МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Филиал федерального государственного бюджетного образовательного учреждения высшего образования «Южно-Уральский государственный университет»
(национальный исследовательский университет)»
Высшая школа экономики и управления
Кафедра «Информационные технологии в экономике»

ПРОЕКТ ПРОВЕРЕН
Рецензент

_____ 2019 г.

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий кафедрой, д.т.н., с.н.с.

_____ Б.М. Суховилов
_____ 2019 г.

Создание модели прогнозирования стоимости жилой недвижимости с использованием методов машинного обучения

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

ЮУрГУ–38.04.05.2019. ПЗ ВКР

Руководитель работы, д.т.н., профессор

_____ В.В. Мокеев
_____ 2019 г.

Автор работы
студент группы ЗЭУ-356
_____ А.О.Казначевский
_____ 2019 г.

Нормоконтролер, к.т.н., доцент

_____ Е.В. Бунова
_____ 2019 г.

Челябинск 2019

АННОТАЦИЯ

Казначевский А.О. Создание модели прогнозирования стоимости жилой недвижимости с использованием методов машинного обучения. – Челябинск: ЮУрГУ, ЗЭУ-356, 2019. 85 с., 16 ил., 1 табл., 1 прил.

Данная работа посвящена актуальной теме, а именно, созданию модели прогнозирования стоимости жилой недвижимости с использованием методов машинного обучения, с последующей коммерциализацией проекта в виде мобильного приложения, предоставляющего возможность расчета стоимости жилой недвижимости.

Цель исследования: Создание модели прогнозирования стоимости жилой недвижимости с применением методов машинного обучения.

Задачи исследования:

1) описать рынок жилой недвижимости РФ;

2) проанализировать методы оценки недвижимости;

3) описать методы машинного обучения и сфер успешного их применения;

4) описать используемые средств разработки;

5) построить модель прогнозирования стоимости для объектов жилой недвижимости;

6) описать коммерциализацию проекта;

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Поскольку все большее число граждан и предприятия вовлекается в процессы на рынке недвижимого имущества, появляется необходимость определенно охарактеризовать свойства недвижимого имущества как особого вида товара. Одним из таких свойств является стоимость. У большого количества граждан нашей страны, принадлежащие им жилое имущество или земля являются самой ценной долей частного владения.

Рынок жилой недвижимости Российской федерации представляет собой сложную структуру, состоящую из миллионов квартир, характеризующихся множеством признаков. При этом любые изменения на рынке могут стать поводом для спекуляций и преднамеренного увеличения цен на недвижимость. Поэтому так важно понимать, где реальная стоимость квартиры, а где завышенная. В большинстве случаев самостоятельная оценка стоимости невозможна, привлекаются специалисты со стороны, что ведет к затратам денег и времени.

Экономическая ситуация последней пары лет сказалась на российском рынке не лучшим образом: снизилась активность, сократились обороты. Нужно искать способы точной оценки, как со стороны продавца, так и со стороны покупателя. Именно поэтому крайне важно независимое, быстрое и точное знание о ценах на рынке жилой недвижимости. Задачи определения и прогнозирования цены решаются достаточно трудоемким способом, требующим обработки большого количества данных.

На данный момент в распоряжении крупных компаний и банков находятся огромные массивы данных о клиентах и покупках. Анализируя эти данные, можно найти пути решения указанных задач. Однако экспертам-аналитикам ручные процедуры анализа сотен и тысяч показателей могут показаться пугающе длительными и трудоемкими, если вообще осуществимыми. И тут на помощь приходят технологии машинного обучения.

Объектом исследования данной работы выступает рынок жилой недвижимости РФ.

Предметом исследования является использование методов машинного обучения для прогнозирования стоимости объектов жилой недвижимости.

Цель данной работы - создание модели прогнозирования стоимости жилой недвижимости, созданной с применением методов машинного обучения.

Для достижения цели необходимо решить следующие задачи:

- дать описание рынка жилой недвижимости РФ;
- описать методы оценки недвижимости;
- описать методы машинного обучения и сфер успешного их применения;
- описать используемые средств разработки;
- построить модель прогнозирования стоимости для объектов жилой недвижимости;
- описать коммерциализацию проекта;

Практическая значимость результатов исследования состоит в том, что в результате работы мы получим актуальную модель прогнозирования стоимости для объектов жилой недвижимости, которую можно использовать для определения стоимости жилого имущества РФ.

1 ОБОСНОВАНИЕ НЕОБХОДИМОСТИ И МЕТОДЫ ПРОГНОЗИРОВАНИЯ СТОИМОСТИ ЖИЛОЙ НЕДВИЖИМОСТИ НА РЫНКЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

1.1 Характеристика рынка недвижимости Российской федерации

Подъем рынка жилой недвижимости поддерживает высокий уровень жизни людей и способствует улучшению демографического положения в стране.

Возмещение потребностей населения в оборудованном жилье становится непременным фактором эффективности проведения экономической и социальной частей метода воздействия государства на рыночную составляющую.

Смысловая база нынешней политики жилищного сектора была определена во время системного изменения экономики России, когда образование рентабельного рынка жилой недвижимости было определено первой из актуальных задач, что нашло проявление в разработке и применении федеральных и региональных целевых программ. В ведение региональных властей был переведен большой объем функций государства в жилищной сфере, среди которых – инвестирование целевых программ в регионах, создание региональных нормативно-правовых актов, проектирование главных направлений застройки территорий, организация принципиальных характеристик для строительства жилой недвижимости, направления расположения жилья и др.

Несмотря на все выше сказанное, проводимые государством действия в урегулировании жилищной проблемы, состояние в рассматриваемой области на уровне регионов остается достаточно напряженной, жилье, как и прежде имеют возможность приобрести группы населения с высоким и средним доходом. Это происходит потому, что вмешательство государства, обычно, не учитывает принципов развития национальной экономики и характеристику регионов, что уменьшает рентабельность рынков жилой недвижимости в функциональном плане и нуждается в поиске новых направлений воздействия, учитывающих индивидуальность различных типов регионов РФ.

Грабовский П.Г. дает определение понятия недвижимого имущества: «Недвижимое имущество — это объект гражданских прав, зарегистрированное согласно определенным законам. Гражданский оборот недвижимого имущества связан со сложным комплексом регистрации прав обладателя государственными органами, и их возможного перехода другому собственнику» [1].

Рынок недвижимости – это комплекс взаимоотношений, непосредственно связанных с операциями проводимыми с субъектами недвижимого имущества (продажи, покупки, оценки, аренды, залога и т. п.) [2].

Главными характерными чертами рынка недвижимого имущества являются: локальный характер рынка недвижимого имущества;

- уникальность каждого земельного участка (различие в ценах);
- невысокий показатель ликвидности недвижимого имущества в сравнении с другими товарами;
- несоответствие завышенных цен и денежных возможностей потенциальных приобретателей, что требует в большинстве случаев привлечения к этому кредита.

Главными субъектами рынка недвижимого имущества наряду с пользователями и собственниками недвижимости (а ими могут быть и юридические лица и физические) являются: вкладчики инвестиций, коммерческие банки, компании по строительству (подрядчики), фирмы-риэлторы, фирмы юридического направления сферы деятельности, агентства по рекламе, компании предоставляющие страховые услуги, комитеты по управлению имуществом, бюро технической инвентаризации, арбитражный суд, конторы предоставляющие нотариальные услуги, инспекции по налоговым сборам.

Рынок недвижимого имущества можно охарактеризовать как комплекс экономических взаимоотношений, посредством которых через ритмичность сил предложения и спроса на определенной территории проводится передача прав собственности и связанных с ней интересов непосредственно от продавца к покупателю.

Передача происходит через институт посредничества (риэлтор, девелопер и др.), подсчитываются и устанавливаются цены и проводится деление пространства между конкурирующими вариантами пользования объектами в границах некоторого замкнутого местного образования (город, населенный пункт) [3].

Принимая во внимание российский опыт образования риэлтерских услуг можно считать, что рынок недвижимого имущества – это механизм, посредством которого соединяются права и интересы сторон, фиксируются цены на недвижимое имущество. Таким образом, рынок недвижимого имущества, а, и как следствие, и риэлтерских услуг – это доля общего рынка, которая дает возможность для взаимодействие юридических или физических лиц, с целью обмена прав на недвижимое имущество которые находятся у них во владении, на денежные средства, либо на другие активы.

Рынок недвижимого имущества, также как базовый элемент этого рынка риэлтерские услуги, представляет собой предпринимательскую деятельность на рынке недвижимого имущества, осуществляя разные сделки с недвижимостью и правами на нее. Это свидетельствует о том, что на рынке недвижимого имущества:

– продавцы и покупатели ведут себя разумным образом, но не владеют абсолютными знаниями. Это означает, что все участники рынка собирают необходимую информацию об характерных особенностях сделки перед тем как принять окончательное решение;

– продавцы и покупатели проводят действия в не зависимости друг от друга, т. е. они проводят действия без сговора или мошенничества. В противном случае некоторые цены сделок могут быть сильно искажены.

