|  |  |
| --- | --- |
| Институт (факультет) | Информационных технологий |
| Направление подготовки (специальность) | 01.03.02 Прикладная математика и информатика  (Прикладная математика и информатика) |
| Выпускающая кафедра | Математики и информатики |

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

|  |  |
| --- | --- |
| Название работы | Построение модели массовой оценки недвижимости в городе Москве методами машинного обучения |

|  |  |
| --- | --- |
| Студента | Ермолиной Валерии Сергеевны |
|  | Ф.И.О. |

ДОПУСТИТЬ К ЗАЩИТЕ

|  |  |  |  |
| --- | --- | --- | --- |
| Директор института (декан факультета) |  |  | Ершов Е.В. |
| Заведующий выпускающей кафедрой |  |  | Лягинова О.Ю. |
| Руководитель выпускной квалификационной работы |  |  | Плотникова Н.В. |
| Нормоконтролер |  |  | Грибкова Ю.В. |
| Выпускник |  |  | Ермолина В.С. |

Череповец

2020 г.

Оглавление

[Введение 2](#_Toc40120466)

[Глава 1. Теоретические аспекты оценки стоимости объектов недвижимости 4](#_Toc40120467)

[1.1 Понятие и виды недвижимого имущества 4](#_Toc40120468)

[1.2 Виды и методы расчёта оценочной стоимости объектов недвижимости 8](#_Toc40120469)

[1.3 Основные факторы влияющие на формирование стоимости объектов недвижимости 12](#_Toc40120470)

[1.4 Постановка задачи работы 16](#_Toc40120471)

[Глава 2. Математические основы теории машинного обучения 18](#_Toc40120472)

[2.1 Определение и категории машинного обучения 18](#_Toc40120473)

[2.1.1 Обучение с учителем 20](#_Toc40120474)

[2.1.2 Обучение без учителя 21](#_Toc40120475)

[2.2 Основные типы задач машинного обучения 22](#_Toc40120476)

[2.2.1 Задача классификации 22](#_Toc40120477)

[2.2.2 Задача кластеризации 24](#_Toc40120478)

[2.2.3 Задача регрессии 26](#_Toc40120479)

[2.3 Основные алгоритмы регрессионного анализа 28](#_Toc40120480)

[2.3.1 Линейная регрессия 28](#_Toc40120481)

[2.3.2 Метод опорных векторов 31](#_Toc40120482)

[2.3.3 Регрессионные деревья 32](#_Toc40120483)

[Глава 3. Построение модели массовой оценки недвижимости в городе Москве методами машинного обучения 38](#_Toc40120484)

[3.1 Выбор инструментов реализации для построения модели 38](#_Toc40120485)

[3.2 Сбор и обработка данных 42](#_Toc40120486)

[3.3 Описание построения модели массовой оценки недвижимости в городе Москве 50](#_Toc40120487)

[3.4 Результат работы модели 57](#_Toc40120488)

[Заключение 62](#_Toc40120489)

# **Введение**

Оценка недвижимости для банка является обязательным этапом кредитования под залог такого имущества, как дома, квартиры, земельные участки, загородные дачи, гаражи, офисы, незавершенное долевое строительство и другие объекты, которые невозможно переместить без ущерба для них.

Банки принимают оцененные объекты в залог при ипотечном кредитовании, кредитовании для развития бизнеса, покупки товаров, жилой и коммерческой недвижимости и любых других крупных проектов залогодателя. Под залог принимается как приобретённое ранее, так и приобретаемое за счет заёмных средств жильё на вторичном рынке, новая или строящаяся недвижимость. Кроме того, недвижимость оказывает влияние на платёжеспособность, а, следовательно, и на итоговый скоринговый балл заёмщика при потребительском кредитовании без обеспечения. В любом случае, только после получения отчёта об оценке банк может заключить кредитный договор.

Большое количество характеристик жилья усложняет процесс оценки. В этой ситуации возникает необходимость в разработке модели, использующей широкий спектр функций для построения массовой оценки недвижимости.

Тема оценки недвижимости актуальна по причине развития рынка жилья и, как следствие, увеличения необходимости компетентной и объективной оценки стоимости объектов недвижимости.

Научная новизна дипломной работы обусловлена современным подходом решения данной проблемы. Предполагается, что математическими методами с помощью машинного обучения можно найти решение задачи массовой оценки недвижимости, опираясь на данные об объектах, представленные интернет-сервисами для размещения объявлений.

Целью выпускной квалификационной работы является построение модели массовой оценки недвижимости в городе Москве методами машинного обучения.

Объект исследования – методы машинного обучения, используемые для анализа данных.

Предмет исследования – применение машинного обучения для построения модели массовой оценки недвижимости в городе Москве.

В соответствии с целью работы определены следующие задачи:

1. Рассмотреть основные понятия, касающиеся проблемы исследования: понятия и виды недвижимого имущества, виды и методы расчёта оценочной стоимости объектов недвижимости, факторы, влияющие на формирование стоимости объектов недвижимости.
2. Сформулировать ряд теоретических положений машинного обучения, необходимых для построения модели.
3. Получить данные об объектах недвижимости с использованием интернет-сервиса размещения объявлений и обработать их.
4. Опираясь на теоретический раздел о машинном обучении, выбрать методы и инструменты для построения модели.
5. Построить модель массовой оценки недвижимости в городе Москве с использованием выбранных методов машинного обучения, учитывая особенности региона.

Структурно работа состоит из трёх глав. В первой главе рассматриваются теоретические основы оценки недвижимости, во второй главе проводится анализ методов машинного обучения, в третьей – осуществляется построение модели массовой оценки недвижимости выбранным методом машинного обучения.

В заключении систематизируются полученные результаты и формулируются выводы в соответствии с предполагаемой гипотезой.

В списке литературы указаны все используемые для написания данной работы источники.

Данные для анализа и программный код представлены в приложении к выпускной квалификационной работе.

# **Глава 1. Теоретические аспекты оценки стоимости объектов недвижимости**

* 1. **Понятие и виды недвижимого имущества**

Понятие недвижимого имущества тесно связано с такими правовыми категориями, как вещь и имущество. Под вещью понимается предмет внешнего (материального) мира, находящийся в естественном состоянии в природе или созданный трудом человека. Также вещь воспринимают не только как физическое тело с его пространственными границами и физическими параметрами, но и как объект права, как материальное благо, способное участвовать в гражданском (имущественном) обороте, обладать потребительской и меновой стоимостью, передаваться по наследству, удовлетворять положительные потребности человека, иными словами, как составную часть юридического понятия имущество [1, с. 87].

Деление вещей на движимые и недвижимые основано на естественных свойствах объектов гражданских прав. Как правило, недвижимые вещи постоянно находятся в одном и том же месте, обладают индивидуальными признаками и являются незаменимыми. Напротив, движимые вещи могут свободно перемещаться вместе с теми лицами, которым они принадлежат, могут быть индивидуально-определёнными или родовыми и, как правило, являются заменимыми. Недвижимые вещи являются индивидуально-определёнными по своей природе [2, с. 225]. Важность данного деления определяется существенными различиями, которые имеются в правовом режиме движимых и недвижимых вещей. Эти различия отражены в нормах Гражданского кодекса Российской Федерации (далее ГК РФ). Указанное деление в строгом смысле распространяется только на вещи и неприменимо к другим видам имущества.

Согласно ст. 130 ГК РФ существует три вида недвижимых вещей. Первую группу составляют вещи, являющиеся недвижимыми по своей природе. К ним закон относит земельные участки и участки недр [3].

Вторую группу недвижимых вещей образуют все вещи, прочно связанные с землёй, т. е. объекты, перемещение которых без несоразмерного ущерба их назначению невозможно, в том числе здания, сооружения, жилые помещения, объекты незавершённого строительства. Приведённый в законе перечень объектов такого рода носит сугубо примерный характер.

Отнесение многих конкретных объектов к недвижимости по признаку прочной связи с землёй весьма проблематично. В частности, современные техника и технологии позволяют перемещать здания и сооружения. Поэтому в данном случае закон следует толковать не буквально, а в соответствии с его смыслом, который состоит в том, что недвижимостью закон признаёт всё то, что может использоваться по своему назначению только в неразрывной связи с землёй [3, с. 129].

К третьей группе недвижимых вещей закон относит объекты, которые признаны недвижимостью не в силу их естественных свойств, а по иным причинам. В частности, недвижимостью считаются подлежащие государственной регистрации воздушные и морские суда, суда внутреннего плавания и космические объекты (искусственные спутники, космические корабли, орбитальные станции и т. д.). Признание их недвижимым имуществом обусловлено высокой стоимостью данных объектов и связанной с этим необходимостью повышенной надёжности правил их гражданского оборота.

Самая востребованная группа объектов недвижимости сегодня – жилые помещения, в частности – квартиры.

Согласно Жилищному кодексу Российской Федерации, жилым помещением признаётся изолированное помещение, которое является недвижимым имуществом и пригодно для постоянного проживания граждан (отвечает установленным санитарным и техническим правилам и нормам, иным требованиям законодательства).

Квартира – это жилое помещение, структурно обособленное в многоквартирном доме, обеспечивающее возможность прямого доступа к помещениям общего пользования в таком доме и состоящее из одной или нескольких комнат, а также помещений вспомогательного использования, предназначенных для удовлетворения гражданами бытовых и иных нужд, связанных с их проживанием в таком обособленном помещении [2, с. 227].

На основании ст. 130, 131 ГК РФ, можно сделать вывод о том, что российский законодатель выделяет следующие признаки отнесения к недвижимому имуществу:

1. естественное свойство вещи (земельные участки, участки недр);
2. прочную связь с земельным участком (здания, строения, сооружения, жилые помещения, имущественные комплексы, прочно связанные с землей, перемещение которых невозможно без несоразмерного ущерба их назначению);
3. нормативно-формальную потребность распространения правового режима недвижимого имущества на некоторые виды движимых вещей с целью укрепления имущественного оборота и защиты имущественных прав собственников (воздушных, морских, речных судов и космических объектов).

Правовой анализ современного гражданского законодательства позволяет заключить, что отличительными признаками недвижимого имущества являются:

1. обязательная государственная регистрация прав на недвижимое имущество и сделок с ним;
2. исполнение обязательств по поводу недвижимого имущества по месту нахождения имущества;
3. действие особых правил совершения сделок в отношении земельных участков и расположенных на них строений;
4. особый порядок приобретения права собственности на недвижимые вещи в силу давностного владения;
5. особый порядок распоряжения недвижимым имуществом, принадлежащим государственным и муниципальным предприятиям;
6. особый порядок обращения взыскания на заложенное недвижимое имущество;
7. наличие такого способа обеспечения исполнения обязательства в отношении недвижимого имущества как ипотека;
8. наследование недвижимых вещей и их правовой режим определяются по нормам права, действующим в месте их нахождения;
9. рассмотрение споров о праве собственности и иных вещных правах на недвижимые вещи происходит по месту нахождения недвижимых вещей.

При этом государственная регистрация прав на недвижимое имущество и сделок с ним является не существенным признаком объекта недвижимости, а требованием закона по укреплению правового режима недвижимого имущества, правовым последствием признания недвижимой вещи объектом права.

Государственная регистрация – это юридический акт признания и подтверждения государством возникновения, ограничения (обременения), перехода или прекращения прав на недвижимое имущество, являющийся доказательством существования зарегистрированного права [1, с. 105].

На основании вышеизложенного, можно сделать вывод о том, что недвижимое имущество – это юридическое понятие, включающее в себя недвижимые вещи в силу их естественных свойств (земельные участки, участки недр), а также вещи, прочно связанные с землёй (здания, строения, сооружения, жилые, нежилые помещения, имущественные комплексы). Кроме того, в состав недвижимого имущества отечественным законодательством включены социально-значимые вещи, признанные недвижимыми не в силу природных (естественных) свойств этих вещей, а в силу закона. Указанные виды недвижимого имущества способны быть объектом гражданского права, принадлежать субъектам права, удовлетворять их законные потребности.

