рублей за квадратный метр. Исходя из этого в выборку попали предложения находящиеся между верхней и нижней границей, что составило 4675 записей от исходных.

Как только датафрейм очистится от лишних записей, из него начинают формироваться выборки. Для этого сначала отделяется зависимая переменная (цена за кв. метр) от независимых, и формируются обучающая и тестовые выборки в соотношение 75% к 25% от всех данных, это означает, что для обучения моделей будут использоваться 3506 объектов, а для их тестирования 1115.

После того как учебная и тестовая выборки были подготовлены можно приступать к созданию и обучению моделей на тренировочной выборке. Первой моделью которая будет построена является «случайный лес».

* 1. Метод случайного леса

Обзор данного метода был произведен в главе 3.2.2. Ниже будет приведен пример программы, отвечающей за создание и обучение модели. Также в коде присутствует прогнозирование цены за квадратный метр и оценка результатов его работы.

Данный подход является популярным выбором при решении задачи регрессионного анализа, о чем говорится в статьях [15] и [16].

Данная модель была построена на 2000 деревьях с MAE в качестве функции разделения, с глубиной равной 50 и количеством вариантов для разделения равным 5.

# Создаем модели

rf\_model = RandomForestRegressor(n\_estimators=2000,

                                 n\_jobs=-1,

                                 bootstrap=False,

                                 criterion='mae',

                                 max\_features=2,

                                 random\_state=1,

                                 max\_depth=50,

                                 min\_samples\_split=5

                                 )

#Проводим подгонку модели на обучающей выборке

rf\_model.fit(train\_X, train\_y)

#Вычисляем предсказанные значения цен на основе валидационной выборки

rf\_prediction = rf\_model.predict(val\_X).round(0)

#Вычисляем и печатаем величины ошибок при сравнении известных цен квартир из валидационной выборки с предсказанными моделью

print("Величины ошибок для случайного леса")

print\_metrics(rf\_prediction, val\_y)

print()

Рисунок 6 – Код использования случайного леса

На рис. 6 представлен код того как сначала создается модель случайного леса с заданными параметрами, после чего ей дается учебная выборка состоящая из выборки независимых признаков «train\_X» и выборки с целевой переменной «train\_y» для обучения модели и потом при помощи команды «predict» производится ее тестирование на тестовой выборке «val\_X».

В ходе выполнения вышеуказанного фрагмента кода были получены следующие результаты:

Коэффициент детерминации (R²) равен 0.53 или же 53%;

Средняя абсолютная ошибка: 3,84 %;

Данные результаты говорят о том, что модель получилась средняя по качеству, а при прогнозировании цены для объекта недвижимости погрешность будет составлять 3,8% от этой цены.

Ниже рис. 7 представлен код построения графика со 100 первыми значениями для тестовой выборки и спрогнозированными моделью значениями. На рисунке 8 по оси X отображены идентификаторы объектов в выборке, а по оси Y зависимая переменная (стоимость за квадратный метр). Синим изображена реальная стоимость объекта (actual), а оранжевым, рассчитанная при помощи модели случайного леса на основании независимых признаков.

df\_0 = pd.DataFrame({'Actual': val\_y, 'Predicted': rf\_prediction})

df1 = df\_0.head(100)

df1.plot(kind='bar', figsize=(10, 8))

plt.title('Случайный лес')

plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')

plt.show()