Экономические функции недвижимого имущества характеризуются такими параметрами, как затраты на содержание, полезность, ликвидность, доходность, стоимость, цена, товар, налогообложение, вложение финансовых средств, спрос и предложение и др.

Любой объект может обладать стоимостью, имея в той или иной мере такие характеристики, как пригодность для эксплуатации и ограниченный характер предложения.

Ограниченность предложения – важная и неотделимая предпосылка образования стоимостной характеристики любых товаров. Социальные идеалы и стандарты, деятельность в экономическом плане и динамики, правовые акты, решения и действия правительственных органов, природные силы все это влияет на поведение населения, и все эти параметры, взаимодействуя между собой, создают, стабилизируют или изменяют стоимость недвижимого имущества.

Физические характеристики объекта недвижимого имущества включают такие данные как его размер и форма, о путях подъезда к нему, услугах коммунального характера, подпочвенный слой и поверхность, ландшафт и пр. Комплексность данных характеристик свидетельствуют о качественных составляющих недвижимого имущества, т.е. материальность и физические характеристики объекта недвижимого имущества измеряют его качественные характеристики. Качество в свою очередь выявляет полезность объекта, которая и составляет базу для расчета ценности недвижимого имущества [4].

Недвижимое имущество является одним из немногих товаров, ценностная составляющая которых не только практически всегда стабильна, но и имеет динамику к постепенному росту по прошествии времени.

Ценообразующая функция – одно из главных функциональных действий рынка – установление равновесных цен, при которых платежеспособный спрос будет соответствовать объему предложений. При ценах ниже равновесных будет иметь место избыточный спрос, а в ситуации превышения равновесных цен – избыточные предложения. В ценовых характеристиках концентрируются и высокие показатели объема информационных данных о насыщенности рынка, предпочтениях приобретателей, затраченных финансовых средствах на возведение (строительство), хозяйственной и социальной политике государственных органов в сфере жилищного строительства и пр.

Коммерческая функция заключается в формировании стоимости и потребительной стоимости недвижимого имущества, и получении дохода на вложенные средства.

Информационная функция – это уникальный способ оперативного сбора и распространения обобщенной объективной информации, которая дает возможность приобретателя и продавцам недвижимого имущества свободно, со знанием дела принять решение в своих интересах.

Функция посредничества выражается в том, что рынок выступает в качестве совокупного посредника и места встречи множества, независимых и обособленных в экономическом плане в результате общественного разделения труда продавцов и приобретателей, устанавливается связь между ними и предоставляется возможность альтернативного выбора партнеров. Действуют на рынке недвижимого имущества и профессиональные посредники: риэлторы, оценщики, агенты, брокеры, страховщики, ипотечные кредиторы и другие лица, оказывающие услуги заинтересованным участникам.

Функция вложения (инвестирования) – рынок недвижимого имущества, это привлекательный способ сохранения и увеличения стоимости капитала. Он дает возможность для перевода сбережений и накоплений людей из пассивной формы запасов в реальный производительный капитал, приносящий прибыль владельцу недвижимости. При этом сама недвижимость становится своеобразной страховой гарантией рисков денежных вложений.

Поскольку все большее число граждан и предприятия вовлекается в процессы на рынке недвижимого имущества, появляется необходимость определенно охарактеризовать свойства недвижимого имущества как особого вида товара. Одним из таких свойств является стоимость. У большого количества граждан нашей страны, принадлежащие им жилое имущество или земля являются самой ценной долей частного владения [5].

Рынок жилой недвижимости Российской федерации представляет собой сложную структуру, состоящую из миллионов квартир, характеризующихся

множеством признаков. При этом любые изменения на рынке могут стать поводом для спекуляций и преднамеренного увеличения цен на недвижимость. Поэтому так важно понимать, где реальная стоимость квартиры, а где завышенная.

Объективная оценка различных видов стоимости (рыночной, инвестиционной, залоговой, страховой, налогооблагаемой и других) недвижимого имущества необходима:

- при операциях купли-продажи или сдачи в аренду;
- при акционировании предприятий и перераспределении имущественных долей;
- при кадастровой оценке для целей налогообложения объектов недвижимости: зданий и земельных участков;
- для страхования объектов недвижимости;
- при кредитовании под залог объектов недвижимости;
- при ликвидации объектов недвижимости;
- при исполнении права наследования, судебного приговора;
- при других операциях, связанных с реализацией имущественных прав на объекты недвижимости.

Экономическая ситуация последней пары лет сказалась на российском рынке не лучшим образом: снизилась активность, сократились обороты. Нужно искать способы точной оценки, как со стороны продавца, так и со стороны покупателя. Именно поэтому крайне важно независимое, быстрое и точное знание о ценах на рынке жилой недвижимости. Задачи определения и прогнозирования цены решаются достаточно трудоемким способом, требующим обработки большого количества данных. Определение справедливой цены на квартиру — большая головная боль как для покупателей, так и для продавцов, потому что это очень многофакторная задача.

Изменчива специфика рынка недвижимости, обусловленная большим количеством влияющих признаков, в том числе и происходящих не только в Российской Федерации, а также большим количеством наблюдаемых объектов. Кто

из субъектов на рынке недвижимости может дать точную оценку стоимости? В большинстве случаев привлекаются оценщики со стороны, которые с помощью общепринятых методов оценивают стоимость.

Рассмотрим методы, с помощью которых можно оценить стоимость жилого недвижимого имущества.

1.2 Методы оценки стоимости жилой недвижимости

Три основных метода оценки рыночной стоимости недвижимости выделяют сегодня:

- метод сравнительного анализа продаж;
- затратный метод;
- метод капитализации доходов;
- экспертные методы прогнозирования.

Первый метод применяется тогда, когда именно рынок формирует цены, т.е. сформировавшийся рынок земли и недвижимости, существуют реальные продажи, и задача оценивающих заключается в том, чтобы проанализировать этот рынок, сравнить аналогичные продажи и таким образом получить стоимость оцениваемого объекта. Этот метод строится на сопоставлении рыночных аналогов с предлагаемым для продажи объектом. Он находит наибольшее применение на Западных рынках, т.е. где присутствует уже сформировавшийся рынок недвижимости и земли [6].

Таим образом этот метод используется при наличии достаточного количества именно достоверной рыночной информации о сделках купли-продажи объектов, являющихся аналогичными интересующему. Применяемый для выбора объектов критерий для является аналогичное наилучшее и наиболее эффективное использование.

Следующий метод подразумевает изучение возможностей инвестора в приобретении недвижимости и исходит из благоразумности инвестора, который проявляя эту благоразумность, сравнивает плату за объект и затраты на получение

соответствующего участка под застройку и строительство аналогичного по назначению и качеству объекта в обозримый период без существенных издержек

Условия, когда данный метод оценки приводит к объективным результатам, это возможность точно оценить величины стоимости и амортизации объекта при условии существующего равновесия спроса и предложения на рынке недвижимости. Этот метод демонстрирует оценку полной восстановительной стоимости объекта за минусом износа, увеличенную на рыночную стоимость земли.

Метод капитализации доходов построен на формировании стоимости объекта оценки с учетом возможных доходов от использования этого объекта, при этом используется текущая стоимость. Основными из методов этого подхода используются: метод прямой капитализации дохода и метод дисконтирования денежного потока. Первый метод возможно применить тогда, когда постоянным является прогнозируемый годовой чистый операционный доход и не наблюдается ярко выраженных тенденций к его изменению в неограниченном по времени периоде его получения. Метод дисконтирования денежного потока (непрямой капитализации) может быть использован в случаях, при которых прогнозируемые на протяжении выбранного периода имеются разные по величине и непостоянные денежные потоки от использования объекта оценки прогнозирования. Целесообразность этих методов зависит от наличия рыночных баз оценки стоимости объекта.

Экспертные методы прогнозирования являются отражение индивидуальных суждений специалистов относительно перспектив развития объекта и они основываются на профессиональном опыте и интуиции последних. Методы таких оценок могут быть использованы для анализа объектов и проблем, когда их развитие частично или полностью не поддается точной математической формализации.

Разработать адекватную для таких объектов модель трудно. Методы экспертной оценки, используемые в прогнозировании, подразделяются на коллективные и

индивидуальные. При этом, индивидуальные экспертные методы основываются на использовании независимых друг от друга мнений экспертов-специалистов соответствующего профиля. Наиболее часто применяются для выяснения таких мнений и формирования прогноза: интервью и аналитические экспертные оценки.