* 1. **Виды и методы расчёта оценочной стоимости объектов недвижимости**

В настоящее время для большинства операций, связанных с недвижимостью, таких, как продажа, обмен, покупка, дарение, завещание, оформление ипотечного кредита или кредита под залог недвижимости, требуется знать её актуальную рыночную стоимость.

Оценкой стоимости недвижимости называют процесс определения стоимости конкретного объекта недвижимости, например, квартиры, частного дома или дачи. Оценкой недвижимости занимаются эксперты-оценщики, деятельность которых регулирует закон «Об оценочной деятельности в Российской Федерации». Большая часть процесса оценки осуществляется экспертами вручную, что отрицательно сказывается на сроках выполнения работ. Рассчитать ориентированную рыночную стоимость объекта можно также и самостоятельно, но результаты такой оценки не примут во внимание государственные органы. Суд, Росреестр, банки принимают во внимание только официальный документ – отчёт об оценке, выданный аккредитованным специалистом [4, с. 120].

Оценка квартиры – самый популярный вариант экспертизы в сфере оценки недвижимости. На сегодняшний день самая распространённая причина провести оценку квартиры – кредитование.

Согласно Федеральному Стандарту Оценки, при осуществлении оценочной деятельности используются четыре вида стоимости объекта оценки: рыночная, ликвидационная, инвестиционная и кадастровая.

Рыночная стоимость объекта недвижимости – это цена, за которую его можно продать в условиях свободного рынка. Она используется, например, покупателями и продавцами при обсуждении стоимости недвижимости, либо юридическими лицами в качестве доказательства наличия определённых активов. Рыночную стоимость определяют, в зависимости от условий рынка: находят аналогичные объекты и рассчитывают среднерыночную цену, которая и будет результатом оценки.

Ликвидационная стоимость – это цена, за которую объект может быть продан на рынке за ограниченные сроки. Эта стоимость объекта недвижимости всегда меньше рыночной, и при её определении учитывают влияние чрезвычайных обстоятельств, которые вынудили собственника продавать объект на условиях, не соответствующих рыночным.

Инвестиционная стоимость – это стоимость объекта в конкретном инвестиционном проекте для конкретного инвестора с учётом его инвестиционных целей. Эта стоимость чаще всего применяется для оценки эффективности проекта или для раздела долей в инвестиционных проектах, когда собственник объекта и инвестор хотят урегулировать свои отношения на основании справедливого раздела будущих инвестиционных доходов.

Инвестиционная стоимость должна быть всегда больше рыночной, иначе собственнику объекта нет никакого смысла участвовать в данном проекте, ему выгоднее продать свой объект по рыночной цене. Если инвестиционная стоимость оказалась меньше рыночной, то выбран неэффективный инвестиционный план.

Кадастровая стоимость – это рыночная стоимость недвижимости, которая определяется в массовом порядке по заказу государственных или муниципальных органов для объектов, состоящих на учёте в государственном кадастре недвижимости.

Реже используют другие типы стоимости при проведении оценки, например, восстановительную.

Восстановительная стоимость объекта недвижимости – это стоимость строительства такого же объекта в текущих условиях. При её расчёте учитывается зарплата рабочих, стоимость строительных материалов, оплата проектных материалов, а также другие прямые и косвенные затраты [4, c. 145-147].

По количеству оцениваемых объектов различают два вида оценки недвижимости: индивидуальную и массовую.

Индивидуальная оценка – оценка конкретного объекта недвижимости на конкретную дату [5, c. 150].

Массовая оценка – это систематическая оценка групп объектов недвижимости по состоянию на определенную дату, с использованием стандартных процедур и статистического анализа [5, c. 151].

Обе методики оценки имеют в своей основе систематические методы сбора, анализа и обработки информации для получения обоснованных результатов. В процессе массовой оценки, осуществляются те же действия, что и при индивидуальной оценке, отличия в основном связаны с разницей в масштабе проведения работ, и методах контроля качества оценки.

Так как модель оценки рассматривается как математическая интерпретация поведения рынка недвижимости в определённое время, при массовой оценке модели сложнее, поскольку они пытаются воспроизвести деятельность рынка на обширной географической территории.

Существует множество методов индивидуальной и массовой оценки недвижимости. Однако не все из них используются одинаково часто.  Выбор метода всегда зависит от конкретной ситуации: вида оцениваемого объекта недвижимости, цели оценки, от имеющейся в распоряжении оценщика информации и ряда других факторов.

Согласно Федеральному Стандарту Оценки, оценщики могут использовать три метода расчёта: сравнительный, доходный и затратный.

Основной тезис сравнительного метода в заключается том, что покупатель не заплатит за недвижимость цену большую той, по которой можно приобрести точно такой же объект. При применении этого метода оценщик должен отобрать практически аналогичную недвижимость, провести сравнительный анализ отобранных и оцениваемого объекта недвижимости, скорректировать цены и найти средний показатель. B зависимости от целей анализа, может быть использован разный период времени, в течение которого продавались объекты. Этот метод требует изучения большого объёма информации и расчётов, но позволяет с высокой точностью определить стоимость.

Суть доходного метода заключается в определении стоимости недвижимости на момент оценки как источника будущих доходов. То есть оценщик определяет потенциальную прибыль объекта c момента оценки до завершения эксплуатации. При применении метода учитываются риски, характерные для имущества региона. Данный метод используют при оценке рыночной ценности, a также в инвестиционной деятельности.

Главный тезис затратного метода состоит в том, что объект недвижимости не может стоить больше, чем сумма, которую придётся потратить на его строительство. Этот метод используют, когда нужно оценить объекты недвижимости вместе с земельными участками, на которых они расположены. При применении затратного метода эксперт оценивает рыночную стоимость земельного участка, определяет восстановительную стоимость здания, в том числе размер предпринимательской прибыли, оценивает выявленные виды износа и рассчитывает итоговую стоимость объекта (для этого корректирует восстановительную стоимость и дoбaвляeт стоимость земельного участка) [5, c. 72-75].

Неразвитость рынка накладывает ограничения на точность индивидуальной и массовой оценок. Так, в ситуации неразвитого рынка приходится делать сильные допущения, которые, если оказываются ошибочными, увеличивают погрешность результата оценки. Таким образом, развитость рынка недвижимости обусловливает точность и надёжность как индивидуальной, так и массовой оценки, при использовании каждого из трёх методов. Если рынок развит, а информация открыта и не искажена, то шансов на аккуратную и точную оценку больше, вне зависимости от применяемых методов. При этом выбор метода (затратный, доходный или сравнения), осуществляемый оценщиком, зависит от типа недвижимости и цели оценки, как при индивидуальной оценке, так и при массовой.

* 1. **Основные факторы влияющие на формирование стоимости объектов недвижимости**

Изменение стоимости любого объекта недвижимости зависит от целого ряда факторов. В процессе оценки недвижимости необходимо установить влияние каждого фактора на итоговую цену, а затем влияние всех факторов в совокупности и сделать общее заключение о стоимости объекта [6, с. 345].

В данном параграфе систематизированы основные факторы, влияющие на стоимость недвижимости.

Таким образом, все факторы можно разделить на две группы: социально-экономические и политико-психологические.

Среди социально-экономических факторов выделяют:

1. Рост и уровень доходов населения.

Теоретически рост доходов населения оказывает влияние на уровень цен. Однако, сопоставляя темпы роста цен и доходов, можно констатировать, что рост доходов, как правило, отстаёт от роста цен.

Уровень доходов – это показатель, относящийся ко всем жителям региона. Потребители такого товара, как недвижимость, – это, в основном, люди с наиболее высокими доходами.

1. Состояние экономики.

Определить, каким образом общее состояние экономики влияет на стоимость недвижимости достаточно сложно. Резкий спад в экономике неизбежно приведёт к снижению цен. Если же спад небольшой, то зависимость между ВВП и ценами на недвижимость может быть и прямо противоположной. При угрозе кризисных ситуаций люди пытаются обрести стабильность, приобретая недвижимость. Спрос при этом растёт, а, следовательно, растут и цены. В случае уверенного роста экономики рядом с потоком инвестиций в недвижимость возникают другие потоки, требующие значительные средства и ресурсы. Интерес к рынку недвижимости при этом снижается, что сказывается и на темпе роста цен.

1. Трудовая и культурная миграция.

Большое количество населения устремляется в крупные города. Среди причин для переселения выделяют наличие рабочих мест, возможности для карьерного роста, развитость науки, культуры, образования. Значительная часть этих людей либо обладают необходимыми для покупки квартир средствами, либо готовы покупать их с помощью кредитов. В этой ситуации наблюдается так же рост цен на недвижимость.

1. Развитие кредитования. Внедрение государственных программ предоставления льготных кредитов.

Всё большее число покупателей пользуются кредитами. Появление дополнительных денег и покупателей повышает спрос на недвижимость и, соответственно, цены.

Понижение ставки по кредитам привлекает значительное число покупателей. Государственные или муниципальные программы льготного кредитования объективно способствуют увеличению платёжеспособного спроса и удорожанию жилья. Отмена же этих программ и ужесточение условий кредитования приводят к снижению спроса и стагнации рынка недвижимости.

1. Недостаток альтернативных областей выгодного и надёжного размещения свободных средств. Недоверие к валютам.

Состояние дел с альтернативными способами сохранения и приумножения накоплений сформировало мнение, что недвижимость – наилучшее место для таких целей. В связи с этим, основные накопления перетекают на рынок жилой недвижимости, риск потерять деньги на котором существенно ниже, рост цен в долгосрочной перспективе весьма вероятен, кроме того, недвижимость можно использовать для улучшения условий жизни, или сдавать в аренду.

Устремление потоков инвестиционно-спекулятивного капитала на рынок недвижимости приводит к росту спроса, а, следовательно, и к росту цен.

1. Изменение статуса микрорайона.

Практически любые изменения (как положительные, так и отрицательные) транспортной доступности и инфраструктуры микрорайона, состояния экологии, благоустройства, озеленения, общественной безопасности, социальной и культурной среды влияют на уровень цен. Строительство современных торговых и развлекательных центров, открытие престижных школ и спортивных и детских учреждений – всё это способствует повышению стоимости недвижимости в микрорайоне. Если люди стремятся в данный микрорайон, то цены в нём будут выше, чем в соседних. Деградация коммунального хозяйства и отдельных домов, разрушение инфраструктуры и социальной сферы, ухудшение экологии, рост преступности в микрорайоне – всё это сказывается на соотношении спроса и предложения, а, следовательно, и на ценах. Если жители стремятся уехать из микрорайона, то цены в нём будут ниже, чем в соседних.

1. Низкие объёмы строительства. Ориентация на строительство многокомнатных просторных квартир.

При недостаточном объёме строительства и фактическом игнорировании строительства малогабаритного жилья постоянно растущий спрос остаётся неудовлетворённым. Не найдя таких квартир в новых домах, люди вынуждены покупать устаревшее жильё, цены на которое растут быстрее всего.

Высокий уровень цен является тем фактором, который сам по себе заставляет потенциальных покупателей отложить покупку жилья и сосредоточиться на накоплении денежных средств. Покупки откладываются, а на рынке начинается стагнация или падение цен [6, с. 340 - 342].

Основными политико-психологическими факторами, которые наиболее сильно влияют на стоимость недвижимости являются:

1. Состояние политической стабильности и, как следствие этого, положительные или негативные ожидания участников рынка.
2. Степень доверия граждан к банковской системе, рынку недвижимости к перспективам развития города и страны.
3. Уровень оптимизма граждан, зависящий от многих факторов, начиная от перспектив развития страны и заканчивая состоянием жилого фонда и криминогенной обстановкой в микрорайоне.