Рисунок 7 – Результаты применения случайного леса

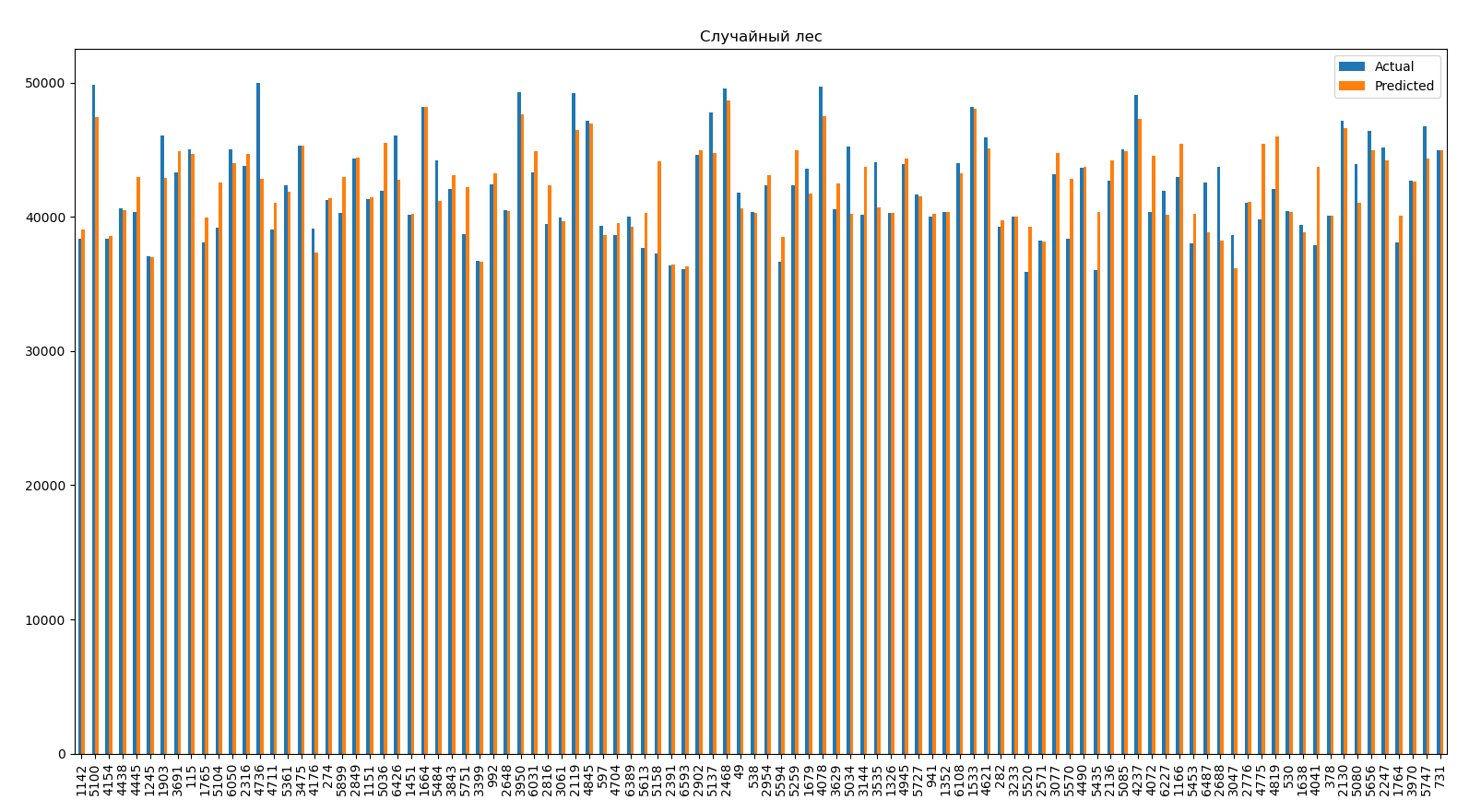


Рисунок 8 – Результаты применения случайного леса

Как видно из рисунка 7 модель более точно прогнозирует значения для обьектов цена за квадратный метр которых ближе всего к медиане равной 41292 рублей за метр.

Данная модель позволяет узнать какие из признаков по мнению модели с какой степенью влияли на цену обьекта недвижимости.

Удаленность от центра (0.290065)

Общая площадь (0.229959)

Азимут (направление от центра) (0.195771)

Количество этажей в здании (0.166247)

Этаж (0.117957)

Как видно из этих показателей, больше всего, по мнению модели, на цену квартиры влияет удаленность от центра города. Это представляется логичным, поскольку центр в любом городе является наиболее экономически развитым и имеет более дорогую жилую недвжимость. Такие низкие резльтаты важности признаков еще раз показывают, как сильно корреляция между целевым и независимыми признакми может влиять на результат.

* 1. Метод бустинга на деревьях

Обзор данного метода был произведен в главе 3.2.3. Ниже так же будет приведен пример программы, отвечающей за создание и обучение модели. При помощи библиотеки XGBoost, создать модель бустинга на деревьях достаточно просто, для это вызывается соответсвующий метод с параметрами как показано на рисунке 9. К видно из рисунка 9, модель для своей работы использует 2000 деревьев с максимальной глубинойравной 50, остальные параметры были выбраны по умолчанию.при таком соотношении количество деревьев является половиной поличество обьектов в тренеровочной выборке и в тоже время их почти в 2 раза больше по отношению к тестовой выборке.

# Создаем модели

xgb\_model = xgb.XGBRegressor(objective ='reg:gamma',

                             learning\_rate = 0.01,

                             max\_depth = 50,

                             n\_estimators = 2000,

                             nthread = -1,

                             eval\_metric = 'gamma-nloglik',

                             )

#Проводим подгонку модели на обучающей выборке

xgb\_model.fit(train\_X, train\_y)

#Вычисляем предсказанные значения цен на основе валидационной выборки

xgb\_prediction = xgb\_model.predict(val\_X).round(0)

#Вычисляем и печатаем величины ошибок при сравнении известных цен квартир из валидационной выборки с предсказанными моделью

print("Величины ошибок для бустинга")

print\_metrics(xgb\_prediction, val\_y)

print()

Рисунок 9 – Код использования XGBoost

Ниже рис. 10 представлен код построения графика со 100 превыми значниями для тестовой выборки и спрогнозированными моделью значениями. На рисунке 11 по оси X отображены идентификаторы объектов в выборке, а по оси Y зависимая переменна (стоимость за квадратный метр). Синим изображена реальная стоимость объекта (actual), а оранжевым, рассчитанная при помощи модели случайного леса на основании независимых признаков.

df\_0 = pd.DataFrame({'Actual': val\_y, 'Predicted': xgb\_prediction})

df1 = df\_0.head(100)

df1.plot(kind='bar', figsize=(10, 8))

plt.title('Бустинг')

plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')

plt.show()