Интервьюирование предполагает беседу эксперта и прогнозиста, на протяжении которой прогнозист, имея заранее разработанную программу, задает эксперту вопросы относительно перспектив прогнозируемого объекта. Успех такой оценки в значительной степени обусловлен способностями опрашиваемого эксперта давать профессиональные заключения по различным фундаментальным вопросам без предварительной подготовки. Аналитические экспертные оценки должны предваряться длительной и тщательной самостоятельной деятельности эксперта над анализом имеющихся тенденций, оценкой состояний и путей развития того или иного объекта. При этом эксперт имеет возможность использовать всю необходимую ему информацию о прогнозируемом объекте. Все свои заключения эксперт оформляет в виде докладной записки.

В результате исследования любым методом на выходе имеется набор данных, который необходимо проанализировать с целью ответа на интересующие исследователя вопросы. На данный момент в распоряжении крупных компаний и банков находятся огромные массивы данных о клиентах и покупках. Анализируя эти данные, можно найти пути решения указанных задач. Однако экспертам-аналитикам ручные процедуры анализа сотен и тысяч показателей могут показаться пугающе длительными и трудоемкими, если вообще осуществимыми. Реализовать такие процедуры можно через методы машинного обучения. Одна из целей машинного обучения — свести к минимуму участие человека в однотипных, повторяющихся действиях.

Особенно много такой рутины — в сегменте оценки недвижимости: оценщику нужно прийти на объект, сделать фотографии, изучить план и техническую документацию, получить информацию о парковочных местах, вспомогательных помещениях, ремонте, состоянии технических систем, строительных материалах

и многом другом. Затем – проверить все данные, рассчитать оценочную стоимость, подготовить отчет и заключение [7]. Методы машинного обучения позволяют автоматизировать процесс решения подобных задач.

1.3 Машинное обучение, анализ тренда и основные понятия

Машинное обучение широко применяется во многих областях, которые, так или иначе, занимаются сбором и анализом данных. Большинство отраслей, работающих с большими объемами данных, признали ценность технологий машинного обучения. Построение точных моделей помогает выявлять решающие факторы и получать полезную информацию, которая может помочь предприятиям увеличить прибыль и избежать рискованных сделок, что повышает эффективность их работы и дает им преимущество перед конкурентами.

Алгоритмы машинного обучения существуют уже достаточно долгое время. В западных странах и США в практику принятия решений финансовых и инвестиционных активно внедряются новые интеллектуальные компьютерные технологий одновременно с развитием теоретических подходов для создания адекватных моделей поведения рынка недвижимости. Вначале это происходит в виде баз знаний и экспертных систем, а с конца восьмидесятых годов - в виде технологий машинного обучения, являющихся адекватным аппаратом для решения задач прогнозирования. В целом в последние годы рынок машинного обучения продолжает расти. Прогнозируется, что к 2021 году только в России объем рынка искусственного интеллекта в промышленности составит $380 млн. Направления развития определены аналитическим центром TAdviser, который опубликовал результаты исследования «Актуальные тенденции рынка машинного обучения и искусственного интеллекта».

Количество проектов, связанных с искусственным интеллектом и машинным обучением, выросло в мире в несколько раз за последние два года. В 2015 г. крупные компании сообщили о существовании 17 таких проектов, в 2016 г. – 71

проекта, в первой половине 2017 г. – уже о 74 проектах. Таким образом по итогам 2015-2017 гг. общее количество инициатив достигло 162. В их реализации участвуют 28 стран и 20 отраслей. К 2018г. 85% указанных проектов уже реализованы, еще 15% находятся на стадии планирования или пилота, причем в государственном секторе на этой стадии прибывает 60% инициатив. В 85% случаев проекты осуществляются по заказу крупного бизнеса.

США определен как лидер по количеству внедрений технологий искусственного интеллекта и машинного обучения. Следом Великобритания, применяющая технологии ИИ в крупных инвестиционных банках, и работающая на иностранных заказчиков Индия[8].

Российский рынок искусственного интеллекта и машинного обучения только начинает развиваться, демонстрируя отставание от зарубежных рынков. На Российском рынке недвижимости технологии машинного обучения появились всего несколько лет назад. В России нет примеров, подтверждающих положительное влияние подобных технологий на бизнес-процессы. Вероятно, причина в том, что компании, не желая раскрывает источники конкурентных преимуществ, не разглашают результаты успешных внедрений. Другая причина медленного внедрения искусственного интеллекта и машинного обучения в России - недостаток вычислительных мощностей и невысокий уровень автоматизации.

Рынок недвижимости консервативен, но новые игроки обнаружили на нем интересные ниши, где можно упростить и улучшить неэффективные бизнес-процессы. Большинство онлайн-сервисов работают над тем, чтобы заменить традиционных посредников – риелторов – и доверить их работу компьютеру, однако на рынке сейчас практически нет возможности получить оценку жилого имущества без привлечения экспертных оценщиков.

Такие эксперты используют классические методы оценки, что означает высокие сроки получения оценки, совершения ряда рутинных операций, и относительно высокую стоимость подобной услуги, поскольку в цену заложены все вышеперечисленные действия. Таким образом, проблемой мы можем назвать

отсутствие возможности получить независимую, своевременную и точную информацию о стоимости жилой недвижимости без привлечения значительных средств и временных затрат. Что собой представляет машинное обучение?

По свой сути машинное обучение – это развитие методов аппроксимации функции, но в качестве точек выступают более сложные объекты, элементы сложно-описанных пространств, а ответами могут быть не только числа, но и множества [9].

Задачами машинного обучения является нахождение неизвестной зависимости между известным множеством объектов и множеством ответов. Таким образом, мы нуждаемся в построении функции, которая бы с достаточной точностью приближала значения множества ответов в точках множества объектов и на всем остальном пространстве.

Объектами для решения таких задач могут быть веб-страницы, страны, люди, изделия, фирмы, т.е. все, что несет какую-либо информацию или имеет какой-то набор признаков, под которыми мы понимаем способы измерения характеристик объектов в исследуемом пространстве. В зависимости от множества допустимых значений все признаки делятся на:

- бинарные;
- номинальные;
- порядковые;
- количественные;
- качественные.

В свою очередь, под задачами машинного обучения понимают разделения задач на различные типы.

В зависимости от природы множества допустимых ответов Y задачи обучения по прецедентам делятся на следующие типы.

Если $Y=\{1,...,M\}$, то это задача классификации на M непересекающихся классов. В этом случае всё множество объектов X разбивается на классы $Ky=\{x \in X:(x) =y\}$, и алгоритм a(x) должен давать ответ на вопрос к какому классу

принадлежит x [10]. В некоторых приложениях классы называют Образами и говорят о задаче распознавания образов.

Если Y={0,1}M, то это задача классификации на M пересекающихся классов. В простейшем случае эта задача сводится к решению M независимых задач классификации с двумя непересекающимися классами [10].

Если Y=R, то это задача восстановления регрессии. Задачи прогнозирования являются частными случаями классификации или восстановления регрессии, когда x∈X описание прошлого поведения объекта x, y∈Y описание некоторых характеристик его будущего поведения [10] .

Модель алгоритмов (модель зависимости) – параметрическое семейство отображений, A={g(x,θ)|θ∈Θ}, где g:XΘY некоторая фиксированная функция, Θ множество допустимых значений параметра, называемое пространством параметров или пространством поиска в котором нужно найти функцию, хорошо приближающую целевую зависимость, функциональную или стохастическую зависимость между объектами и ответами. Задача нахождения модели зависимости сводится к построению алгоритма, который бы одинаково точно приближал неизвестную целевую зависимость, как на элементах выборки, так и на всем пространстве объектов. Это задача получила название – обучение с учителем.

Исходя из законов исследуемой области, например, физической модели можно по-разному строить зависимость параметров. В самом простом случае используется линейная модель.

Метод обучения – отображение, которое ставит в соответствие обучающей выборке алгоритм из заданной модели. Другими словами, алгоритм строится по выборке.

Процесс обучения подразделяется на два этапа:

- этап обучения – на этапе выявляется зависимость в эмпирических данных;
- этап применения – выявленная зависимость проверяется на точность.

На этапе обучения для того, чтобы выявить зависимость необходимо использовать обучающую выборку, и уже по ней оптимизировать параметры.

Чтобы оценить, насколько модель качественна необходимо использовать тестовую, или контрольную, выборку. Чтобы предотвратить смещение оценки, обучающая и контрольная выборки должны быть независимыми. Чтобы выбрать наилучшую модель из множества моделей, построенных на обучающей выборке, используется проверочная выборка. Важным условием является то, чтобы обучающая выборка обладала достаточной полнотой, т.е. должна покрывать все возможные случаи.