В каждой конкретной ситуации всегда можно выделить превалирующие факторы. В большинстве случаях, действуют они сообща, усиливая или ослабляя друг друга [6, с. 343 – 344].

Подводя итоги, отметим основные факторы, которые могут снизить цены на недвижимость.

Положительные факторы:

1. резкое увеличение объёма строительства: ориентация на строительство целых микрорайонов со своей инфраструктурой, привлечение крупных иностранных строительных компаний;
2. развитие регионов, создание благоприятных условий для жизни и работы как в городах, так и в сельской местности;
3. создание альтернативных способов размещения денежных средств.

Отрицательные факторы:

1. кризис в экономике и, как следствие, резкое падение доходов населения и сокращение платёжеспособного спроса на недвижимость;
2. форс-мажорные обстоятельства (стихийные бедствия, национальные конфликты);
3. деградация микрорайонов, коммунального хозяйства, отдельных домов;
4. ухудшение условий ипотечного кредитования, сворачивание льготных программ приобретения жилья.

В заключение следует отметить, что сложность анализа регионального рынка недвижимости обусловлена именно необходимостью правильного выбора не только факторов, влияющих на стоимость объектов, но и дополнительных нюансов, присущих анализируемому региону.

Таким образом, на основании предложенных факторов представляется возможным построение достоверной математической модели массовой оценки объектов недвижимости.

* 1. **Постановка задачи работы**

Главной проблемой в рассматриваемой теме выпускной квалификационной работы является то, что существует перечень услуг, предоставляемых банком, обязательным условием для оформления которых является определение рыночной стоимости объектов недвижимости:

* 1. Ипотечное кредитование.

Ипотека или ипотечный кредит – это целевой кредит, предоставляемый для приобретения недвижимости под залог этой или другой недвижимости. Именно такие кредиты составляют сегодня основную массу выдаваемых банками кредитов, обеспеченных залогом недвижимости [5, c. 154].

Банк при формировании суммы займа учитывает рыночную и ликвидационную стоимости объекта недвижимости и принимает во внимание меньшую, обычно ликвидационную стоимость. Сумма ипотеки, как правило, равна ликвидационной стоимости жилья или меньше её, потому что определённую часть средств вносит заёмщик. Также оценка стоимости необходима для расчёта суммы, которую банк сможет выручить, если заёмщик перестанет исполнять свои обязательства, то есть платить по ипотеке [6, c.342].

* 1. Кредитование под залог недвижимости.

Кредит под залог недвижимости – это кредит под залог имеющегося объекта недвижимости на различные цели (развитие бизнеса, приобретение дорогостоящих покупок, рефинансированиедолгов, потребительские нужды и т.д.), но не на покупку жилья [5, c. 155].

В случае кредитования под залог, как и в случае c ипотекой, оценка стоимости необходима для формирования суммы займа и для расчёта суммы, которую банк сможет выручить, если заёмщик не будет платить по кредиту, в расчёт принимается меньшая стоимость, то есть ликвидационная [6, c. 343].

* 1. Потребительское кредитование без обеспечения.

Потребительский кредит без обеспечения – это услуга выдачи заёмных средств, которая не требует предоставления дополнительных гарантий, таких как оформление поручительства или передача имущества в залог [5, c. 155].

Наличие недвижимости, главной характеристикой которой является стоимость, влияет на платёжеспособность, а, следовательно, и на итоговый скоринговый балл заёмщика (показатель, от величины которого зависит решение банка о предоставлении кредита).

В связи с вышепредставленным перечнем услуг возникает необходимость в разработке модели, которая позволит автоматизировать рутинный процесс оценки объектов недвижимости, включающий в себя подпроцесс подбора объектов-аналогов, и как следствие, сократить сроки проведения оценки при одновременном повышении её качества. Кроме того, за счёт использования разработанной модели унифицируется методология и подходы к оценке.

Для выбора наиболее подходящих аналогов модель должна обработать информацию о характеристиках объекта (этажность дома, этаж, на котором находится квартира, материал дома, общая площадь, cрок сдачи) и расстояниях до центров позитивного (станции метро, центр города, Московская Кольцевая Автомобильная дорога, озеленённые зоны отдыха и спортивные комплексы общего пользования) и негативного (промышленные предприятия, очистные сооружения) влияния.

Предполагается, что при совмещении таких областей, как математика и машинное обучение можно найти решение задачи массовой оценки недвижимости, опираясь на данные об объектах недвижимости, представленные интернет-сервисом для размещения объявлений Avito, и данные о центрах позитивного и негативного влияния, полученные посредством применения поисково-информационного картографического сервиса Яндекс.Карты. Главным инструментом в решении данной проблемы является математический аппарат, на котором и строятся методы машинного обучения.

# **Глава 2. Математические основы теории машинного обучения**

## **2.1 Определение и категории машинного обучения**

На сегодняшний день построение систем машинного обучения является одной из самых актуальных областей человеческой деятельности на стыке информационных технологий, математического анализа и статистики. Для построения методов машинного обучения используются средства математической статистики, численных методов, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными.

Машинное обучение в основном рассматривается как часть сферы искусственного интеллекта.

Под искусственным интеллектом будем понимать научное направление, занимающееся вопросами имитации процесса мышления человека с помощью компьютера, в рамках которого ставятся и решаются задачи аппаратного и программного [моделирования](https://wiki.loginom.ru/articles/modelling.html) тех видов человеческой деятельности, которые традиционно считаются интеллектуальными.

При таком подходе, машинное обучение – это направление искусственного интеллекта, связанное с разработкой и построением аналитических моделей, которые способны автоматически обнаружить в данных скрытые и ранее неизвестные закономерности, а также самостоятельно приобретать свойства, необходимые для распознавания этих закономерностей [7, c. 191].

Отличительной чертой машинного обучения является наличие в распоряжении исследователя набора данных, который описывает исследуемые объекты или процессы, и отражает присущие им свойства и закономерности. Такой набор данных называется обучающим множеством.

Таким образом, в машинном обучении обучающее множество представляет собой набор структурированных данных, используемый для обучения аналитических моделей [7, c. 192].

Фактически, обучающее множество представляет собой функцию, заданную таблично парами входных и выходных векторов . После обучения модель должна реализовывать эту функцию [7, c. 193].

Обучающее множество должно удовлетворять нескольким требованиям:

1. отражать правила и закономерности исследуемого процесса, которые должна обнаружить модель и по которым должно строится отображение;
2. быть репрезентативным, т.е. содержать достаточное количество уникальных примеров, как можно более полно отражающих закономерности исследуемого процесса;
3. удовлетворять установленным критериям качества данных;
4. не содержать дубликатов, противоречий, пропусков и аномальных значений.

Обучающее множество может быть получено как совокупность наблюдений за развитием объекта или процесса в прошлом, создана экспертом или аналитиком на основе некоторых гипотез, аналогий или личного опыта.

Наблюдения из обучающего множества, называемые обучающими примерами, последовательно предъявляются обучаемой модели, и в процессе этого она приобретает необходимые свойства. Данный процесс называется обучением модели, является итеративной процедурой, где на каждом шаге предъявляется объект данных из обучающей выборки и в соответствии с правилом, которое называется алгоритмом обучения, производится корректировка параметров модели. Обучение продолжается до тех пор, пока модель не достигнет способности выполнять требуемое преобразование с достаточной точностью. После того как модель обучена и протестирована, её можно использовать для анализа рабочих данных.

В зависимости от типа обучаемой модели и задачи анализа данных, которую она выполняет, обучение может вестись двумя способами: с учителем и без учителя. Кроме того, существуют метода частичного обучения, располагающиеся примерно по середине между машинным обучением с учителем и машинным обучением без учителя [7, c. 195].

* + 1. **Обучение с учителем**

Данное направление [машинного обучения](https://wiki.loginom.ru/articles/machine-learning.html) предполагает наличие полного набора размеченных данных для тренировки модели на всех этапах её построения, означающее, что каждому примеру в обучающем наборе соответствует ответ, который должен получить алгоритм.

Значение, которое должна выдавать модель после обучения называется целевым. Разность между целевым и фактическим выходами модели называется [ошибкой обучения](https://wiki.loginom.ru/articles/training-error.html), которая минимизируется в процессе обучения и выступает в качестве «учителя». Значение выходной ошибки затем используется для вычисления коррекций параметров модели на каждой итерации обучения [8, c. 70].

Формальная постановка задачи обучения с учителем выглядит следующим образом:

Пусть – множество описаний объектов, – множество допустимых ответов. Существует неизвестная целевая зависимость - отображение , значения которой известны только на объектах конечной обучающей выборки . Требуется построить алгоритм , приближающий неизвестную целевую зависимость как на элементах выборки, так и на всём множестве .

Данная постановка является обобщением классических задач аппроксимации функций. В классической аппроксимации объектами являются действительные числа или векторы. В прикладных же задачах входные данные об объектах могут быть неполными, неточными, неоднородными, нечисловыми. Эти особенности приводят к большому разнообразию методов обучения с учителем.

В анализе данных машинное обучение с учителем используется в задачах классификации и регрессии, изучению которых посвящён раздел об основных типах задачах машинного обучения.

* + 1. **Обучение без учителя**

 Часто перед алгоритмом стоит задача найти заранее не известные ответы. В этом случае возникает необходимость в обучении без учителя.

В данной технологии машинного обучения для коррекции параметров обучаемой модели не используется целевая функция. Иными словами, в обучающих примерах при обучении без учителя не нужно иметь заранее заданные выходы модели.

В алгоритмах обучения без учителя сложно вычислить точность, так как в данных отсутствуют «правильные ответы», поэтому выходная ошибка модели на обучающем множестве не вычисляется. Вместо неё используется информация о текущем состоянии параметров модели и примеров обучающего множества. Например, это может быть Евклидово расстояние между вектором признаков примера и вектором весов нейрона, которое и будет управлять коррекцией параметров модели в ходе обучения [8].

 Как правило, такое обучение пригодно только для задач, в которых известны описания множества объектов (обучающей выборки), и требуется обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами. Так, например, обучение без учителя применяется для решения таких задач, как поиск правил ассоциации, заполнение пропущенных значений, сокращение размерности, визуализация данных [7].

Типичными примерами моделей, использующих обучение без учителя, являются сети и карты Кохонена, широко используемые в анализе данных.

Тем не менее, основное применение обучения без учителя – построение моделей для кластеризации, подробнее о которой рассказывается в разделе об основных типах задачах машинного обучения.

## **2.2 Основные типы задач машинного обучения**

### **2.2.1 Задача классификации**

Классификация является популярной задачей анализа данных, которая применяется в кредитно-финансовой сфере, торговле, медицинской диагностике и многих других областях.

В области машинного обучения и статистики классификация представляет собой проблему определения того, к какой из категорий (подгрупп, классов) относится новое наблюдение, на основе обучающего набора данных, содержащего наблюдения, принадлежность к категории, которых известна [7].

Таким образом, классифицировать объект – значит, указать номер или наименование класса, к которому данный объект относится.

Когда новое наблюдение относится к определённому классу, свойства, характерные объектам этого класса, автоматически распространяются и на него. Если число классов ограничено двумя, то имеет место бинарная классификация, к которой могут быть сведены многие более сложные задачи.

В математической статистике задачи классификации называются также задачами дискриминантного анализа.

Как было сказано ранее, в терминологии машинного обучения классификация считается примером обучения с учителем, т. е. обучения, когда имеется обучающий набор правильно идентифицированных наблюдений.

Алгоритм, реализующий классификацию, известен как классификатор. Этот термин часто относится к математической функции, реализуемой алгоритмом классификации, которая отображает входные данные в категорию.