Рисунок 10 – Код вывода результата для бустинга

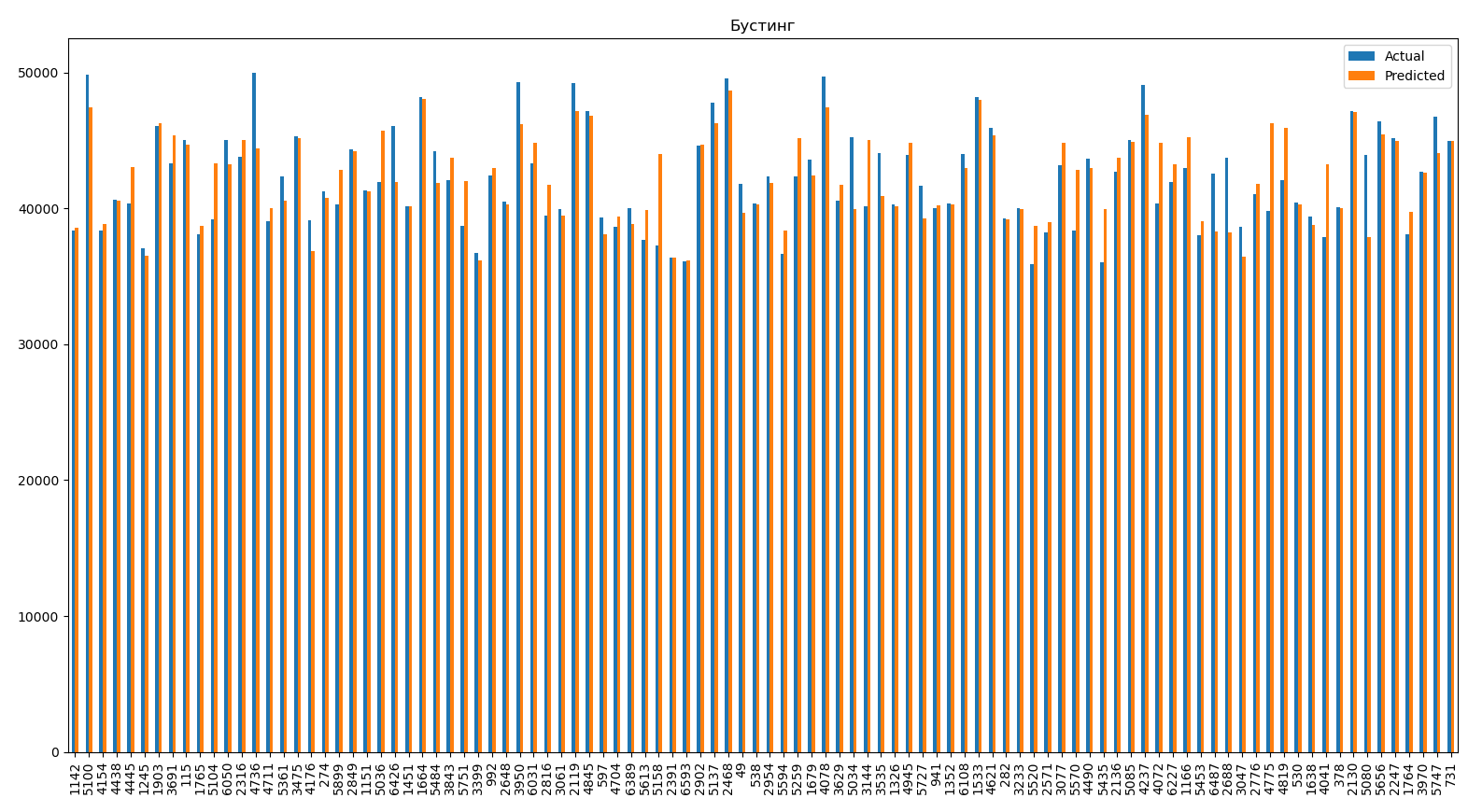


Рисунок 11 – Результаты применения XGBoost

На рисунке 11 види примерно такую же ситуацию как и на рисунке 8, только в этот раз количество обьектов для которых модель достаточно точно прогнозирует стоимость квадратного метра значительно меньше, так видно что если фактическая стоимость обьекта достаточно сильно отклоняется от медианы равной 41292 рублей за метр, то ошибка получается куда значительнее чем в методе случайного леса, это обьясняет более высокое значение MAPE, равное 3.97 %, а так же более низкую совместимость модели с данными, которая равна 50%

* 1. Метод линейной регрессии

Обзор метода линейной регрессии был произведен в главе 3.2.1. Ниже так же приведен пример программы, отвечающей создание и обучение модели, а также прогнозирование цены этим методом.

Для использования данного метода не требуется указывать как либо дополнительных параметров. Как и в источнике [17] в данной работе создается модель линейной регресси, после чего обучается на тренировочных выборках и получает тестовую выборку для оценки стоимости квадратного метра для уже известных нам квартир.

# Создаем модели

reg\_model = LinearRegression()

#Проводим подгонку модели на обучающей выборке

reg\_model.fit(train\_X, train\_y)

#Вычисляем предсказанные значения цен на основе валидационной выборки

reg\_prediction = reg\_model.predict(val\_X).round(0)

#Вычисляем и печатаем величины ошибок при сравнении известных цен квартир из валидационной выборки с предсказанными моделью

print("Величины ошибок для линейной регрессии")

print\_metrics(reg\_prediction, val\_y)

print()

Рисунок 12 – Код использования линейной регрессии

df\_0 = pd.DataFrame({'Actual': val\_y, 'Predicted': reg\_prediction})

df1 = df\_0.head(100)

df1.plot(kind='bar', figsize=(10, 8))

plt.title('Линейная регрессия')

plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')

plt.show()

Рисунок 13 – Код вывода результата для линейной регрессии

На рисунке 14 ситуация совершенно отличается от предыдущих двух моделей. Поскольку линейная регрессия сильно зависит от данных, то ей для работы нужны, как можно более усредненные данные. В наших же имеется существенный разброс по цене за квадратный метр. Данный недочет выливается в ярко выраженное на графике несоответствие. Предсказанные значения стремятся к истинным, но при этом сильно прижаты к медиане. Из этого вытекает, что модель выдает относительно точные результаты только в случаях, когда искомая квартира находится примерно на уровне 41292 рублей за квадратный метр. При таком разбросе модель линейной регрессии достигает лишь 11% соответсвия реальным данным, а ее погрешность находится в районе 6,7%, что тоже в два раза больше, чем у предыдущих двух моделей.