Чтобы оценить алгоритм, а точнее его применение на выборке, необходимо сначала произвести оценку применения алгоритма на отдельном объекте. Для того, чтобы формализировать величину ошибки алгоритма на объекте используется понятие – функция потерь. Функция потерь характеризует величину отклонения ответа модели от правильного ответа на произвольном объекте.

Эмпирический риск – функционал алгоритма на обучающей выборке, характеризует качество приближения заданной функции на выборке и равна среднему значению потерь (сумма значений функции потерь по обучающей выборке, деленная на длину выборки). Другими словами, эмпирический риск– средняя величина ошибки алгоритма на обучающей выборке. Таким образом, задачи обучения сводятся к задачам оптимизации, нахождению функций приближения с минимальным значением функции потерь, т.е. сводить задачи к численным методам оптимизации.

Обучающая способность алгоритма определяется соотношением точности ответов на тестовой и контрольной выборках. Если значения ошибок на тестовой выборке малы и близки к значению ошибок на обучающей выборке, то можно говорить о хорошей обученности алгоритма, иначе мы сталкиваемся с такими явлениями, как недообучение и переобучение.

Недообучение возникает, если мы наблюдаем большую вероятность и величину значений ошибки на обучающей выборке, и это говорит о недостаточной сложности моделей.

Переобучение можно наблюдать, когда вероятность ошибки на объектах тестовой выборки существенно выше вероятности на обучающей, это говорит, что использовались чрезмерно сложные модели.

Переобучение возникает из-за избыточной сложности пространства параметров, лишние степени свободы в модели «тратятся» на чрезмерно точную подгонку под обучающую выборку. Переобучение есть всегда, когда есть оптимизация параметров по конечной (заведомо неполной) выборке

Совсем избавиться от переобучения нельзя, можно его лишь минимизировать.

1.4 Методы машинного обучения

Рассмотрим методы машинного обучения, которые мы будем использовать в своем исследовании. Существует большое множество методов машинного обучения. Для построения модели прогнозирования стоимости жилой недвижимости воспользуемся методами ансамбля моделей. Под обучением ансамбля моделей понимается процедура обучения конечного набора отдельных классификаторов, результаты прогноза которых затем мы объединим и сформируем прогноз (агрегированного) общего классификатора. Ансамбли моделей создаются с целью повышения точности прогноза общего классификатора при сравнении с точностью прогноза каждого отдельного основного классификатора.

Интуитивно понятно, что комбинирование основных классификаторов даст более точный результат, чем каждый отдельный классификатор, если базовые классификаторы с одной стороны достаточно точны, а с другой – если они дают разные результаты, т.е. ошибаются на разных множествах [11].

На основании общих рассуждений можно выделить три причины, по которым объединение классификаторов может быть успешным:

1. Статистическая причина. Классификационный алгоритм можно рассматривать как процедуру поиска в пространстве гипотез H о распределении данных с целью поиска наилучшей гипотезы. Обучаясь на конечном наборе данных, алгоритм может найти множество различных гипотез одинаково

хорошо описывающих обучающую выборку. Строя ансамбль моделей, мы «усредняем» ошибку каждой индивидуальной гипотезы и уменьшаем влияние нестабильностей и случайностей при формировании гипотез.

2. Вычислительная причина. Большинство обучающих алгоритмов используют методы нахождения экстремума некой целевой функции. Например, нейронные сети используют методы градиентного спуска для минимизации ошибки прогноза, деревья решений – жадные алгоритмы роста дерева, минимизирующие энтропию данных и т.д. Эти алгоритмы оптимизации могут «зависнуть» в точке локального экстремума. Ансамбли моделей, поскольку комбинируют результаты прогноза базовых классификаторов, обученных на различных подмножествах исходных данных, имеют больший шанс найти глобальный оптимум, так как ищут его из разных точек исходного множества гипотез.

3. Репрезентативная причина. Комбинированная гипотеза может не находиться в множестве возможных гипотез для базовых классификаторов, т.е. строя комбинированную гипотезу, мы расширяем множество возможных гипотез.

Одним из наиболее распространенных подходов к формированию ансамбля моделей является манипулирование обучающим множеством с последующим построением базовых классификаторов на различных его подмножествах. Каждый базовый классификатор использует один и тот же алгоритм, но обучается на различных данных. Затем прогноз ансамбля производится комбинированием результатов прогноза каждого отдельного классификатора при помощи (взвешенного) усреднения для непрерывной целевой переменной или (взвешенного) голосования – для дискретной. Такой подход особенно успешен, когда базовый алгоритм является нестабильным, т.е. даёт ощутимо различный результат при небольшом изменении данных в обучающей выборке. Примером нестабильного алгоритма может служить деревья решений, в процессе построения которых небольшие изменения данных могут послужить причиной разбиения

узлов дерева по разным атрибутам и границам их значений [12]. Мы рассмотрим два алгоритма, использующих этот подход: Bagging и Boosting.

В алгоритме Bagging исходные данные случайно разбиваются на одинаковые по размеру подмножества, каждое из которых используется для обучения одного отдельного классификатора. Прогноз ансамбля определяется большинством голосов или средним.

Алгоритм Boosting итеративно учится распознавать примеры на границах классов. Каждой записи данных на каждой итерации алгоритма присваивается вес, который соответствует, например, вероятности попадания этой записи в обучающую выборку на следующей итерации или соответствует частоте, с которой эта запись будет «размножена» в обучающей выборке на следующей итерации. Первый базовый классификатор обучается на всех данных с равномерными весами. Затем, на каждой последующей итерации, веса правильно классифицированных примеров уменьшаются, а неправильно классифицированных – увеличиваются. Следующий классификатор, таким образом, будет «уделять больше внимания» неправильно классифицированным примерам, т.е. все больше учиться исправлять ошибки классификатора на прошлой итерации. Прогноз ансамбля представляет собой взвешенное голосование прогнозов базовых классификаторов.

Вес, с которым учитывается результат прогноза каждого базового классификатора при голосовании, соответствуют точности его прогнозирования.

Рассмотрим подробнее методы ансамблей моделей:

1.4.1 Bagging

Bagging (Bootstrap Aggregating, Leo Breiman) – улучшающее объединение. Исходные данные D, состоящие из N строк, разделяются на t подмножеств D1 ,…, Dt с тем же числом строк в каждом при помощи равномерной случайной выборки с возвратом. Затем, t базовых классификаторов, используя один и тот же алгоритм, обучаются на этих подмножествах. Результаты прогнозирования базовых классификаторов усредняются или выбирается класс на основании большинства голосов [13]. Схема работы представлена на рисунке 1.1.

Каждая запись из D имеет одинаковую вероятность быть выбранной в множество Di, т.е. 1/N, значит, вероятность не быть выбранной равна (1- 1/N). Так как каждое из множеств Di формируется независимо друг от друга, то вероятность того, что определенная запись из D не попадет ни в какое Di, будет равна (1- 1/N)t, следовательно, вероятность, что запись хотя бы раз будет выбрана, равна 1- (1-1/N)t.