Рассмотрим математическую постановку задачи классификации.

Пусть – множество описаний объектов, – конечное множество номеров (имён, меток) классов. Существует неизвестная целевая зависимость – отображение , значения которой известны только на объектах конечной обучающей выборки . Требуется построить алгоритм , способный классифицировать произвольный объект [8].

Более общей считается вероятностная постановка задачи. Предполагается, что множество пар «объект, класс» является вероятностным пространством с неизвестной вероятностной мерой . Имеется конечная обучающая выборка наблюдений , сгенерированная согласно вероятностной мере . Требуется построить алгоритм , способный классифицировать произвольный объект [7].

Признаком называется отображение , где – множество допустимых значений признака. Если заданы признаки , то вектор называется признаковым описанием объекта . Признаковые описания допустимо отождествлять с самими объектами. При этом множество называют признаковым пространством [8].

Формально классификация производится на основе разбиения многомерного пространства признаков на области, в пределах каждой из которых многомерные векторы рассматриваются как идентичные. Иными словами, если объект попал в область пространства, ассоциированную с определённым классом, моделью относит его к этому классу.

Для решения задачи классификации используются такие алгоритмы, как нейронные сети, деревья решений, метод опорных векторов и др., при построении которых применяется обучение с учителем, когда выходная переменная (метка класса) задана для каждого наблюдения [8].

Однако, это деление не строгое поскольку, например, нейронные сети также могут быть адаптированы для решения задачи классификации.

Существует также обучение без учителя, когда разделение объектов обучающей выборки на классы не задаётся, и требуется классифицировать объекты только на основе их сходства друг с другом. В этом случае принято говорить о задаче кластеризации, изучению которой посвящён следующий раздел данной работы.

**2.2.2 Задача кластеризации**

Часто разделение объектов обучающей выборки на классы не задаётся, и объекты необходимо классифицировать только на основе их сходства друг с другом. Такого вида задачи относятся к типу задач кластеризации, а классы называются, соответственно, кластерами.

Таким образом, задача кластеризации – задача разбиения заданной выборки объектов (ситуаций) на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались [8]. При этом чем сильнее сходство объектов внутри кластера и их отличие от объектов в других кластерах, тем лучше кластеризация.

Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Рассмотрим формальную постановку задачи кластеризации.

Пусть – множество описаний объектов, – конечное множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами . Имеется конечная обучающая выборка объектов . Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, таким образом, чтобы каждый кластер состоял из объектов, близких по метрике , а объекты разных кластеров существенно отличались. При этом каждому объекту  приписывается номер кластера [8].

Алгоритм кластеризации – это функция , которая любому объекту  ставит в соответствие номер кластера . Множество  в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного критерия качества кластеризации [8].

Кластеризация позволяет увеличить понятность данных, упростить дальнейшую обработку и принятие решений, компактно представлять и хранить данные, обнаружить нетипичные объекты, которые не попали ни в один кластер [9].

Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы:

1. Отбор выборки для кластеризации. Подразумевается, что имеет смысл кластеризовать только количественные данные.
2. Определение множества переменных, по которым будут оцениваться объекты в выборке, то есть признакового пространства.
3. Вычисление значений той или иной меры сходства (или различия) между объектами.
4. Применение метода кластерного анализа для создания групп схожих объектов.
5. Проверка достоверности результатов кластерного решения [9].

Входные данные для кластерного анализа могут быть представлены в виде признакового описания объектов, матрицы расстояний между объектами или матрицы сходства объектов. Но в любом случае должны выполняться два фундаментальных требования к входным данным: однородность и полнота. Однородность требует, чтобы все кластеризуемые сущности были одной природы, описывались сходным набором характеристик, а полнота – достаточность информации для принятия решения [9].

В анализе данных кластеризация используется для сегментации клиентов и рынков, медицинской диагностики, социальных и демографических исследований, для определения кредитоспособности заёмщиков и во многих других областях.

Популярными алгоритмами обучения без учителя для решения задачи кластеризации являются: графовые алгоритмы, метод k-средних, EM-алгоритм, нейронная сеть Кохонена, иерархическая кластеризация.

Кластеризация может играть вспомогательную роль при решении задачи регрессии, изучению которой посвящён следующий раздел данной работы.

**2.2.3 Задача регрессии**

Регрессионный анализ является одним из наиболее распространённых методов обработки результатов экспериментов при изучении зависимостей в естественных науках, экономике, технике и других областях. В аналитических технологиях элементы регрессионного анализа широко используются для решения задач прогнозирования, оценивания, классификации, выявления зависимостей между [признаками](https://wiki.loginom.ru/articles/attribute.html).

Задача регрессии, как и задача классификации, относится к классу задач обучения с учителем, когда по заданному набору признаков наблюдаемого объекта необходимо найти некоторую целевую переменную. Отличие этих двух задач состоит в том, что задача классификации сводится к определению класса объекта по его характеристикам, а задача регрессии позволяет определить по известным характеристикам объекта значение некоторого его параметра, причём значением параметра является не конечное множество классов, а множество действительных чисел [10].

Рассмотрим постановку задачи регрессионного анализа.

Часто в задачах один из показателей выделяется в качестве результирующего и изучается влияние на него других показателей как факторов.

Между случайной величиной результата и случайной величиной фактора имеется стохастическая (случайная) зависимость, т.е. существует корреляционная зависимость.

Таким образом, корреляционную зависимость результата от факторов, проявляющуюся приблизительно и лишь в массе наблюдений, требуется отобразить с помощью функциональной зависимости результата от факторов, проявляющуюся определённо и точно в каждом конкретном случае. Задача состоит в определении случайной величины результата, если известны конкретные значения случайных величин факторов, от которых статистически зависит результат [10].

Функциональная зависимость результата от факторов представляется уравнением регрессии. Парная корреляция выражается зависимостью между двумя случайными величинами – результатом и одним фактором :

(1)

Множественная корреляция характеризует стохастическую связь между результатом и несколькими факторами :

(2)

Замена корреляционной зависимости на функциональную может привести к искажению отображения влияния факторов на результат. Поэтому общая задача регрессионного анализа состоит в определении такого вида и параметров уравнения регрессии, при которых наиболее точно представляется корреляционная зависимость.

Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих:

(3)

Присутствие в модели аддитивной случайной величины с нулевым математическим ожиданием , также называемой случайным членом, обусловлено рядом причин: ошибки спецификации (игнорирование важных объясняющих переменных, объединение переменных); ошибки измерения; ошибки, связанные со случайностью человеческих реакций [10].

Критерием качества приближения в задачах регрессии обычно является среднеквадратическая ошибка:

(4)

В формуле (4) – фактический результат, – предсказанный моделью результат, – количество объектов.

Решением проблемы оценки недвижимости является регрессионный анализ, для реализации которого в настоящее время разработано множество алгоритмов, некоторые из которых будут рассмотрены в следующем параграфе данной работы.

## **2.3 Основные алгоритмы регрессионного анализа**

### **2.3.1 Линейная регрессия**

Регрессионный анализ с использованием линейной функции, называется линейным регрессионным анализом.

Широкое применение линейной регрессии обусловлено тем, что большое количество реальных процессов в науке, экономике и бизнесе можно описать линейными моделями.

В математической статистике линейная регрессия представляет собой метод аппроксимации зависимостей между входными и выходными переменными на основе линейной модели.

Если рассматривается зависимость между одной входной переменной и одной выходной переменной , то имеет место простая линейная регрессия. Для этого определяется уравнение регрессии:

(4)

После определения уравнения регрессии строится соответствующая прямая, известная как линия регрессии (рис. 2.1).

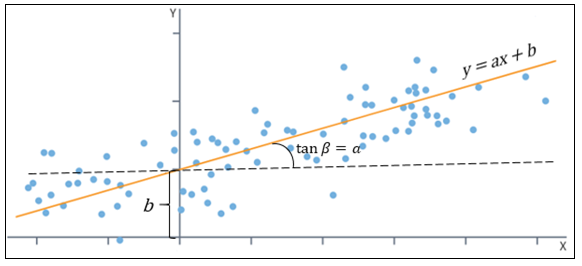


Рисунок 2.1 – Одномерная (простая) линейная регрессия

[Коэффициенты](https://wiki.loginom.ru/articles/coefficient-of-regression.html)  и , называемые также **параметрами модели,** определяются таким образом, чтобы сумма квадратов отклонений точек, соответствующих реальным наблюдениям данных, от линии регрессии, была бы минимальной. Коэффициенты обычно оцениваются методом наименьших квадратов, формально критерий которого можно записать следующим образом:

(5)

В формуле (5) – фактический результат, – предсказанный моделью результат, – количество объектов.

Более распространённой моделью является множественная линейная регрессия, которая предполагает установление линейной зависимости между множеством входных независимых и одной выходной зависимой переменной. Такая модель остается линейной, поскольку выходное значение представляет собой линейную комбинацию входных значений.

Уравнение множественной линейной регрессии имеет вид:

(5)

Как и в простой линейной регрессии, параметры модели  вычисляются при помощи метода наименьших квадратов. Отличие между простой и множественной линейной регрессией заключается в том, что вместо линии регрессии в ней используется гиперплоскость.

Параметр {\displaystyle b\_{0}}, при котором нет факторов, часто называют константой. Формально – это значение функции при нулевом значении всех факторов.

Коэффициент множественной регрессии показывает, на какую величину в среднем изменится результат , если переменную увеличить на единицу измерения, т. е. является нормативным коэффициентом.

Одномерная регрессионная модель является частным случаем более сложной модели – полиномиальной, которая используется для нелинейно разделяемых данных.

Один из способов, позволяющих применить линейную регрессию к нелинейным отношениям между переменными, – преобразование данных в соответствии с новыми базисными функциями. Идея состоит в том, чтобы взять многомерную модель (5) и построить на основе имеющегося одномерного входного значения . В таком случае, – функция, выполняющая преобразование данных.

Например, если , модель преобразуется в полиномиальную регрессию, в которой уравнение регрессии представляет собой полином -ой степени:

(6)

Причём модель по-прежнему является линейной – линейность относится к тому, что коэффициенты никогда не умножаются и не делятся друг на друга. Фактически осуществляется проекция одномерных значений на более многомерное пространство [10]. Таким образом, с помощью линейной регрессии можно отражать и более сложные зависимости между переменными. На рис.2.2 продемонстрирован пример использования линейной полиномиальной регрессии для более сложного распределения данных.

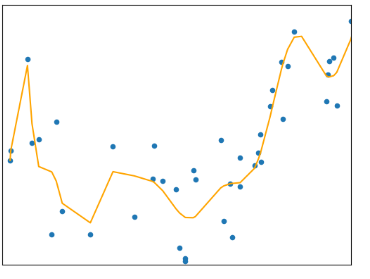


Рисунок 2.2 – Полиномиальная регрессия

Данный алгоритм не может быть применим для построения модели массовой оценки недвижимости, т.к. для него характерна низкая точность предсказания в задачах со сложной нелинейной зависимостью ответов от признаков.

**2.3.2 Метод опорных векторов**

Ещё один популярный и гибкий класс алгоритмов обучения с учителем как для классификации, так и регрессии – метод опорных векторов.

Методы этого класса особенно хорошо подходят для работы со сложными, но небольшими или средними наборами данных. В данном разделе объясняются ключевые концепции методов опорных векторов и особенности их использования для решения задач регрессионного анализа.

Метод опорных векторов довольно универсален, так как поддерживает не только линейную и нелинейную классификацию, но также линейную и нелинейную регрессию.