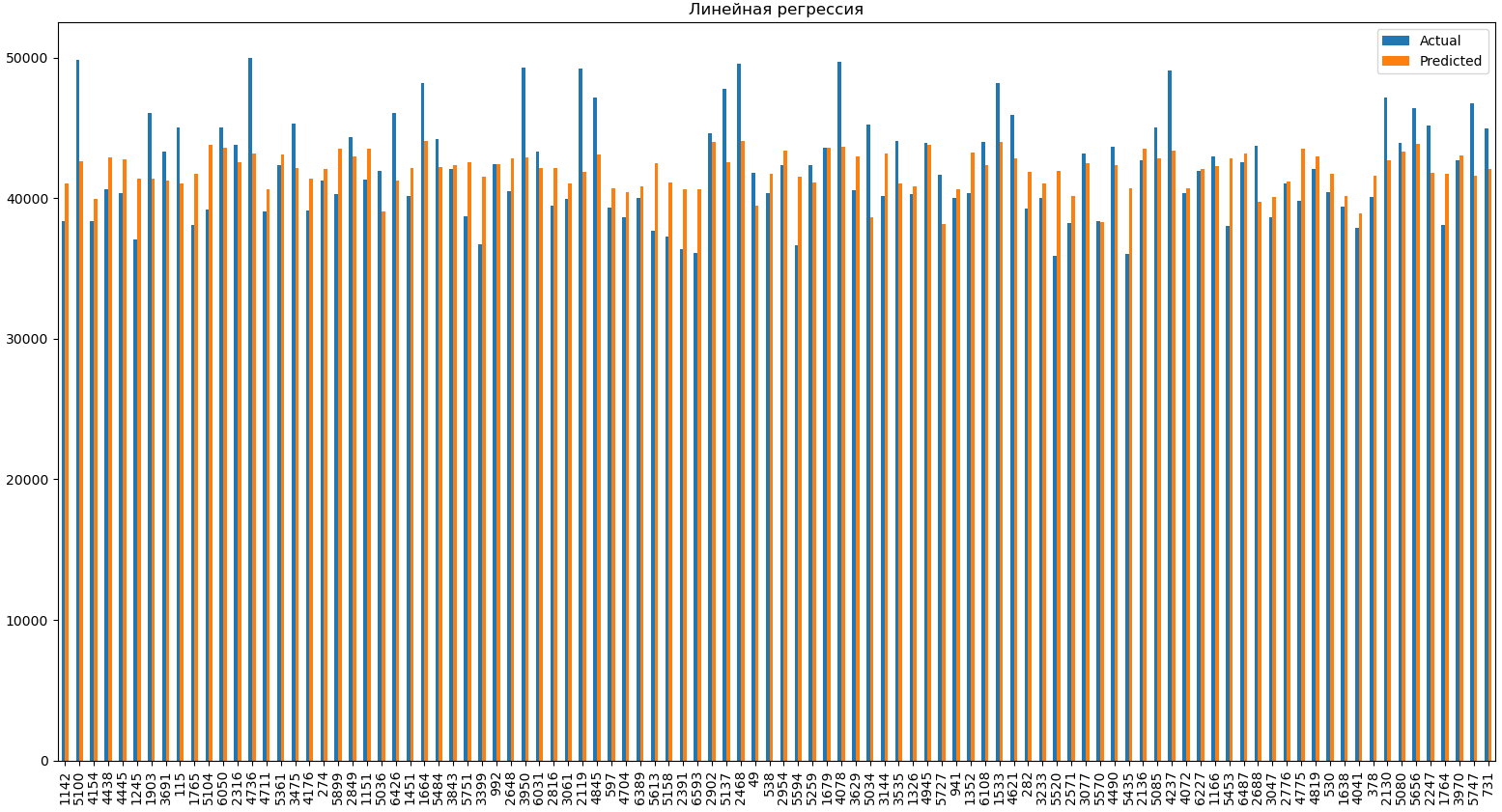


Рисунок 14 – Результаты применения линейной регрессии

* 1. Выводы по регрессионному анализу

Для оценки качества работы методов, в первую очередь будем смотреть на коэффициент детерминации (R²).

Из всех трех методов хуже всего с задачей справилась модель множественной линейной регрессии, поскольку ее достоверность реальным данным составила всего 11%. В случае с этой моделью по итогам ее работы можно построить уравнение линейной регрессии на основе известных коэффициентов. Данное уравнение будет выглядеть следующим как в формуле 10:

|  |  |
| --- | --- |
|  | (10) |

X1 – этаж

X2 – этажность здания

X3 – общая площадь

X4 – удалённость от центра города

X5 – направление от центра города

Получение настолько плохого результата произошло по следующим причинам:

* Большой разброс в значениях данных (полное отсутствие линейности в данных);
* Большое количество выбросов, которые сильно влияли на работу модели;
* Недостаточное количество данных в тренировочной выборке. Зачастую тренировочные данные насчитывают десятки тысяч объектов, в нашем же случае вся выборка (включая тестовую) насчитывает всего 4600 объектов недвижимости.

Лучше себя показал бустинг, поскольку его результат составил 50% соответствия с реальными данными, а его погрешность при расчетах не превышала 3,9%. На его работу так же могли повлиять следующие факторы:

Большой разброс в значениях данных;

Недостаточное количество данных в тренировочной выборке.

Плохое качество данных. Большое количество пропусков в независимых признаках.

Лучший результат показала модель случайного леса, которая сама по себе является более универсальной, чем линейная регрессия, и более простая, чем бустинг. Точность модели случайного леса составила 53%, а его погрешность при расчетах не превышала 3,8%. Больше всего на уменьшение точности модели повлиял размер выборки, поскольку точность данного метода напрямую зависит от размера обучающей выборки.

Как было сказано выше на все методы повлиял раброс в значениях стоимости за метр жилья. Уменьшение диапазона доступных значений повлияет как на точность моделй (она увеличится), так и на их соответствие реальным данным. Ниже приведена таблица, показывающая влияние уменьшения разброса в стоимости за квадратный метр, на коэффициент детерминации и MAPE.