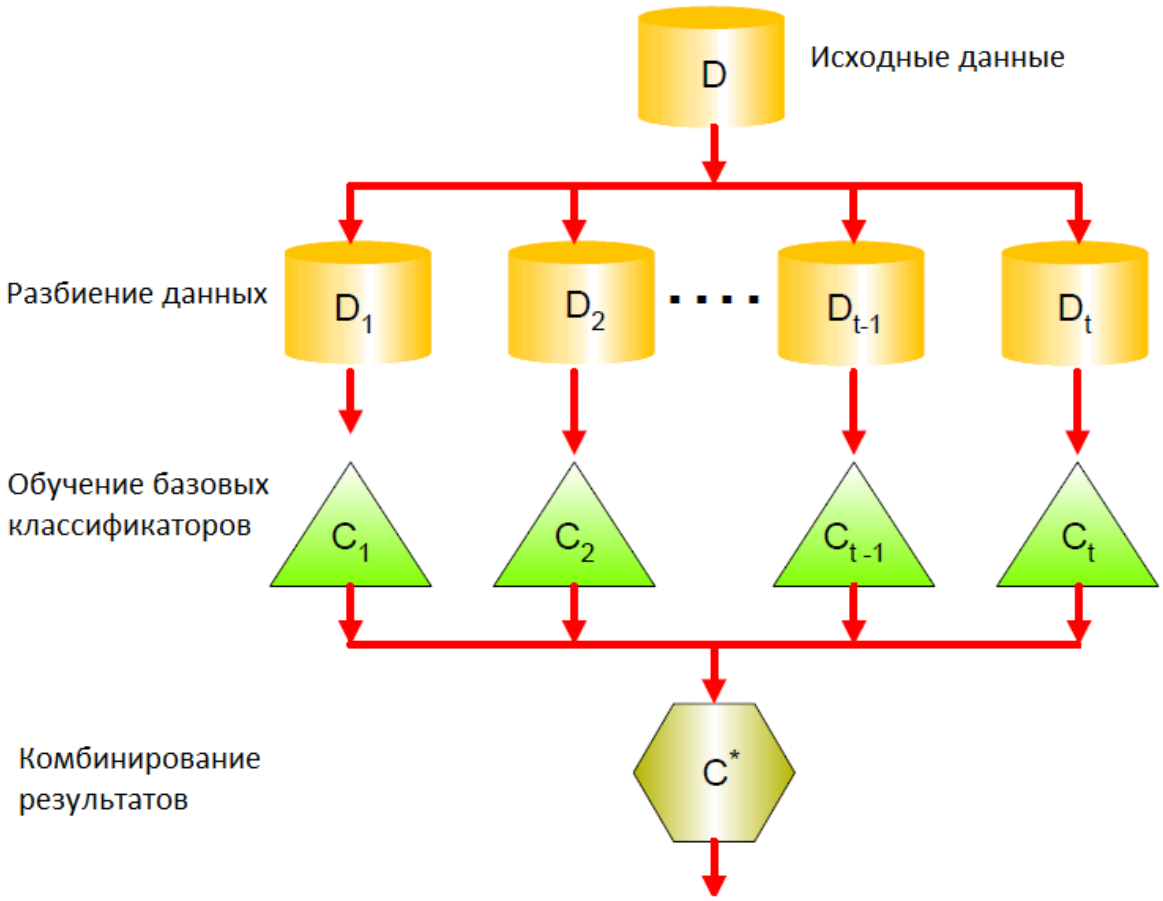

Рисунок 1.1 – Схема работы метода Bagging

На практике Bagging хорошо улучшает точность результатов (5-40% если использовать в основе алгоритм CART) по сравнению с отдельными основными классификаторами классификаторами если алгоритм, используемый базовым классификатором достаточно точен, но нестабилен. Точность прогноза повышается из-за того, что уменьшается разброс нестабильных прогнозов отдельных классификаторов, т.е. уменьшается дисперсия.

Преимущество алгоритма Bagging в том, что его просто реализовать, а также тем, что возможно распараллеливание вычислений по обучению каждого основного классификатора по различным вычислительным узлам. Недостатками являются:

− отсутствие строгого математического обоснования условий улучшения прогноза ансамбля;

− недетерминированность результата (обучающие выборки формируются случайно);

− сложность интерпретации результатов по сравнению с индивидуальными моделями.

Одним из самых популярных способов реализации метода Bagging является Random Forest.

1.4.2 Boosting

В алгоритме Boosting базовые классификаторы обучаются последовательно, а не параллельно, как в алгоритме Bagging. Набор данных, на которых обучается каждый последующий базовый алгоритм в Boosting, зависит от точности прогнозирования предыдущего базового классификатора. Одним из важных понятий в Boosting – это вес строки данных, который обновляется после каждого выполнения обучения основного классификатора, т.е. итерации. Значение веса строки – это важность ее для обучения следующего основного классификатора и основывается на величине значения ошибки предыдущего классификатора на этой строке. Вес строки можно объяснить, как вероятность выбора этой строки для обучения следующего классификатора или как число, пропорциональное относительной доли этой строки в следующей обучающей выборке. Второй подход представляется предпочтительней, так как является детерминистическим (мы «размножаем» строки для обучения на следующей итерации пропорционально их весу) [14]. Различные модификации алгоритма Boosting имеют одинаковую структуру:

1. Присвоить всем N строкам обучающей выборки одинаковые веса 1/N.

2. В цикле от m=1 до M.

2.1. Обучить базовый классификатор fm на данных, распределение строк в которых соответствует весам этих строк.

2.2. Вычислить взвешенную ошибку прогноза fm.

2.3. Пересчитать веса каждой строки в соответствии с ошибкой прогноза на этой строке: для правильно классифицированных строк вес уменьшается, для неправильных – увеличивается.

3. Прогноз ансамбля вычисляется по (взвешенному) среднему или (взвешенному) большинству прогнозов базовых классификаторов. В случае использования весов, их значения больше у более точных базовых классификаторов.

Схема работы представлена на рисунке 1.2.

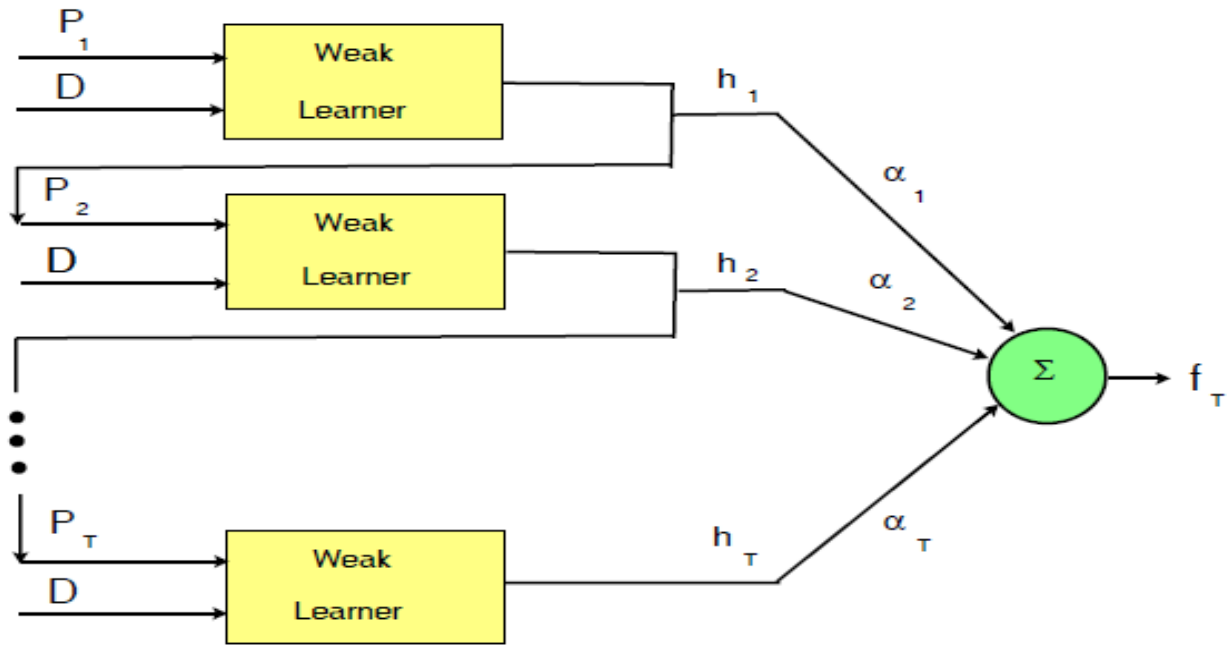

Рисунок 1.2 – Схема работы метода Boosting

Непосредственно из описания шаблона алгоритма Boosting видны его недостатки:

1. Так как на каждой итерации алгоритм уделяет все большее внимание неправильно классифицированным строкам, то в случае зашумления данных,

алгоритм на больших итерациях будет полностью сконцентрирован на попытках получить прогноз на имеющихся ошибочных записях, а не на выявлении объективных закономерностей.

2. Последовательность выполнения обучения базовых классификаторов не позволяет использовать распараллеливание вычислительных мощностей.

3. Результаты классификации ансамбля трудны для интерпретации.

Однако алгоритм Boosting обладает существенными преимуществами по сравнению с другими алгоритмами, работающими с ансамблями моделей:

1.Алгоритм Boosting основывается на вероятностной модели и является аналогом хорошо исследованной логистической регрессии.

2. Каждая итерация любой модификации алгоритма Boosting представляет собой шаг по минимизации функции взвешенного штрафа ошибки или шаг по максимизации функции условного ожидаемого правдоподобия данных. Экстремумы этих функций являются истинными значениями классификационной функции модели. Т.е. алгоритм Boosting имеет статистическое обоснование.

3. Ошибка классификационной функции ансамбля на обучающей выборке при достаточно слабых условиях на точность базовых классификаторов стремится к нулю при увеличении числа итераций.

4. Boosting уменьшает не только ошибку разброса, но и ошибку смещения в терминологии равенства.

Все это делает Boosting наиболее теоретически исследованным и обоснованным методом из существующих.

Одним из самых популярных на сегодняшний день механизмов реализации метода Boosting является механизм XGBoost.