Метод опорных векторов для регрессии, называемый также опорным вектором регрессии, строится на основании опорного метода классификации, основная идея которого заключается в построении гиперплоскости или набора гиперплоскостей в многомерном или бесконечномерном пространстве. Алгоритм работает в предположении, что чем больше расстояние (зазор) между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше будет средняя ошибка классификатора.

Исходная задача может быть сформулирована в конечномерном пространстве, но часто данные для задачи не являются линейно разделимыми в этом пространстве. Поэтому было предложено исходное конечномерное пространство отображать в пространство большей размерности, что делает разделение намного проще.

Способ применения метода опорных векторов для решения задачи регрессии заключается в инвертировании цели: вместо попытки приспособиться к самой широкой (из возможных) области (полосе) между двумя классами, одновременно ограничивая нарушения зазора, регрессия пробует уместить как можно больше образцов на полосе наряду с ограничением нарушений зазора (т.е. образцов вне полосы). Ширина полосы управляется специальным гиперпараметром.

**2.3.3 Регрессионные деревья**

Ещё один инструмент, который активно используется в машинном обучении, анализе данных и статистике – дерево принятия решений.

Деревья решений являются одним из наиболее эффективных методов интеллектуального анализа данных и предсказательной аналитики, которые применяются для решения задач регрессии.

Данный раздел посвящён изучению деревьев принятия решений и трёх основных алгоритмов, использующих эти деревья для построения регрессионных моделей.

В основе работы дерева решений лежит процесс рекурсивного разбиения исходного множества объектов на подмножества, ассоциированные с предварительно заданными классами. Разбиение производится с помощью решающих правил, в которых осуществляется проверка значений атрибутов по заданному условию [13].

В связи с тем, что деревья решений являются моделями, строящимися на основе обучения с учителем, в обучающем множестве для примеров должно быть задано целевое значение. При этом, если целевая переменная непрерывная, то модель называют деревом регрессии [10].

Структурно дерево решений состоит из объектов двух типов – узлов и листьев. В узлах расположены решающие правила и подмножества наблюдений, которые им удовлетворяют.

В простейшем случае, в результате проверки, множество примеров, попавших в узел, разбивается на два подмножества, в одно из которых попадают примеры, удовлетворяющие правилу, а в другое – не удовлетворяющие. Затем к каждому подмножеству вновь применяется правило, и процедура рекурсивно повторяется пока не будет достигнуто некоторое условие остановки алгоритма. В результате в последнем узле проверка и разбиение не производится, и он объявляется листом. Когда дерево построено, каждому листу ставится в соответствие ответ. Для дерева регрессии — это соответствующее листу среднее значение, медиана или другая функция от целевых переменных объектов в листе [13].

Таким образом, в отличие от узла, в листе содержится не правило, а подмножество объектов, удовлетворяющих всем правилам, лежащим на пути к этому листу.

Поскольку путь в дереве к каждому листу единственный, то и каждый пример может попасть только в один лист, что обеспечивает единственность решения.

Такая иерархическая структура данных, в которой каждый узел имеет не более двух потомков называется бинарным (двоичным) деревом (рис. 2.3).

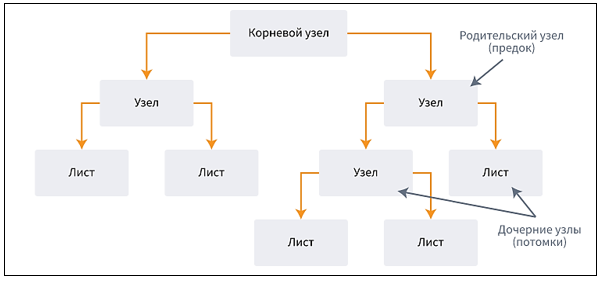


Рисунок 2.3 – Бинарное дерево принятия решений в общем виде

Деревья решений являются жадными алгоритмами, т.е. на каждом шаге ищут локально-оптимальное решение, предполагая, что конечное общее решение, являющееся суперпозицией локальных, также будет оптимальным.

Широкая популярность деревьев решений обусловлена следующими их преимуществами:

1. правила в них формируются практически на естественном языке, что делает объясняющую способность деревьев решений очень высокой;
2. могут работать как с числовыми, так и с категориальными данными;
3. требуют относительно небольшой предобработки данных, в частности, не требуют нормализации, создания фиктивных переменных, могут работать с пропусками;
4. могут работать с большими объёмами данных [13].

Рассмотрим несколько основных методов, которые используют деревья принятия решений:

* 1. CART (англ. Classification and regression trees – Классификационные и регрессионные деревья).

Данный алгоритм построения деревьев решений работает как с дискретной, так и с непрерывной выходной переменной, т.е. применяется для решения задач классификации и регрессии.

Суть алгоритма CART состоит в построении бинарного дерева принятия решений, содержащего два потомка в каждом узле. На каждом шаге построения дерева правило, формируемое в узле, делит заданное множество примеров (обучающую выборку) на две части – часть, в которой выполняется правило (правый потомок) и часть, в которой правило не выполняется (левый потомок) [13].

Для выбора оптимального правила используется функция оценки качества разбиения. Оценочная функция, используемая алгоритмом CART, базируется на интуитивной идее уменьшения неопределённости в узле [10].

Основная идея алгоритма заключается в том, чтобы выбрать такое разбиение из всех возможных в данном узле, чтобы полученные дочерние узлы были максимально однородными. При этом каждое разбиение производится только по одному атрибуту [13].

Так, например, если атрибут  является непрерывным с различными значениями, существует  различных разбиений по .

В результате работы этого алгоритма получается одно дерево принятия решений. За счёт этого, он удобен для первичного анализа данных, например, проверить наличие связей между переменными.

* 1. Случайный лес.

В основе данного метода лежит использование ансамбля деревьев принятия решений.

В статистике и машинном обучении под ансамблем моделей понимают комбинацию нескольких алгоритмов обучения, которые, работая вместе, позволяют построить модель более эффективную и точную, чем любая из моделей, построенная с помощью отдельного алгоритма. Модель, построенную на основе ансамбля, называют метамоделью [11].

Данный алгоритм использует две ключевые концепции:

1. случайная выборка образцов из набора данных при построении деревьев;
2. при разделении узлов выбираются случайные наборы параметров.

Суть концепции о случайной выборке тренировочных образцов состоит в том, что в процессе тренировки каждое дерево случайного леса обучается на случайном образце из набора данных. Выборка образцов происходит с возмещением (в статистике этот метод называется бутстреппинг), что даёт возможность повторно использовать образцы одним и тем же деревом [12].

Хотя каждое дерево может быть высоковариативным по отношению к определённому набору тренировочных данных, обучение деревьев на разных наборах образцов позволяет понизить общую вариативность леса, не жертвуя точностью. При тестировании результат выводится путём усреднения прогнозов, полученных от каждого дерева [14].

Подход, при котором каждый обучающийся элемент получает собственный набор обучающих данных (с помощью бутстреппинга), после чего результат усредняется, называется бэггинг [11].

Вторая базовая концепция случайного леса заключается в использовании определённой выборки параметров образца для разделения каждого узла в каждом отдельном дереве. Обычно размер выборки равен квадратному корню из общего числа параметров. Хотя обучение случайного леса можно провести и с полным набором параметров, как это обычно делается при регрессии [11].

Таким образом, случайный лес включает в себя множество деревьев принятия решений, каждое из которых обучается на отдельной выборке данных, узлы в каждом дереве разделяются с использованием ограниченного или полного (в задачах регрессии) набора параметров. Итоговый прогноз делается путём усреднения прогнозов всех деревьев (рис. 2.4).

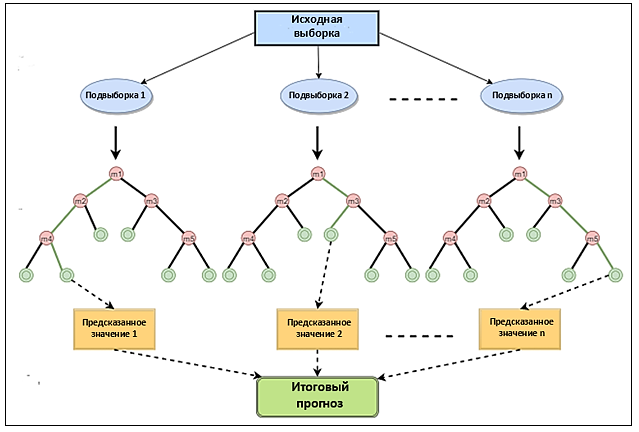


Рисунок 2.4 – Схема функционирования случайного леса

* 1. Градиентный бустинг.

Данная техника машинного обучения для решения задачи регрессии (к которой можно свести классификацию) является разновидностью бустинга и строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений [11].

В отличие от бэггинга, на котором основан алгоритм случайного леса, бустинг представляет собой такую технику построения ансамблей, в которой предсказатели строятся не независимо, а последовательно. Кроме того, бустинг использует идею о том, что следующая модель будет учится на ошибках предыдущей [12].

Данный алгоритм на каждом шаге пытается исправить ошибки предыдущего дерева, т.е. обучает новые модели по остаточной ошибке прошлых, двигаясь к минимуму функции потерь, в качестве которой может выступать среднеквадратичная ошибка (MSE).

(8)

В формуле (8) – фактический результат, – предсказанный моделью результат, – количество объектов.

На первой итерации этого метода осуществляется построение первого базового алгоритма - ограниченного по количеству узлов дерева принятия решений. После чего вычисляется ошибка, как разность между тем, что предсказало полученное дерево и искомой переменной на этом шаге. Затем строится новая модель с учётом ошибок как целевых переменных. При этом требуется найти лучшее разделение для минимизации ошибки. Предсказания, сделанные с помощью этой новой модели, сочетаются с предсказаниями предыдущих. Затем снова вычисляются ошибки.  
Процесс продолжается до тех пор, пока результат не перестанет улучшаться (рис. 2.5). Стоит отметить, что критерий остановки следует выбирать таким образом, чтобы избежать переобучения модели [12], [14].

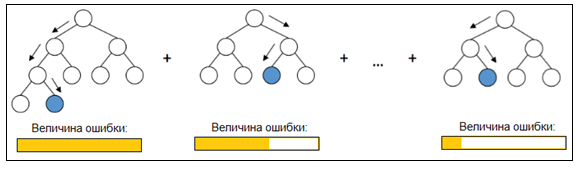


Рисунок 2.5 - Схема функционирования градиентного бустинга

Для построения модели оценки недвижимости будут использованы два метода – случайный лес и градиентный бустинг, и усреднены их предсказания.

# **Глава 3. Построение модели массовой оценки недвижимости в городе Москве методами машинного обучения**

## **3.1 Выбор инструментов реализации для построения модели**

Выбор языка программирования является важным моментом, от которого зависит скорость создания программы, скорость тестирования, возможность переноса на другие платформы, возможность быстрого внесения изменений, быстрота выполнения конечного продукта и так далее.

Обычно вопрос о выборе языка возникает на этапе перевода модели с математического языка на язык программирования, на котором будет реализована соответствующая прикладная программа. Однако в данной работе определиться с выбором языка программирования необходимо уже на этапе сбора данных, так как возникает необходимость в создании программы, реализующей сбор и систематизирование информации об объектах недвижимости, опубликованной на сайте для размещения объявлений, и центрах влияния на их стоимость, посредством применения поисково-информационного картографического сервиса Яндекс. Карты.

Таким образом, необходимо выбрать удобный в использовании язык программирования, применимый для сбора и анализа большого объёма данных и поддерживающий библиотеки для работы в области машинного обучения.

На данный момент самыми популярными инструментами программирования для работы с данными являются Python и R.

Язык программирования R обладает обширным каталогом статистических и графических методов, поэтому считается лучшим инструментом для решения статистических задач.

Язык программирования Python в свою очередь выполняет почти ту же работу, что и R, но специалисты в области аналитики для работы с данными предпочитают его по причине сравнительной простоты и высокой производительности [15].