Таблица 2 – Таблица влияния выбросов

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Количество выбросов | Границы | | RF | | XGBoost | | LR | |
| Нижняя | Верхняя | R² | MAPE | R² | MAPE | R² | MAPE |
| 0 | 0 | 142857.0 | 48% | 8,8% | 42% | 8,9% | 5,6% | 14,5% |
| 1322 | 34091.0 | 52567.6 | 59% | 4,3% | 53% | 4,5% | 13% | 8,1% |
| 2644 | 36752.2 | 48184.0 | 55% | 3,1% | 50% | 3,2% | 11% | 5,5% |
| 2647 | 38773.0 | 45238.0 | 53% | 2,12% | 47% | 2,15% | 9,1% | 3,5% |
| 5287 | 40211.4 | 43160.0 | 51% | 0.928% | 45% | 0.998% | 7.4% | 1.75 % |

Как видно из таблицы 2, чем больше мы сужаем разброс в значениях стоимости за квадратный метр, тем больше падает коэффициент детрминации и тем точнее становится модель. Для эксперимента это значит, что на текущих данных мы можем получить максимально точный прогноз только при условии что оцениваемый обьект по признакам будет находится на медиане, т.е если его стоимость за квадратный метр, в рассматриваемых условиях, будет приближаться к 41292. Чем дальше от медианы будет отклоняться оцениваемый объект, тем больше будет вырастать погрешность при расчете стоимости.

По результатам изучения данных таблицы 2 можно сделать вывод, что удаление выбросов из датасета позволит сделать модели более точными, но менее универсальными, что в свою очередь сделает их пригодными для оценки только малого количества категорий квартир.

Для проверки работоспособности моделей проведем эксперимент на продаваемой в настоящий момент времени квартире.

Рассмотрим обьявление [18], в нем продается двухкомнатная квартира на третьем этаже в пятиэтижном доме, общей площадью 47,6 м². Преобразуем адрес в координаты при помощи любого сервиса геокодирования и в итоге соберем все требующиеся данные о картире.

На рисунке 13 показан фрагмент кода, который отвечает за оценку данной квартиры обученными моделями и получение итогового результата.



Рисунок 15 – Код создания и оценки квартиры

Согласно модели случайного леса данная квартира должна стоить 1 981 000 рублей, а если учитывать погрешность равную величине ошибки 0.928%, то получим интервал в рамках которого должна варьироваться стоимость данной квартиры, а именно от 1 962 616,32 до 1 999 383,68 рублей.

Модель бустинга на деревьях показала, что данная квартира стоит 1 992 000 рублей без учета погрешности. С учетом ошибки , равной 0.998%, бдет получен интервал от 1 972 080 до 2 011 920 рублей, внутри которого должна находится стоимость оцениваемой квартиры.

Линейная регрессия оценила данную квартиру в 1 976 000 рублей, а с учетом ошибки в 1.75% ее стоимость должна находится в диапазоне от 1 956 240 до 1995760 рублей.

Настоящая же цена, за которую продается квартира на момент просмотра данного обьявления составляет 1 950 000 рублей. Это означает что наиболее близкий результат был получен методом линейной регрессии за счет большей погрешности.

Калькулятор циан, на основе тех же параметров оценил данную квартиру в 2 000 000 рублей, с погрешностью в 7 процентов был дан интервал от 1 800 000 до 2 100 000 рублей.

Как результат эксперимента метод линейной регресси ближе всего подобрался к цене продажи квартиры, но при этом дальше всех оказался от цены который выдает калькулятор сайта «Циан» с большой базой данных. Алгоритм бустинга в своем резльтате с учетом погрешности так же предполагает стоимость квартиры в 2 000 000, но от изначальных 1 950 000 он дальше всех из алгоритмов. В случае случайного леса, то его результат оказался чуть ниже предсказаного цианом и выше цены предлагаемого продавцом, возможной причиной этого могла послужить плохая универсальность модели из-за небольшого количества тренировочных данных, по сравнению с калькулятором сайта «Циан». Как итог можно сказать, что наиболее предпочтительным алгоритмом на этом наборе данных является «случайный лес», за счет его универсальности и приемлемости резльтатов.

Заключение

В ходе написания данной работы были рассмотрены технологии машинного обучения для прогнозирования стоимости жилой недвижимомти города иваново при помощи стравнительно подхода. Чтобы проверить возможность использвоания данного инструмента для оценки жилой недвижимости и тем самым сократить время требующееся для проведения оценки.

В настоящий момент подбор аналогов оцениваемого обьекта и сам рассчет его стоимости производится специалистами компании вручную на основе принятых методик и использвания открытых источников информации. Такое решение является дорогостоящим и отнимающем много времени, но при этом явялющимся единственным возможным из-за отсутвия алтернатив. Поэтому было принято решение проверить возможность создания модели для прогнозирования стоимости жилой недвижимости, используя данные собранные из открытого источника, тем самым имитирую реальную деятельность.

На основе изученного и проанализированного теоретического материала были выбраны несколько моделей и обучены на наборе данных о предложениях по продаже недвижимости. Наилучший результат по значению ошибоки MAPE и R² показала модель случайного леса.

После проведения эксперимента с реальным обьектом недвижимости было дополнительно подтверждено что модель случайного леса является наиболее потимальной.

После анализа рельтатов был сделан вывод, что модели машинного обучения не могут выдавать нужную точность при прогнозировании стоимости жилой недвижимости. Это происходит из-за недостатка данных, увеличить объем которых практически невозможно из-за относительно небольшого размера рынка недвижимости города Иваново.