1.5 Анализ степени изученности темы

Основополагающие научные разработки в области построения моделей прогнозирования стоимости жилого имущества с помощью методов машинного

обучения представлены трудами таких авторов, как Ritchie Ng, Gul Md Ershad, Neha Chanu, Fatima Hamdan, Lainey Liu, Andrew Caplin, Sumit Chopra, John Leahy, Yann LeCun, Trivikrmaman Thampy, Свирчков Д.В., Стерник Г.М., Стерник С.Г.,Свиридов А.В., Левченко В.В.· Их труды представлены работами «Machine Learning for a London Housing Price Prediction Mobile»[15], «Boston Home Prices Prediction and Evaluation»[16], «Machine Learning and the Spatial Structure of House Prices»[17], «Методология прогнозирования российского рынка недвижимости»[18], « Модель определения привлекательности инвестирования в сфере недвижимости с помощью методов машинного обучения»[19].

Изучение литературы на эту тематику показало, что в указанных источниках не найдено подробного описания эффективного построения модели прогнозирования стоимости жилой недвижимости с помощью методов машинного обучения для рынка Российской Федерации. Большинство публикаций являются описаниями возможностей машинного обучения и их потенциальных преимуществ перед другими способами решения подобных задач. Часть работ посвящена именно построению моделей, однако и модели, и выводы по работе были сделаны без учета особенностей рынка недвижимости Российской Федерации, и применимы только к западным рынкам.

Что касается практических реализаций методов машинного обучения в сфере жилой недвижимости, то можно привести несколько примеров:

- PurpleBricks;

Совершенно точно можно повысить эффективность работы риелторов, лучше выстроить процессы коммуникации с клиентами и устранить лишние расходы, как делает данная британская компания .

По сути, это цифровое-агентство недвижимости, использующее умные математические алгоритмы для автоматизации оценки стоимости и аудита объектов, оптимизации маршрутов передвижения агентов и других задач. Особенно интересно, что с продавца берется фиксированная комиссия,

а не процент от сделки, так что риелторам нет смысла навязывать покупателю более дорогой объект, чтобы больше на нем заработать [20].

- HomeApp;

В России похожий сервис предлагает проект HomeApp. Он использует ИИ, чтобы определить цену для выставления объекта на продажу. Система сама просматривает доски объявлений и анализирует аналогичные предложения на рынке. Затем она выделяет целевые группы клиентов-покупателей под каждую конкретную квартиру и им показывает таргетированную рекламу [21].

- Skyline AI;

Израильский стартап пытается занять место не столько риелтора, сколько инвестиционного советника. Разработанная им платформа анализирует рынок коммерческой недвижимости и составляет прогнозы динамики арендной ставки по конкретному объекту, его рыночной стоимости через несколько лет, подсказывает, когда нужно повысить арендную плату или запланировать вложения на капитальный ремонт. Для этого сервис собирает данные из 130 источников и анализирует около 10 тыс. характеристик каждого объекта за последние 50 лет. Сейчас платформа находится в стадии финального тестирования, но когда она «дозреет», станет мощным подспорьем инвестора [22].

- Bowery Valuation;

Компания из Нью-Йорка модернизирует эту нишу с помощью умного поиска информации и анализа больших данных. Она автоматически собирает из открытых источников всю необходимую информацию об объекте, считает рыночную стоимость по специальному алгоритму, а также генерирует типовой отчет с результатами [23].

- Мобильный оценщик;

Российский стартап берет на себя некоторые простые, но затратные по времени функции: обрабатывает загруженные фото в соответствии с шаблонными требованиями и помогает свести к минимум ручной ввод информации при

составлении отчета. Акцент при использовании смещен на автоматизацию рутинных процессов, нежели на предоставление информации [24].

- ЦИАН Онлайн калькулятор;

Продукт российских разработчиков, позволяющий использовать алгоритм оценки стоимости жилой недвижимости. Позволяет риэлторам и собственникам выставить приближенную к рынку цену при попытке продажи жилья, Оценивается недвижимость по аналогии с ценами объявлений последних лет. Как достоинство компания преподносит предварительную фильтрацию объявлений от дублей. В среднем для оценки каждой конкретной квартиры используется порядка 350 объявлений по аналогичным объектам [25].

1.6 Разработка модели исследуемого объекта

По ходу исследования мы приходим к выводу, что мы своим исследованием займём пустую нишу: подробно опишем построение модели для задачи прогнозирования стоимости жилой недвижимости, подробно опишем шаги анализа данных и опишем сравнение моделей к опирающихся на установленных особенностях в данных рынка Российской Федерации. Наша задача представляется собой задачу регрессии, и ее решение будет состоять из будет состоять из следующих этапов:

- выбор инструментов реализации;
- формирование данных для исследования;
- предобработка данных;
- первичный анализ данных;
- построение моделей различными методами;
- выводы по модели.

Выводы по первой главе:

В результате проделанной работы мы провели обзор актуальности исследуемой темы, дали обоснование важности рассматриваемого вопроса, убедились в важности знания стоимости на рынке жилой недвижимости, пришли к выводу о

необходимости машинного обучения в данной сфере и рассмотрели примеры успешно работающих в этой области проектов. Были сформулированы цели и задачи диссертационного исследования, определена актуальность темы. В процессе разработки были решены следующие задачи:

- произведено описание проблемы;
- проведен Анализ изученности темы;
- проведен обзор методов оценки стоимости жилой недвижимости;
- проведен обзор методов машинного обучения;
- определены этапы построения модели прогнозирования стоимости жилой недвижимости.

2 ПОСТРОЕНИЕ МОДЕЛИ ПРОГНОЗИРОВАНИЯ СТОИМОСТИ ЖИЛОЙ НЕДВИЖИМОСТИ

2.1 Выбор инструментов реализации

Концепция «больших данных» приобретает все большую популярность, и распространение получает раздел информатики, который называется наука о данных (data science). Изучает наука о данных как проблемы анализа, так и обработки и представления данных больших объемов [26].

Одними из самых распространенных инструментов для работы с большими данными на текущий момент являются R и Python.

1. Язык программирования R

R – язык программирования, предназначенный для статистической обработки данных и работы с графикой, но в то же время это свободная программная среда с открытым исходным кодом, развиваемая в рамках проекта GNU. R получил широкое распространение в сферах, где проводится работа с данными. Основная вычислительная мощь R заключается в статистическом анализе, однако он также обладает обширным функционалом для первичного анализа данных (построение графиков и таблиц сопряженности) и математического моделирования. Язык создавался специально для анализа данных: запись конструкций языка понятна многим специалистам в области [27].

Многие функции, необходимые для анализа данных, являются встроенными функциями языка. Проверка статистических гипотез зачастую занимает лишь несколько строк кода. К преимуществам R относятся:

- установка IDE (RStudio) и пакетов для работы упрощена до предела;
- удобный репозиторий пакетов и обилие готовых тестов практически под все методы Data Science и машинного обучения;
- эффективная работа с векторами и матрицами;
- несколько качественных пакетов для визуализации данных для различных задач (ggplot2, lattice, ggvis, googleVis, rCharts и т.д.);

К недостаткам R относят:

− низкая производительность. Однако в системе присутствуют пакеты, позволяющие повысить скорость работы (pqR, renjin, FastR, Riposte и т. д.). При работе с большими массивами данных рекомендуется использовать библиотеки data.table and dplyr;

− язык специфичен если сравнивать со стандартными языками программирования, так как язык узкоспециализированный (например, индексация векторов начинается вместо нуля с единицы);

− так как большая часть кода на R написана людьми, не знакомыми с программированием, иногда код слабочитаем. К рекомендациям по оформлению не всегда прислушиваются, что приводит к еще более слабой читаемости;

− R прекрасный инструмент для статистики и соответствующих stand-alone приложений, но слабо конкурирует там, где традиционно применяются языки общего назначения;

− синтаксис решения не всегда очевиден, поскольку есть возможность решить одну задачу разными методами;

− в силу большого количества библиотек, документация некоторых менее популярных из них нельзя считать полной.

2. Язык программирования Python

Python – высокоуровневый универсальный интерпретируемый язык сценариев. При разработке на языке Python большое внимание уделяется простоте и понятности синтаксиса, что не только сокращает время изучения его основ, но и повышает скорость разработки в целом [28]. Это далеко не все преимущества данного языка, основные из них:

− объектно-ориентированность;

− свободное распространение и широкая поддержка;

− кроссплатформенность;

− развитые функциональные возможности.

К достоинствам Python можно отнести:

− универсальный многоцелевой язык: можно найти данные, обработать их, и тут же применить результаты обработки в веб-приложении;

− Python отлично подходит в качестве первого для изучения языка программирования. Если изучающий потеряет интерес к науке о данным, то навыки использования языка пригодятся и в других областях.