Оба языка программирования являются бесплатными, с открытыми исходными кодами.

Сравним языки по следующим критериям:

1. Простота использования.

R является сравнительно сложным языком не только для людей без опыта программирования, но и для программистов, привыкших работать с другими языками. Это обусловлено тем, что ему присущи такие особенности, как начало индексирования с единицы, использование нескольких операторов присваивания, нетрадиционные структуры данных. Python же в свою очередь по причине читаемости кода очень прост в изучении, что делает его предпочтительным как начинающих, так и опытных программистов [16].

1. Производительность.

Несмотря на появление в R пакетов, повышающих его производительность, он уступает Python по этому критерию и считается достаточно медленным языком. Так, например, R потребуется почти в два раза больше времени, чтобы загрузить csv-файл, чем Python, используя библиотеку Pandas.

1. Универсальность.

R является специфичным и узкоспециализированным языком и больше подходит для статистических исследований. В то время как Python считается универсальным языком общего назначения и применяется во многих сферах.

1. Графика и визуализация.

R состоит из многочисленных пакетов, которые обеспечивают расширенные графические возможности, такие как ggplot2 и ggvis. Python в свою очередь тоже имеет некоторые библиотеки визуализации, такие как seaborn и bokeh, но они читаются более сложными. На сегодняшний день Python уступает R в наглядности визуализации.

1. Популярность.

На сегодняшний день Python стал более популярным у массы пользователей по сравнению с R. Кроме того, по статистике, процент специалистов, переключившихся с R на Python, в два раза превышает процент пользователей, переключившихся с Python на R.

Результаты сравнения проиллюстрированы в табл. 3.1.

Таблица 3.1

Сравнение языков программирования R и Python

|  |  |  |
| --- | --- | --- |
| Критерий | R | Python |
| Простота использования | - | + |
| Производительность | - | + |
| Универсальность | - | + |
| Графика и визуализация | + | - |
| Популярность | - | + |

Анализируя приведённую выше таблицу, можно сделать вывод, что Python является сравнительно простым в использовании, мощным, универсальным языком, который можно использовать для различных задач в области компьютерных наук. Таким образом, все прикладные программы в практической части данной работы будут реализованы на языке программирования Python.

Кроме сравнительного анализа на выбор языка повлияло так же наличие в Python множества библиотек, что позволяет значительно упростить программирование на этом языке без написания лишнего кода. Стоит отметить, что в число библиотек также входят библиотеки, которые содержат необходимые инструменты для сбора данных об объектах и реализации построения модели массовой оценки недвижимости посредством методов машинного обучения.

Таким образом, для реализации процесса сбора данных будут использованы библиотеки:

1. Peewee - библиотека для работы с базами данных, поддерживающая выбранную для промежуточного хранения информации об объектах недвижимости систему управления базами данных [SQLite](https://python-scripts.com/sqlite);
2. Requests – библиотека для создания HTTP-запросов;
3. Selenium – библиотека для автоматизированного управления браузерами;
4. Beautifulsoup4 – библиотека для синтаксического разбора HTML-файлов.

Для реализации процесса построения модели массовой оценки недвижимости методами машинного обучения будут использованы библиотеки:

1. Pandas – библиотека для анализа и обработки данных;
2. Scikit-learn – высокоуровневая библиотека, содержащая множество инструментов для машинного обучения, в том числе и алгоритм случайного леса, выбранный для построения модели оценки недвижимости;
3. Matplotlib – библиотека, представляющая собой двумерную числовую структуру, которая используется для визуализации данных двумерной графикой;
4. XGBoost – библиотека машинного обучения, реализующая алгоритм градиентного бустинга, выбранный для построения модели оценки недвижимости.

Также для построения модели массовой оценки недвижимости будет использован сервис Google Colab, среди преимуществ которого выделяют то, что является бесплатным, работает в браузере и предоставляет для расчётов GPU – мощный графический процессор NVIDIA Tesla K80, благодаря которому можно кратно ускорить обучение моделей. Кроме того, данный сервис позволяет запускать Jupyter Notebook, представляющий собой удобную для научных расчётов оболочку языка Python, в которой можно писать код блоками, запуская их поэтапно.

## **3.2 Сбор и обработка данных**

Перед тем, как приступить к построению модели необходимо осуществить сбор данных об объектах недвижимости и центрах позитивного и негативного влияния на их стоимость.

В качестве источника данных об объектах недвижимости будет использован интернет-сервис для размещения объявлений Avito. Данные о центрах позитивного и негативного влияния будут получены с помощью поисково-информационного картографического сервиса Яндекс.Карты.

В первую очередь необходимо собрать информацию о центрах влияния.

В качестве центров позитивного влияния в данной работе выделены:

1. станции метро;
2. центр города;
3. МКАД;
4. озеленённые зоны отдыха;
5. спортивные комплексы общего пользования.

В качестве центров негативного влияния:

1. промышленные предприятия;
2. очистные сооружения.

Сбор информации о таких центрах влияния, как озеленённые зоны отдыха, спортивные комплексы общего пользования, промышленные предприятия, очистные сооружения, осуществляется с помощью программы, которая посылает API-запросы на сервис Яндекс.Карты с помощью методов библиотеки Requests, т.е. напрямую обращается к API-серверу и получает наименование, адрес, широту и долготу каждого центра влияния. Затем полученные данные сохраняются в csv-файлы, причём каждой группе центров влияния соответствует отдельный файл. В результате сформированы четыре scv-файла, содержащих информацию о центрах влияния: озеленённые зоны отдыха – 500 объектов, спортивные комплексы общего пользования – 500 объектов, промышленные предприятия – 500 объектов, очистные сооружения – 9 объектов.

На рис. 3.1 представлен фрагмент данных об озеленённых зонах отдыха.

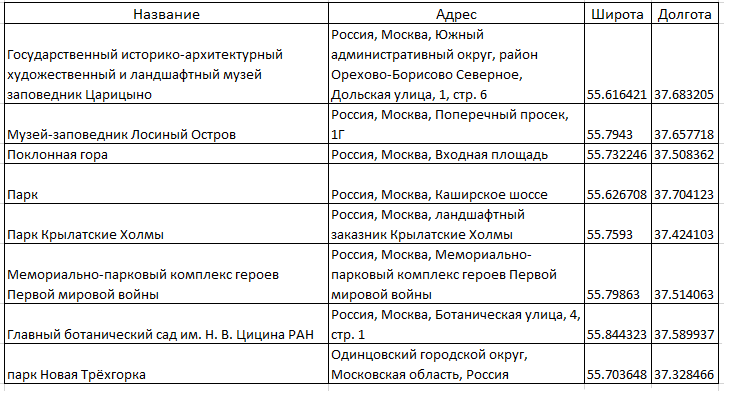


Рисунок 3.1 – Фрагмент данных об озеленённых зонах отдыха г. Москвы

На рис. 3.2 представлен фрагмент данных о спортивных комплексах общего пользования города.

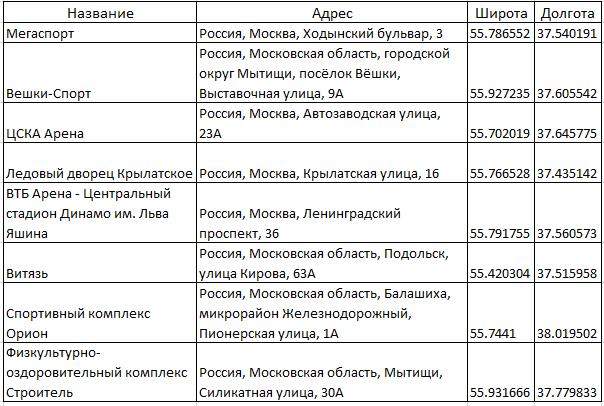


Рисунок 3.2 – Фрагмент данных о спортивных комплексах общего пользования г. Москвы

На рис. 3.3 представлен фрагмент данных об очистных сооружениях города.



Рисунок 3.3 – Фрагмент данных об очистных сооружениях г. Москвы

На рис. 3.4 представлен фрагмент данных о промышленных предприятиях города.

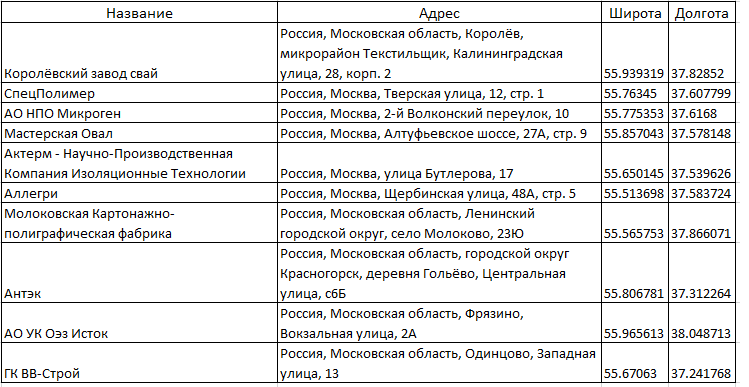


Рисунок 3.4 - Данные о промышленных предприятиях города г. Москвы

Программный код сбора данных о центрах влияния представлен в Приложении 1.

Данные обо всех центрах влияния города представлены в Приложении 2.

На следующем шаге сбора данных необходимо получить данные о таких центрах положительного влияния, как центр города, МКАД и станции метро.

В качестве центра города в данной работе выступает Московский Кремль, координаты которого можно получить вручную с помощью сервиса Яндекс.Карты. Таким образом, координаты широты и долготы центра города – (55.751637; 37.618530).

МКАД представляет собой автомобильную кольцевую дорогу (рис. 3.5).

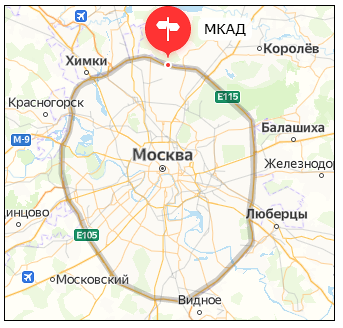


Рисунок 3.5 – МКАД в Яндекс.Картах

Координаты каждого километра МКАДа получены вручную также с помощью сервиса Яндекс.Карты и представлены в Приложении 2.

В процессе сбора данных, получение координат широты и долготы центров влияния и объектов недвижимости необходимо для того, чтобы рассчитать расстояния между ними. Станции метро являются положительным центром влияния, но получать их координаты нет необходимости, так как на сайте для размещения объявлений Avito для каждого объекта недвижимости уже указано расстояние до ближайшей станции метро, которое будет собрано вместе с остальными данными об объектах недвижимости.

После того как процесс сбора данных о центрах позитивного и негативного влияния завершён, следует перейти к сбору данных об объектах недвижимости.

Так как модель будет ориентирована на оценку квартир нового рынка жилья города Москвы, необходимо осуществить сбор информации только о новостройках.

Интернет-сервис для размещения объявлений Avito борется с массовым сбором данных, так как они могут быть использованы в мошеннических целях. По этой причине сервис блокирует большое количество API запросов.

С помощью методов библиотеки для автоматизированного управления браузерами Selenium программа сможет максимально точно воспроизвести поведение пользователя на этом сайте, и блокировки удастся избежать.

Сначала программа осуществляет сбор ссылок на объекты недвижимости и сохранение их в базу данных SQLite. Промежуточное хранение информации об объектах недвижимости в базе данных необходимо для обеспечения надёжности сохранения данных в случае прерывания работы программы по различным причинам, и для исключения дубликатов из выборки в процессе обработки данных.

Затем осуществляется сбор полной информации об объектах недвижимости на основе ссылок, собранных на предыдущем этапе, и сохранение её в базу данных SQlite. На этом этапе программа собирает абсолютно всю информацию об объектах, так как выбор характеристик, влияющих на стоимость квартир, будет сделан на этапе формирования итогового scv-файла.