Список используемых источников

1. Рынок недвижимости [Электронный ресурс]. — URL: https://ru.wikipedia.org/wiki/Рынок\_недвижимости (дата обращения: 01.06.2020).
2. Оценка недвижимости [Электронный ресурс]. — URL: https://ru.wikipedia.org/wiki/Оценка\_недвижимости (дата обращения: 01.06.2020).
3. «Общие понятия оценки, подходы к оценке и требования к проведению оценки (ФСО № 1)» [Электронный ресурс]. — URL: http://www.ocenchik.ru/docsf/2229-ponyatiya-ocenki-podhody-ocenki-fso1.html (дата обращения: 01.06.2020).
4. «Оценка недвижимости (ФСО № 7)» [Электронный ресурс]. — URL: http://www.ocenchik.ru/docs/1961-standart-ocenki-nedvizhimosti-fso7-prikaz611.html (дата обращения: 01.06.2020).
5. RoadRunner: Towards Automatic Data Extraction from Large Web Sites/ сост.: Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo// VLDB '01 Proceedings of the 27th International Conference on Very Large Data Bases Pages 109-118
6. Коэффициент детерминации и линейная регрессия [Электронный ресурс]. — URL: http://statistica.ru/theory/koeffitsient-determinatsii-i-lineynaya-regressiya/ (дата обращения: 01.06.2020).
7. Регрессионная модель [Электронный ресурс]. — URL: http://www.machinelearning.ru/wiki/index.php?title=Регрессионная\_модель (дата обращения: 01.06.2020).
8. Линейная регрессия [Электронный ресурс]. — URL: https://ru.wikipedia.org/wiki/Линейная\_регрессия (дата обращения: 01.06.2020).
9. Случайные леса: обзор// С. П. Чистяков// Труды Карельского научного центра РАН № 1. 2013. С. 117-136
10. Bethany J Wolf 1, Elizabeth G Hill, Elizabeth H Slate. — Logic Forest: An Ensemble Classifier for Discovering Logical Combinations of Binary Markers// National Library of Medicine — URL: https://pubmed.ncbi.nlm.nih.gov/20628070/ (дата обращения: 01.06.2020).
11. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес [Электронный ресурс]. — URL: https://habr.com/ru/company/ods/blog/324402/ (дата обращения: 01.06.2020).
12. Introduction to Boosted Trees [Электронный ресурс]. — URL: https://xgboost.readthedocs.io/en/latest/tutorials/model.html (дата обращения: 01.06.2020).
13. GradientBoostingRegressor [Электронный ресурс]. — URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (дата обращения: 01.06.2020).
14. Implementing Gradient Boosting in Python [Электронный ресурс]. — URL: https://blog.paperspace.com/implementing-gradient-boosting-regression-python/ (дата обращения: 01.06.2020).
15. Understanding Random Forest [Электронный ресурс]. — URL: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (дата обращения: 01.06.2020).
16. A beginner’s guide to Linear Regression in Python with Scikit-Learn [Электронный ресурс]. — URL: https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f (дата обращения: 01.06.2020).
17. Как правильно выбрать метрику оценки [Электронный ресурс]. — URL: https://www.machinelearningmastery.ru/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0/ (дата обращения: 01.06.2020).
18. 2-комн. квартира [Электронный ресурс]. — URL: https://ivanovo.cian.ru/sale/flat/234203151/ (дата обращения: 01.06.2020).

ПРИЛОЖЕНИЕ А

Исходный код, анализирующий данные в датасете

import numpy as np

import matplotlib

import pandas as pd

import matplotlib.pyplot as plt

plt.style.use('ggplot')

pd.set\_option('display.max\_rows', 500)

pd.set\_option('display.max\_columns', 500)

pd.set\_option('display.width', 1000)

from subprocess import check\_output

data1 = pd.read\_csv('ivDom2.csv',encoding = "cp1251")

categorical\_columns = [c for c in data1.columns if data1[c].dtype.name == 'object']

numerical\_columns = [c for c in data1.columns if data1[c].dtype.name != 'object']

print("Категориальные признаки {}".format(categorical\_columns))

print("Количественные признаки {}".format(numerical\_columns))

print()

print()

data = pd.read\_csv('ivanovoDom.csv',encoding = "cp1251")

data\_labels = data['13']

print(data)

print()

print(data.corr())

print()

from feature\_selector import FeatureSelector

# Признаки - в train, метки - в train\_labels

fs = FeatureSelector(data = data, labels = data\_labels)

fs.identify\_missing(missing\_threshold=0.6)

print(fs.missing\_stats.head())

plt.show(fs.plot\_missing())

fs.identify\_collinear(correlation\_threshold=0.01)

correlated\_features = fs.ops['collinear']

correlated\_features[:29]

plt.show(fs.plot\_collinear())

# список признаков для удаления

collinear\_features = fs.ops['collinear']

# датафрейм коллинеарных признаков

print(fs.record\_collinear.head())

fs.identify\_zero\_importance(task = 'classification', eval\_metric = 'auc',

n\_iterations = 10, early\_stopping = True)

one\_hot\_features = fs.one\_hot\_features

base\_features = fs.base\_features

print('There are %d original features' % len(base\_features))

print('There are %d one-hot features' % len(one\_hot\_features))

print(fs.data\_all.head(10))

zero\_importance\_features = fs.ops['zero\_importance']

plt.show(fs.plot\_feature\_importances(threshold = 0.90, plot\_n = 7))

fs.identify\_single\_unique()

plt.show(fs.plot\_unique())

ПРИЛОЖЕНИЕ Б

Исходный код, применения методов машинного обучения

import xgboost as xgb

import pandas as pd

import seaborn as sns

import numpy as np

from geopy.distance import geodesic

from feature\_selector import FeatureSelector

import math

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_absolute\_error, r2\_score, median\_absolute\_error

from sklearn.metrics import mean\_squared\_error

from sklearn.model\_selection import train\_test\_split, GridSearchCV, RandomizedSearchCV

from sklearn.ensemble import RandomForestRegressor

from sklearn.preprocessing import LabelEncoder

import matplotlib.pyplot as plt

import warnings

warnings.filterwarnings('ignore')