К недостаткам Python можно отнести:

− нет общего репозитория, и нет такой специализации для обработки данных, как в R. Однако за последние несколько лет данный недостаток перестает быть актуальным: если раньше аналитикам приходилось использовать несколько языков для разных задач, то теперь инструментарий Python пополняется, и становиться никак не беднее конкурента. Использование языка облегчают также сборки, например Anaconda.

− Python – язык с динамической типизацией. Работа программы существенно ускоряется, но одновременно усложняется поиск ошибок.

В языке Python объединяются две парадигмы программирования – объектно-ориентированная, которая сама по себе мощное средство структурированного программного кода для многократного использования, и процедурная, которая позволяет расширить круг решения задач, разрешая использование средства Python при решении тактических задач с отсутствием фазы проектирования.

Объектная модель Python поддерживает понятия полиморфизма, перегрузки операторов и множественного наследования.

Кроссплатформенность языка достигнута благодаря его реализации на переносимом ANSI C, благодаря чему программы, написанные на языке Python, одинаково хорошо компилируются и выполняются на любых платформах, где устанавливают совместимую версию Python.

Python – это удобное средство разработки приложений самых разных типов благодаря простоте и удобству языков сценариев и мощности компилирующих языков. Такой эффект достигается благодаря его гибридной природе. Однако при решении задач автоматизации процессов и обработки данных эффективность

проявляется наилучшим образом. Python активно используют в исследовательских работах. В язык программирования Python заложен мощный встроенный инструментарий (встроенные типы объектов и динамическая типизация, автоматическое управление памятью), а также возможность использования внешних библиотек и утилит сторонних разработчиков при решении узкоспециализированных задач. Благодаря широкому распространению Python собрал вокруг себя активное сообщество разработчиков, которые в рамках различных проектов разрабатывают модули для узкоспециализированных задач.

Одним из таких проектов стал Google Summer of Code 2007, в рамках которого David Cournapeau разработал библиотеку scikit-learn [29]. Разработка данной бибилиотеки является одной из причин популяризации применения языка Python в области анализа данных с помощью методов машинного обучения.

Библиотека scikit-learn предоставляет реализацию ряда алгоритмов как для обучения с учителем (Supervised learning), так и для обучения без (Unsuoervised learning). Scikit-learn построена на основе стека SciPy (Scientific Python), который включает в себя:

- numPy [30] добавляет поддержку больших многомерных массивов и матриц, а также библиотеку высокоруровневых математических функций для операций с ними;
- sciPy [31] - открытая библиотека высококачественных научных инструментов для языка программирования Python;
- matplotlib [32]– библиотека для визуализации двумерной и трехмерной графики;
- jupyter Notebook [33]– интерактивная оболочка для языка программирования Python, которая предоставляет расширенную интроспекцию, дополнительный командный синтаксис, подсветку кода и автоматическое дополнение Sympy – библиотека для работы с символьными вычислениями;
- pandas [34]- реализует различные структуры данных и анализ;

Оба языка поддерживаются open-source лицензиями (в отличие от коммерческих инструментов SAS и SPSS или проприетарного MATLAB) и традиционно рассматриваются как наиболее востребованные. Анализируя рисунке 2.1 статистику Google Trends с января 2012-го года по январь 2018-го о запросах о применении Python и R в задачах машинного обучения можно наблюдать, что в течение последних пары лет интерес к Python стал преобладать и на начало 2018-го года по числу запросов Python примерно в 2 раза превышает запросы, связанные с R.

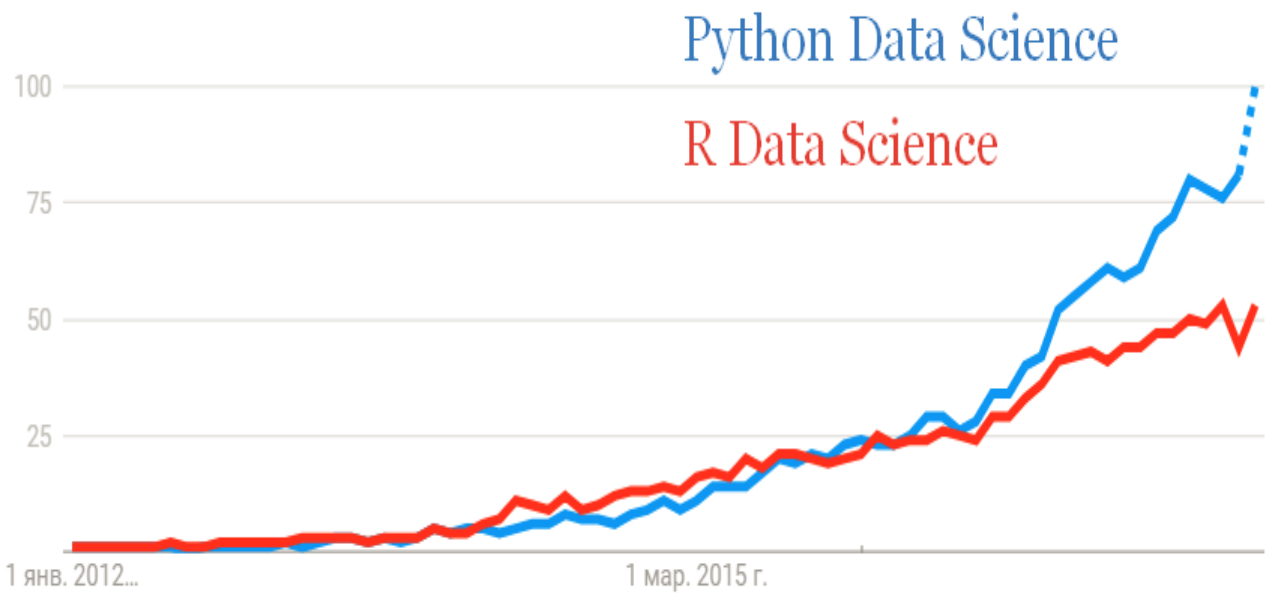

Рисунок 2.1 – Статистика применения языков R и Python

По данным сообщества kaggle.com на рисунке 2.2 показано, как распределяются решения задач с использованием языков Python и R на протяжении 2017 года [35]. На рисунке 2.2 мы можем наблюдать явное повышение количества решений задач с применением языка Python относительно языка R с течением времени.