В результате собраны все представленные на сайте данные о квартирах в новостройках: студии – 452 объекта, однокомнатные квартиры – 3 218 объектов, двухкомнатные квартиры – 4 045 объектов, трёхкомнатные квартиры – 2 883 объекта, четырёхкомнатные квартиры – 693 объекта, пятикомнатные квартиры – 159 объектов. Таким образом, всего 11 450 квартир нового рынка г. Москвы.

После того, как все необходимые данные собраны, производится расчёт расстояний до всех центров влияния. Опишем этот процесс подробнее.

Каждый объект недвижимости и центр влияния имеют две координаты, измеряемые в градусах, для вычисления расстояний между такими объектами существует формула, которая выглядит следующим образом:

(9)

В формуле (9) - широта и долгота двух точек в радианах, - разница координат по долготе, - угловая разница, - радиус Земли.

Единицы конечного расстояния будут равны единицам, в которых выражен радиус (в данном случае - метры).

Таким образом, используя, координаты широты и долготы для всех объектов недвижимости и центров влияния, по формуле (9) осуществляется расчёт расстояний между каждым объектом недвижимости и каждым центром влияния, причём если, например, необходимо вычислить расстояние между конкретной квартирой и центром влияния определённого типа, вычисляются расстояния от этой квартиры до всех центров влияния этого типа, и выбирается минимальное.

Особенность вычисления до такого центра влияния, как МКАД состоит в том, что осуществляется расчёт расстояний от объекта недвижимости до каждого километра МКАДа, и выбирается минимальное. Причём, для объектов, находящихся внутри МКАДа, вычисленное расстояние необходимо сделать отрицательным. Это объясняется тем, что чем больше расстояние от объекта недвижимости, находящегося за пределами МКАД, до МКАД, тем меньше стоимость, но при этом, чем больше расстояние от объекта недвижимости, находящегося в пределах МКАД, тем меньше расстояние до центра города, а значит выше стоимость.

После того, как все расстояния вычислены, формируется итоговый csv-файл, содержащий информацию об объектах недвижимости и расстояниях до центров влияния, на основании которого будет построена модель.

На рис. 3.6, 3.7 представлен фрагмент данных с характеристиками квартир города.

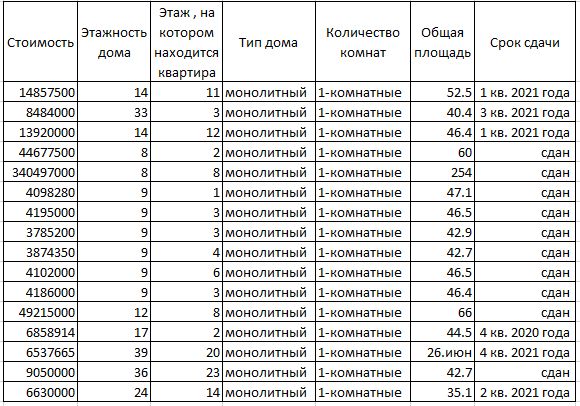


Рисунок 3.6 – Фрагмент данных о квартирах г. Москвы



Рисунок 3.7 – Продолжение фрагмента данных о квартирах

Программный код сбора данных об объектах недвижимости представлен в Приложении 1. Данные обо всех объектах недвижимости представлены в Приложении 2.

## **3.3 Описание построения модели массовой оценки недвижимости в городе Москве**

На основе собранных данных с использованием выбранных средств для реализации, будет построена модель массовой оценки недвижимости в городе Москве, описанию процесса построения которой посвящён данный раздел.

В первую очередь требуется подключить библиотеки, описанные в разделе 3.1, с помощью команды import (рис. 3.8).

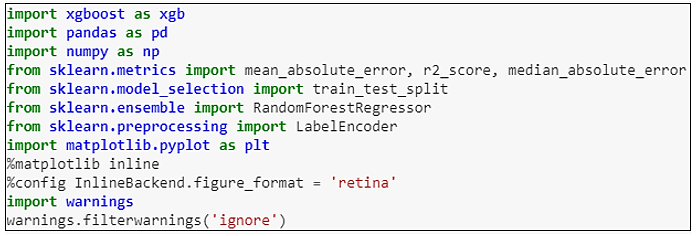


Рисунок 3.8 – Подключение библиотек для построения модели

Затем необходимо объявить функции для расчёта метрик и коэффициента детерминации.

Метрики качества нужны, чтобы сравнивать модели между собой и понимать, насколько хорошо модель описывает реальность. Для оценки результатов модели будут использованы следующие две метрики: средняя абсолютная процентная ошибка и медианная абсолютная процентная ошибка.

Средняя абсолютная процентная ошибка (MAPE) - показывает среднее арифметическое значение всех абсолютных (взятых по модулю) процентных ошибок прогноза и рассчитывается по формуле:

(10)

В формуле (10) – практическое значение стоимости, – предсказанная моделью стоимость.

Медианная абсолютная процентная ошибка (MedAPE) основана на той же идее вычисления абсолютной процентной ошибки, но вместо среднего арифметического используется более устойчивая к выбросам медиана.

Кроме того, следует рассчитать коэффициент детерминации (), представляющий собой долю дисперсии зависимой переменной, объясняемой рассматриваемой моделью зависимости, то есть объясняющими переменными. Чем ближе коэффициент детерминации к единице, тем лучше модель описывает реальность.

Объявляем функции для расчёта метрик и коэффициента детерминации (рис. 3.9).

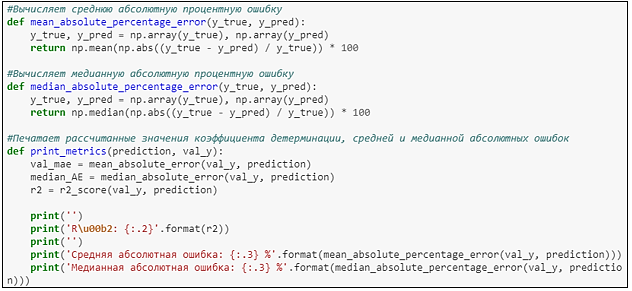


Рисунок 3.9 – Объявление функций

На следующем этапе с помощью библиотеки Pandas необходимо загрузить данные и сделать первичную их обработку (рис. 3.10). Библиотека Pandas позволяет произвести считывание csv-файла и преобразование его в новый формат данных - датафрейм (таблицу).

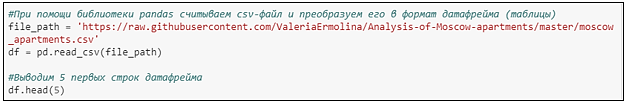
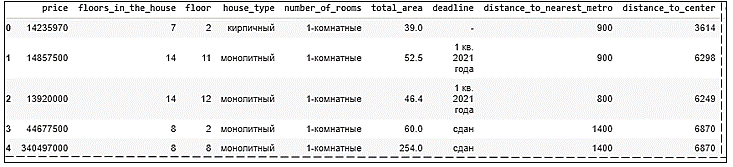
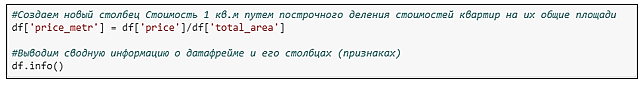


Рисунок 3.10 – Загрузка данных

Последняя строка кода df.head(5) выводит пять первых строк таблицы, чтобы визуально оценить информацию (рис. 3.11).

Рисунок 3.11 – Фрагмент таблицы данных

Далее необходимо создать дополнительный столбец: цену одного квадратного метра квартиры (именно это значение будет предсказывать модель) (рис. 3.12).

Рисунок 3.12 – Создание дополнительного столбца

Последняя строка кода df.info() выводит сводную информацию о датафрейме и его столбцах (рис. 3.13).

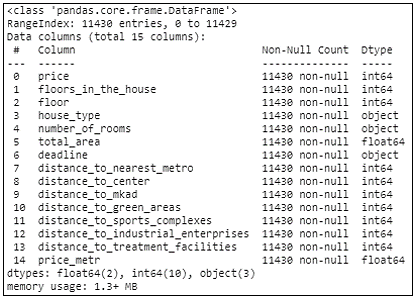


Рисунок 3.13 – Сводная информация о датафрейме и его столбцах

В сводной информации о датафрейме и его столбцах видно, что такие столбцы, как тип дома (house\_type), количество комнат (number\_of\_rooms) и срок сдачи дома (deadline) имеют тип object, т.е. не являются числовыми, в отличие от остальных столбцов, поэтому их формат необходимо заменить при помощи функции **LabelEncoder()** библиотеки Scikit-learn, (рис. 3.14).

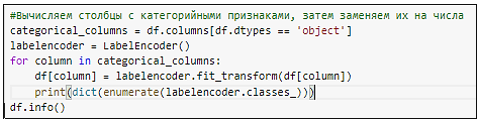


Рисунок 3.14 – Вычисление столбцов с категорийными признаками и замена их на числа

Затем необходимо вывести сводную информацию о датафрейме и его столбцах(признаках), чтобы убедиться, что теперь все они содержат цифровые значения (рис. 3.15).

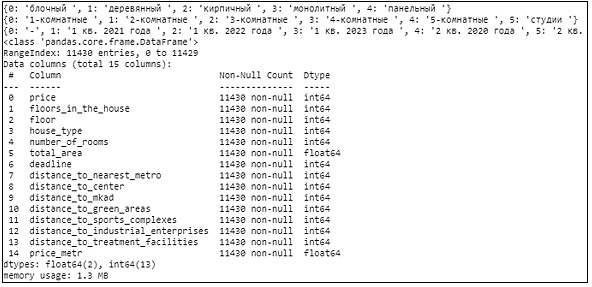


Рисунок 3.15 – Сводная информация о датафрейме и его обновлённых столбцах (признаках)

Последний шаг подготовки к обучению - выделение целевой переменной, которую будет предсказывать модель. Следует обозначить эту переменную как и присвоить ей столбец датафрейма «Цена квадратного метра».

Затем необходимо отобрать признаки, на основании которых будет строится модель: количество этажей в доме (floors\_in\_the\_house); этаж, на котором находится квартира (floor); тип дома (house\_type); количество комнат (number\_of\_rooms); общая площадь (total\_area); срок сдачи (deadline); расстояние до метро (distance\_to\_nearest\_metro); расстояние до центра (distance\_to\_nearest\_center); расстояние до МКАД (distance\_to\_nearest\_mkad); расстояние до озеленённых зон отдыха (distance\_to\_green\_areas); расстояние до спортивных комплексов общего пользования (distance\_to\_sports\_complexes); расстояние до промышленных предприятий (distance\_to\_industrial\_enterprises); расстояние до очистных сооружений (distance\_to\_treatment\_facilities).

После чего необходимо создать новый датафрейм , состоящий из этих признаков, и провести случайное разделение данных на выборки для обучения (train) и валидации (val), по умолчанию в пропорции 75% на 25%. Функция train\_test\_split()автоматически разбивает  и  на четыре группы: тренировочный , валидационный (проверочный) , тренировочный , валидационный  (рис. 3.16). Это позволит проверить качество модели на незнакомых ей данных.