#Вычисляет азимут по двум заданным точкам, код взят отсюда: https://pastebin.com/PHeWmiEN

def get\_azimuth(rad,city\_lat,city\_long, latitude, longitude):

    llat1 = city\_lat

    llong1 =city\_long

    llat2 = latitude

    llong2 = longitude

    lat1 = llat1\*math.pi/180.

    lat2 = llat2\*math.pi/180.

    long1 = llong1\*math.pi/180.

    long2 = llong2\*math.pi/180.

    cl1 = math.cos(lat1)

    cl2 = math.cos(lat2)

    sl1 = math.sin(lat1)

    sl2 = math.sin(lat2)

    delta = long2 - long1

    cdelta = math.cos(delta)

    sdelta = math.sin(delta)

    y = math.sqrt(math.pow(cl2\*sdelta,2)+math.pow(cl1\*sl2-sl1\*cl2\*cdelta,2))

    x = sl1\*sl2+cl1\*cl2\*cdelta

    ad = math.atan2(y,x)

    x = (cl1\*sl2) - (sl1\*cl2\*cdelta)

    y = sdelta\*cl2

    z = math.degrees(math.atan(-y/x))

    if (x < 0):

        z = z+180.

    z2 = (z+180.) % 360. - 180.

    z2 = - math.radians(z2)

    anglerad2 = z2 - ((2\*math.pi)\*math.floor((z2/(2\*math.pi))) )

    angledeg = (anglerad2\*180.)/math.pi

    return round(angledeg, 2)

#Вычисляет среднюю абсолютную процентную ошибку

def mean\_absolute\_percentage\_error(y\_true, y\_pred):

    y\_true, y\_pred = np.array(y\_true), np.array(y\_pred)

    return np.mean(np.abs((y\_true - y\_pred) / y\_true)) \* 100

#Вычисляет медианную абсолютную процентную ошибку

def median\_absolute\_percentage\_error(y\_true, y\_pred):

    y\_true, y\_pred = np.array(y\_true), np.array(y\_pred)

    return np.median(np.abs((y\_true - y\_pred) / y\_true)) \* 100

#Печатает рассчитанные значения коэффициента детерминации, средней и медианной абсолютных ошибок

def print\_metrics(prediction, val\_y):

    val\_mae = mean\_absolute\_error(val\_y, prediction)

    median\_AE = median\_absolute\_error(val\_y, prediction)

    rmse = math.sqrt(mean\_squared\_error(val\_y, prediction))

    r2 = r2\_score(val\_y, prediction)

    print('')

    print('R\u00b2: {:.2}'.format(r2))

    print('RMSE: {:.2}'.format(rmse))

    print('Средняя абсолютная ошибка: {:.3} %'.format(mean\_absolute\_percentage\_error(val\_y, prediction)))

    print('Медианная абсолютная ошибка: {:.3} %'.format(median\_absolute\_percentage\_error(val\_y, prediction)))

#При помощи библиотеки pandas считываем csv-файл и преобразуем его в формат датафрейма (таблицы)

file\_path = 'C:/Users/ilyaa/Desktop/dat2.csv'

df\_I = pd.read\_csv(file\_path)

#Выводим 5 первых строк датафрейма

print(df\_I.head(5))

#Создаем новый столбец Стоимость 1 кв.м путем построчного деления стоимостей квартир на их общие площади

df\_I['priceMetr'] = df\_I['price']/df\_I['totalArea']

#Задаем широту и долготу центра города и рассчитываем для каждой квартиры расстояние от центра и азимут

Ivanovo\_center\_coordinates = [57.000348, 40.973921]

df\_I['distance'] = list(map(lambda x, y: geodesic(Ivanovo\_center\_coordinates, [x, y]).meters, df\_I['latitude'], df\_I['longitude']))

df\_I['azimuth'] = list(map(lambda x, y: get\_azimuth(6372795,Ivanovo\_center\_coordinates[0],Ivanovo\_center\_coordinates[1],x, y), df\_I['latitude'], df\_I['longitude']))

#Выбираем из датафрейма только те квартиры, которые расположены не дальше 40 км от центра города с панельными стенами

df\_I = df\_I.loc[(df\_I['distance'] < 40000)]

#Округляем значения стоблцов Стоимости метра, расстояния и азимута

df\_I['priceMetr'] = df\_I['priceMetr'].round(0)

df\_I['distance'] = df\_I['distance'].round(0)

df\_I['azimuth'] = df\_I['azimuth'].round(0)

#Выводим сводную информацию о датафрейме и его столбцах (признаках)

df\_I.info()

mediane = df\_I['priceMetr'].quantile(q=0.5)

print(f'Медиана: {mediane}')

#Вычисляем строки со значениями-выбросами

first\_quartile = df\_I['priceMetr'].quantile(q=0.4)

print(f'Нижняя граница: {first\_quartile}')

third\_quartile = df\_I['priceMetr'].quantile(q=0.6)

print(f'Верхняя граница: {third\_quartile}')

outliers = df\_I[(df\_I['priceMetr'] > third\_quartile) | (df\_I['priceMetr'] < first\_quartile )].count(axis=1)

# Удаляем из датафрейма 3000 строк, подходящих под критерии выбросов

print(f'Кол-во выбросов: {outliers.count}')

df\_I.drop(outliers.index, inplace=True)

df\_I.drop(['latitude'], axis=1,inplace = True)

df\_I.drop(['longitude'], axis=1,inplace = True)

df\_I.drop(['price'], axis=1,inplace = True)

#Выводим сводную информацию о датафрейме и его столбцах (признаках)

df\_I.info()

cor = df\_I.corr()

sns.heatmap(cor, annot=True, cmap=plt.cm.Reds)

plt.show()

#Назначаем целевой переменной цену 1 кв. метра, а можно и цену всей квартиры, тогда будет y = df['price']

y\_I = df\_I['priceMetr']

#Создаем список признаков, на основании которых будем строить модели

features = [

            'floorNumber',

            'floorsTotal',

            'totalArea',

            'distance',

            'azimuth',

           ]

#Создаем датафрейм, состоящий из признаков, выбранных ранее

X\_I = df\_I[features]

#Проводим случайное разбиение данных на выборки для обучения (train) и валидации (val), по умолчанию в пропорции 0.75/0.25

train\_X, val\_X, train\_y, val\_y = train\_test\_split(X\_I, y\_I, random\_state=1)

 #Создаем регрессионную модель случайного леса

# Создаем модели

rf\_model = RandomForestRegressor(n\_estimators=2000,

                                 n\_jobs=-1,

                                 bootstrap=False,

                                 criterion='mae',

                                 max\_features=2,

                                 random\_state=1,

                                 max\_depth=50,

                                 min\_samples\_split=5

                                 )

xgb\_model = xgb.XGBRegressor(objective ='reg:gamma',

                             learning\_rate = 0.01,

                             max\_depth = 50,

                             n\_estimators = 2000,

                             nthread = -1,

                             eval\_metric = 'gamma-nloglik',

                             )

reg\_model = LinearRegression()

#Проводим подгонку модели на обучающей выборке

rf\_model.fit(train\_X, train\_y)

xgb\_model.fit(train\_X, train\_y)

reg\_model.fit(train\_X, train\_y)

#Вычисляем предсказанные значения цен на основе валидационной выборки

rf\_prediction = rf\_model.predict(val\_X).round(0)

xgb\_prediction = xgb\_model.predict(val\_X).round(0)

reg\_prediction = reg\_model.predict(val\_X).round(0)

#Вычисляем и печатаем величины ошибок при сравнении известных цен квартир из валидационной выборки с предсказанными моделью

print("Величины ошибок для случайного леса")

print\_metrics(rf\_prediction, val\_y)

print()

print("Величины ошибок для бустинга")

print\_metrics(xgb\_prediction, val\_y)

print()

print("Величины ошибок для линейной регрессии")

print\_metrics(reg\_prediction, val\_y)

print()

# #Усредняем предсказания обоих моделей

# prediction = rf\_prediction \* 0.5 + xgb\_prediction \* 0.5

# #Вычисляем и печатаем величины ошибок для усредненного предсказания

# print("Величины ошибок для усредненного предсказания")

# print\_metrics(prediction, val\_y)

#Рассчитываем важность признаков в модели Random forest

importances = rf\_model.feature\_importances\_

std = np.std([tree.feature\_importances\_ for tree in rf\_model.estimators\_],

             axis=0)

indices = np.argsort(importances)[::-1]

#Печатаем рейтинг признаков

print()

print("Рейтинг важности признаков:")

for f in range(X\_I.shape[1]):

    print("%d. %s (%f)" % (f + 1, features[indices[f]], importances[indices[f]]))

df\_0 = pd.DataFrame({'Actual': val\_y, 'Predicted': rf\_prediction})

df1 = df\_0

df1.plot(kind='bar', figsize=(10, 8))

plt.title('Случайный лес')

plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')

plt.show()

df\_0 = pd.DataFrame({'Actual': val\_y, 'Predicted': xgb\_prediction})

df1 = df\_0

df1.plot(kind='bar', figsize=(10, 8))

plt.title('Бустинг')

plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')

plt.show()

df\_0 = pd.DataFrame({'Actual': val\_y, 'Predicted': reg\_prediction})

df1 = df\_0

df1.plot(kind='bar', figsize=(10, 8))

plt.title('Линейная регрессия')

plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')

plt.show()

#Строим столбчатую диаграмму важности признаков

plt.figure()

plt.title("Важность признаков")

plt.bar(range(X\_I.shape[1]), importances[indices], color="g", yerr=std[indices], align="center")

plt.xticks(range(X\_I.shape[1]), indices)

plt.xlim([-1, X\_I.shape[1]])

plt.show()

#Создаем датафрейм с параметрами квартиры https://ivanovo.cian.ru/sale/flat/225753580/,

flat = pd.DataFrame({

                     'floorNumber':[3],

                     'floorsTotal':[5],

                     'totalArea':[47.6],

                     'latitude':[56.9704019],

                     'longitude':[40.9945961],

                  })

# #Вычисляем предсказанное значение стоимости по двум моделям

rf\_prediction\_flat = rf\_model.predict(flat).round(0)

xgb\_prediction\_flat = xgb\_model.predict(flat).round(0)

reg\_prediction\_flat = reg\_model.predict(flat).round(0)

rf\_price\_flat = rf\_prediction\_flat\*flat['totalArea'][0]

xgb\_price\_flat = xgb\_prediction\_flat\*flat['totalArea'][0]

reg\_price\_flat = reg\_prediction\_flat\*flat['totalArea'][0]

mean\_price\_flat = (rf\_prediction\_flat \* 0.3 + xgb\_prediction\_flat \* 0.3+ reg\_prediction\_flat \* 0.3)\*flat['totalArea'][0]

print(f'Предсказанная случайным лесом цена предложения: {int(rf\_price\_flat.round(-3))} рублей')

print(f'Предсказанная бустингом цена предложения: {int(xgb\_price\_flat.round(-3))} рублей')

print(f'Предсказанная регерссией цена предложения: {int(reg\_price\_flat.round(-3))} рублей')

print(f'Усредненная цена предложения: {int(mean\_price\_flat.round(-3))} рублей')

ПРИЛОЖЕНИЕ В

Описание содержания прилагаемого диска:

* Пояснительная записка.docx
* Презентация «Защита дипломного проекта».pptx
* Скрипт анализа данных.py
* Скрипт обучения моделей.py
* Исходные данные.xlsx
* Данные используемы при обучении.csv