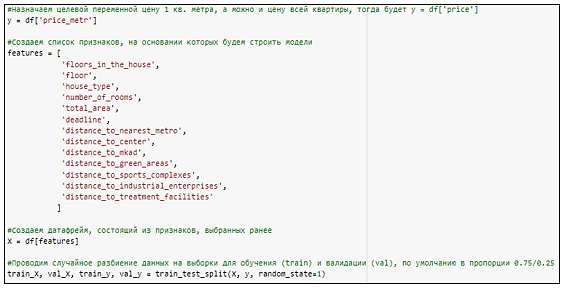


Рисунок 3.16 – Определение целевой переменной и признаков, создание нового датафрейма, разбиение данных на обучающую и валидационную выборки

Далее следует приступить к обучению первой модели, основанной на алгоритме случайного леса деревьев решений. Необходимо создать объект RandomForestRegressor, ранее импортированный из библиотеки Scikit-learn, задать ему ряд гиперпараметров (однако если гиперпараметры не будут заданы, то алгоритм установит их по умолчанию), после чего сохранить его как rf\_model.

За обучение модели отвечает метод fit, который применяется к созданной модели rf\_model. После того, как модель обучится, с помощью метода predict будут предсказаны цены для тестовой выборки квартир (val\_X). А с помощью объявленной в начале функции print\_metrics(), на основании предсказанных и фактических значений стоимостей, будут напечатаны значения метрик.

Процессы создания регрессионной модели случайного леса, обучения её на обучающей выборке, вычисления предсказанных значений на основе валидационной выборки, вычисления и вывода величин ошибок продемонстрированы на рис. 3.17.

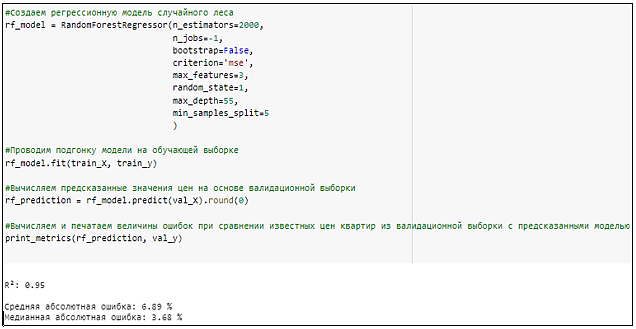


Рисунок 3.17 – Создание и результат работы модели случайного леса

Затем следует обучить модель градиентного бустинга. Необходимо создать объект XGBRegressor, импортированный из библиотеки градиентного бустинга XGBoost, задать ему ряд гиперпараметров (однако если гиперпараметры не будут заданы, то алгоритм установит их по умолчанию), после чего сохранить его как xgb\_model.

За обучение модели, так же как и в модели случайного леса, отвечает метод fit, который применяется к созданной модели xgb\_model. После того, как модель обучится, с помощью метода predict будут предсказаны цены для тестовой выборки квартир (val\_X). С помощью функции print\_metrics() будут напечатаны значения метрик.

Процессы создания регрессионной модели градиентного бустинга, обучения её на обучающей выборке, вычисления предсказанных значений, вычисления и вывода величин ошибок продемонстрированы на рис. 3.18.

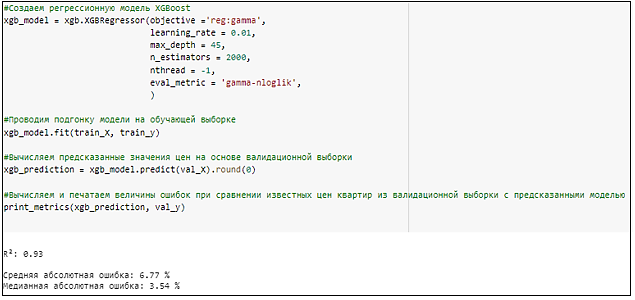


Рисунок 3.18 – Создание и результат работы модели случайного леса

На последнем этапе построения модели с целью улучшения предсказания необходимо усреднить результаты моделей случайного леса и градиентного бустинга (рис. 3.19).

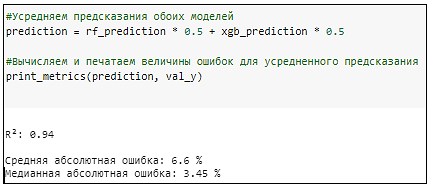


Рисунок 3.19 – Усреднение предсказаний моделей

## **3.4 Результат работы модели**

Перед тем, как перейти к оценке результатов работы модели на конкретном примере, необходимо рассчитать важность признаков, построить их рейтинг и отразить его в виде столбчатой диаграммы важности признаков (рис. 3.20). Это позволит выделить как признаки, оказывающие сильное влияние на стоимость квартир, так и те факторы, которые незначительно влияют на стоимость (рис. 3.21).

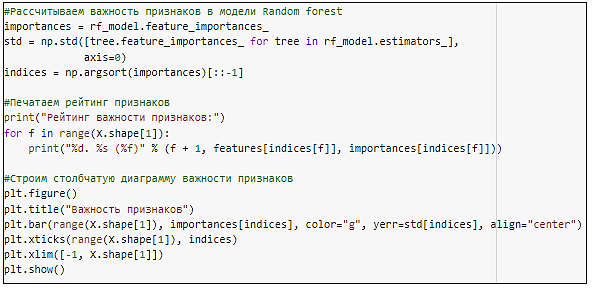


Рисунок 3.20 – Расчёт важности признаков, построение рейтинга и диаграммы важности

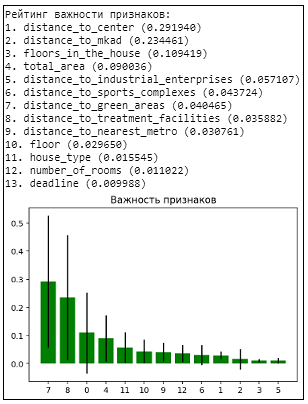


Рисунок 3.21 – Рейтинг и диаграмма важности признаков

Таким образом рейтинг признаков, влияющих на стоимость квартир на новом рынке города Москвы, имеет вид:

* 1. Расстояние до центра – 29,2%.
  2. Расстояние до МКАД – 23,4%.
  3. Количество этажей в доме – 10,9%.
  4. Общая площадь – 9%.
  5. Расстояние до промышленных предприятий – 5,7%.
  6. Расстояние до спортивных комплексов – 4,4%.
  7. Расстояние до озеленённых зон отдыха – 4%.
  8. Расстояние до очистных сооружений -3,6%.
  9. Расстояние до метро – 3,1%.
  10. Этаж, на котором находится квартира – 3%.
  11. Тип дома – 1,6%.
  12. Количество комнат – 1,1%.
  13. Срок сдачи – 1%.

Таким образом, построенный рейтинг признаков является подтверждением факта, что главным определяющим фактором стоимости недвижимости является её месторасположение. А региональная особенность столицы состоит в том, что самыми влияющими факторами являются расстояния до центра города и МКАД.

Как правило, квартиры в многоэтажных домах в целом обычно дешевле, чем в среднеэтажных и малоэтажных, поэтому этот признак лидирует в группе характеристических факторов.

Далее следует такой характеристический фактор, как общая площадь, логично, что чем больше квартира, тем она дороже, но стоит отметить, что по мере увеличения площади жилья снижается стоимость квадратного метра.

Расстояния до остальных центров влияния оказывают примерно одинаковое влияние на стоимость объектов.

Такие характеристики квартир как тип дома, количество комнат и срок сдачи оказывают сравнительно меньшее влияние на формирование стоимости.

С помощью модели оценки недвижимости, процесс построения которой описан в предыдущем разделе, вычислим стоимости двадцати различных квартир в новостройках города Москвы (рис. 3.22, 3.33).

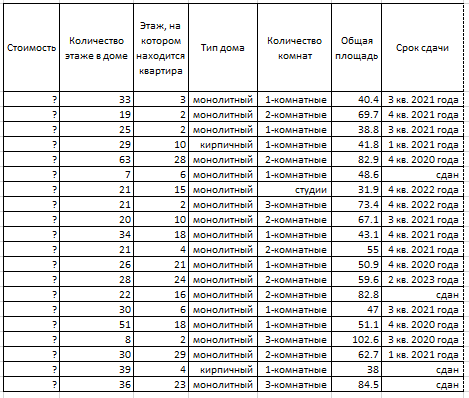


Рисунок 3.22 – Первая часть данных о квартирах, которые необходимо оценить

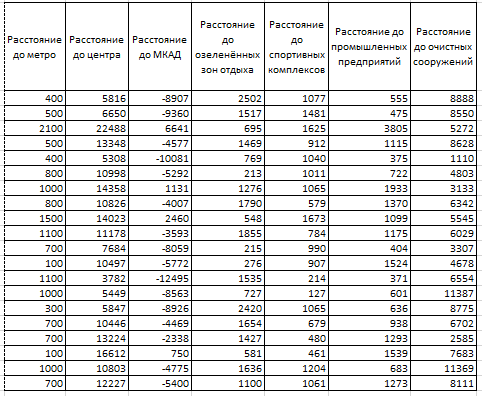


Рисунок 3.22 – Вторая часть данных о квартирах, которые необходимо оценить

Данные об объектах, которые необходимо оценить, должны быть сохранены в csv-формате, после чего их можно импортировать и приступить непосредственно к предсказанию стоимости (рис. 3.23).

Для того, чтобы сравнить фактическое и предсказанное значения стоимости каждого оцениваемого объекта, необходимо оценивать квартиры, стоимость которых известна оценщику и неизвестна модели.

Таким образом, результате работы модели будут предсказаны не только значения стоимостей объектов, но и величина отклонений их от фактических значений.



Рисунок 3.23 – Массовая оценка квартир нового рынка г. Москвы

Результат работы модели представлен на рис. 3.24, где столбец price – фактическая стоимость объектов, price\_prediction – предсказанная стоимость объектов, mistake – модуль разности фактической и предсказанной стоимостей, т.е. величина отклонения, mistake\_percentage – процент отклонения.

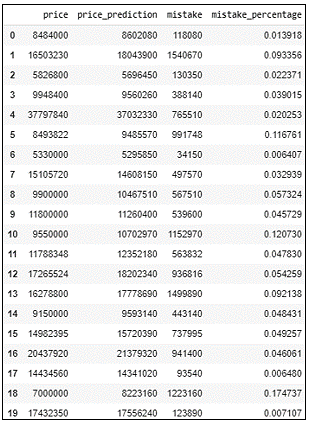


Рисунок 3.24 – Результат работы модели массовой оценки квартир нового рынка г. Москвы

Таким образом, была построена модель массовой оценки недвижимости в городе Москве, объединяющая в себе два ансамблевых алгоритма – случайный лес и градиентный бустинг, которые показывают сравнительно хорошие результаты по отдельности. С помощью усреднения результатов этих методов удалось добиться повышения точности предсказания.

# **Заключение**

В результате выполнения выпускной квалификационной работы были рассмотрены основные аспекты оценки недвижимости, проанализированы возможности использования методов машинного обучения для построения модели массовой оценки недвижимости. Применение методов машинного обучения заключается в проведении серии экспериментов, с целью анализа полученных результатов и сравнения их с реальными данными.

В качестве источника данных об объектах недвижимости был использован интернет-сервис по размещению объявлений Avito. Данные о центрах влияния были получены с помощью поисково-информационной картографической службы Яндекс.Карты. На основании региональных особенностей рынка недвижимости в городе Москве были отобраны факторы, влияющие на стоимость объектов.

Для предсказания стоимости применялись два ансамблевых алгоритма – случайный лес и градиентный бустинг, которые показывают сравнительно хорошие результаты по отдельности. С помощью усреднения результатов этих методов удалось добиться повышения точности предсказания.

Результатом выполнения выпускной квалификационной работы является модель массовой оценки недвижимости в городе Москве, с использованием которой можно произвести оценку рыночной стоимости большого количества объектов недвижимости. Данная модель позволит автоматизировать процесс оценки объектов недвижимости, сократить сроки проведения оценки при одновременном повышении её качества. Кроме того, за счёт использования разработанной модели унифицируется методология и подходы к оценке.

Все задачи данной выпускной квалификационной работы выполнены. Поставленная цель достигнута.