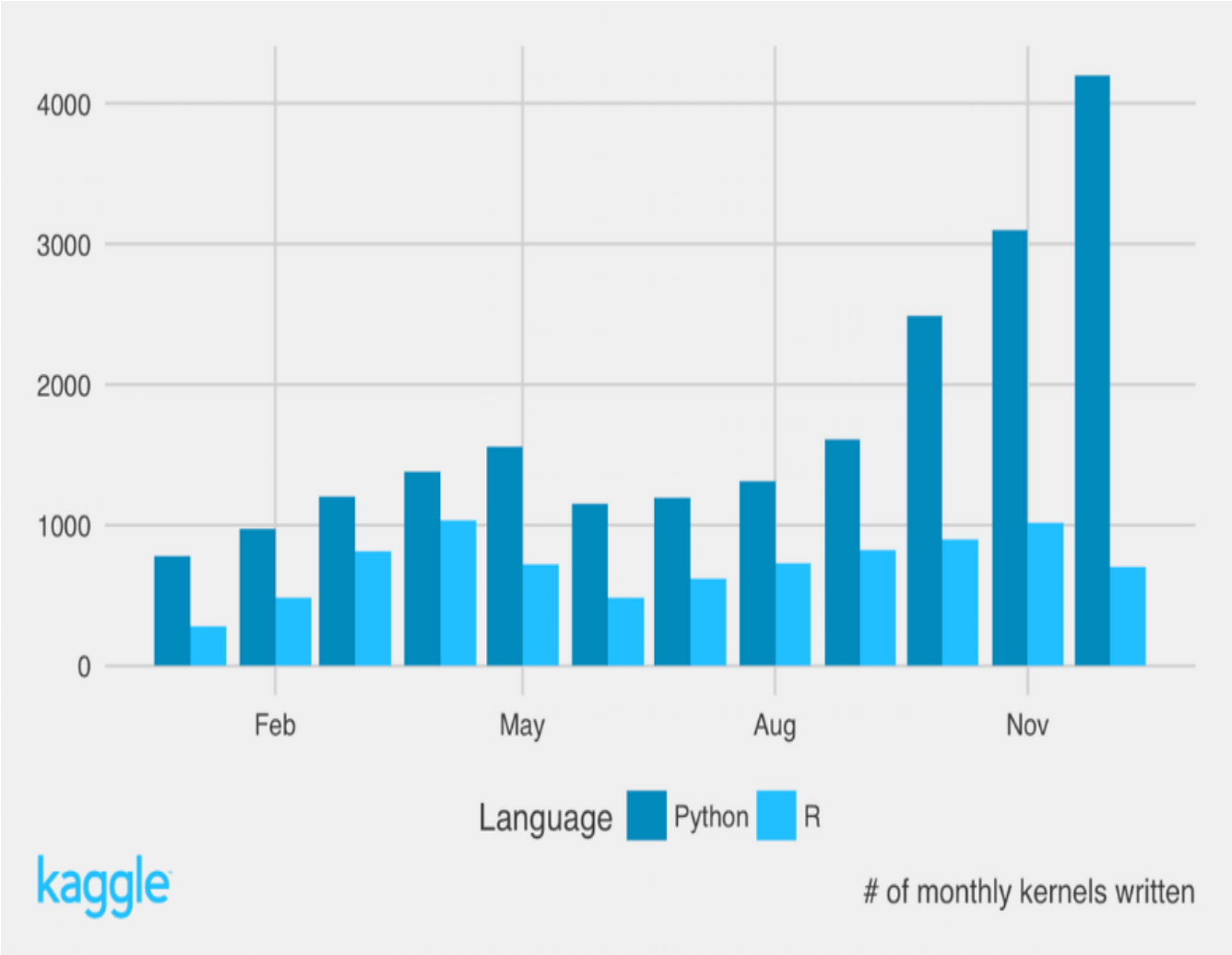

Рисунок 2.2 – Количества решений задач на языках R и Python на ресурсе Kaggle

На основании вышеизложенного, с учетом всех достоинств и недостатков, а также популярности применения, инструментом для построения нашей модели станет язык программирования Python.

2.2 Построение моделей

Наша цель – построить модель, которая сможет правильно оценить стоимость жилого недвижимого имущества на основе заданных значений признаков. Очевидно, что это задача регрессии: целевая переменная численная.

Также это задача обучения с учителем: целевая переменная явно определена в тренировочном наборе данных, и нам необходимо получить ее значения для каждой записи тестового набора. Задачу обучения по прецедентам при Y=R принято называть задачей восстановления регрессии. Задано пространство

объектов X и множество возможных ответов Y. Существует неизвестная целевая зависимость y∗:X→Y, значения которой известны только на объектах обучающей выборки Xℓ= (xi,yi)ℓi=1,yi=y∗(xi). Требуется построить алгоритм, который в данной задаче принято называть функцией регрессии a: X→Y, аппроксимирующий целевую зависимость y∗. В целом задачу регрессии можно наблюдать на рисунке 2.3

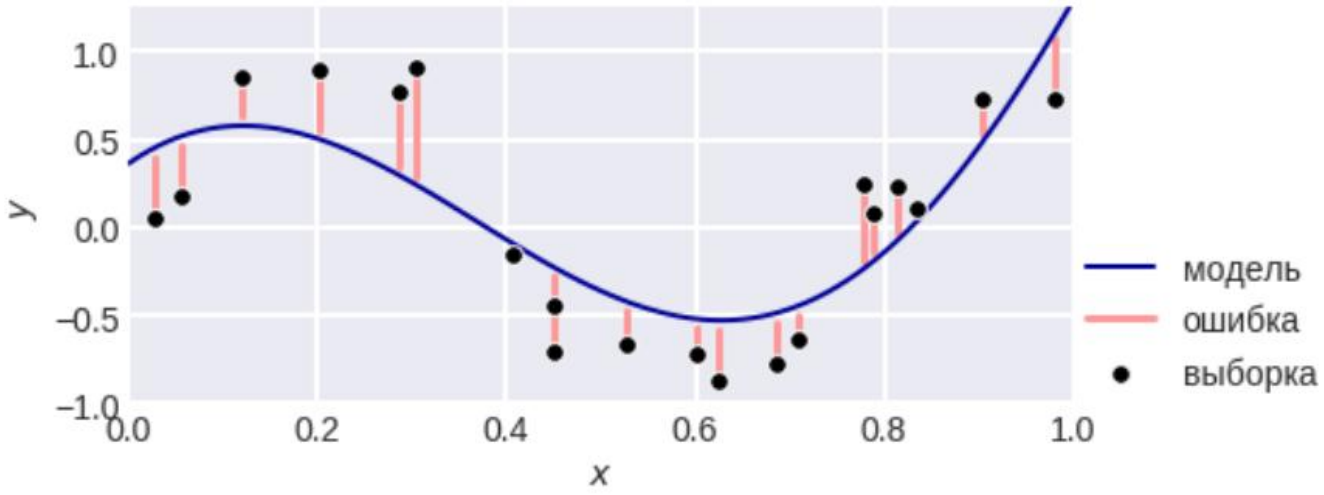

Рисунок 2.3 – Графическое представление задачи регрессии

Построим модель прогнозирования стоимости жилого недвижимого имущества, используя язык программирования Python. Используемыми методами машинного обучения будут метод Bagging, реализованный библиотекой Random Forest, и метод Boosting, реализованный библиотекой XGBoost. Для этого применим следующий подход:

- подготовить данные;
- выбрать функцию потерь;
- проверить целостность данных, пропущенные значения;
- обучить выбранные алгоритмы – XGBoost и Random Forest;
- осуществить настройку алгоритмов, используя параметры каждого из них;
- сравнить полученные результаты.

2.2.1 Подготовка данных

Начнем с импортирования данных, рабочей средой будет выступать Jupyter notebook. Загрузим необходимые библиотеки, а именно:

- Numpy,
- Pandas,
- Matplotlib,
- Seaborn.

Для использования методов машинного обучения для прогнозирования стоимости жилой недвижимости была сформирована выборка заключенных сделок за период с 2011 по 2016 г. Данные, по которым будем строить модель, сохранены в файле train.csv (Comma-Separated Values) для каждой зондировки в следующем формате: Наименование признака, значение для объекта. Наше построение модели начнется с изучения имеющихся данных.

Вся наша выборка состоит из 30471 элементов. Набор данных содержит 292 различных признаков (включая индекс и целевую переменную). Это достаточно большое количество признаков. Признаки нашей выборки включают в себя такие параметры как: Дата совершения сделки, общая площадь квартиры, жилая площадь квартиры, год постройки дома и многие другие. На рисунке 2.4 представлен фрагмент выборки.

id	timestamp	full_sq	life_sq	floor	max_floor	material	build_year	num_room	kitch_sq	state	product_type	sub_area	area_m
30469	2015-06-30	44	27.0	7.0	9.0	1.0	1975.0	2.0	6.0	3.0	Investment	Otradnoe	1.005305e+07
30470	2015-06-30	86	59.0	3.0	9.0	2.0	1935.0	4.0	10.0	3.0	Investment	Tverskoe	7.307411e+06
30471	2015-06-30	45	NaN	10.0	20.0	1.0	NaN	1.0	1.0	1.0	OwnerOccupier	Poselenie Vnukovskoe	2.553630e+07
30472	2015-06-30	64	32.0	5.0	15.0	1.0	2003.0	2.0	11.0	2.0	Investment	Obruchevskoe	6.050065e+06
30473	2015-06-30	43	28.0	1.0	9.0	1.0	1968.0	2.0	6.0	2.0	Investment	Novogireevo	4.395333e+06

Рисунок 2.4 – Частичный фрагмент выборки данных

Обратим внимание, что 276 признаков числовые, 16 категориальные. Нам нужно будет закодировать эти 16 признаков, так как большинство алгоритмов машинного обучения не могут корректно обрабатывать категориальные переменные. Также мы видим, что 292 признаком выступает наша целевая

переменная «price_doc», содержащая значения стоимости каждой квартиры. Прогнозирование значения целевой переменной и будет целью построения наших моделей.

2.2.2 Функция потерь

Метрика или функция потерь характеризует величину отклонения ответа модели от правильного ответа на произвольном объекте выборки [36]. Для задач регрессии существует множество методов измерения ошибки алгоритма, для нашей задачи наилучшим образом подходит «Среднеквадратичная логарифмическая ошибка» (Root Mean Squared Logarithmic Error). выбранная функция является производной от популярной функции «Среднеквадратичная ошибка» (Root Mean Squared Error), и показывает соотношение между фактическим и прогнозируемым значением целевой переменной. Данная метрика используется, когда в ответах выборки оба значения являются большими числами, и абсолютная разница между прогнозируемым и фактическим значением также выражается большим числом. И другие метрики, видя большую абсолютную разницу, в значениях, будут неточно оценивать результаты нашего алгоритма. Кроме того, мы используем данную функцию чтобы больше штрафовать за недооценку, чем за переоценку.

2.2.3 Проверка и преобразование данных

На рисунке 2.5 показано распределение целевой переменной согласно закону нормального распределения, и, поскольку мы в качестве функции потерь берем среднеквадратичную логарифмическую ошибку, то и распределение будет логарифмическое.

В целом распределение соответствует нормальному распределению, что позволяет нам говорить о достаточном качестве исходных данных выборки.

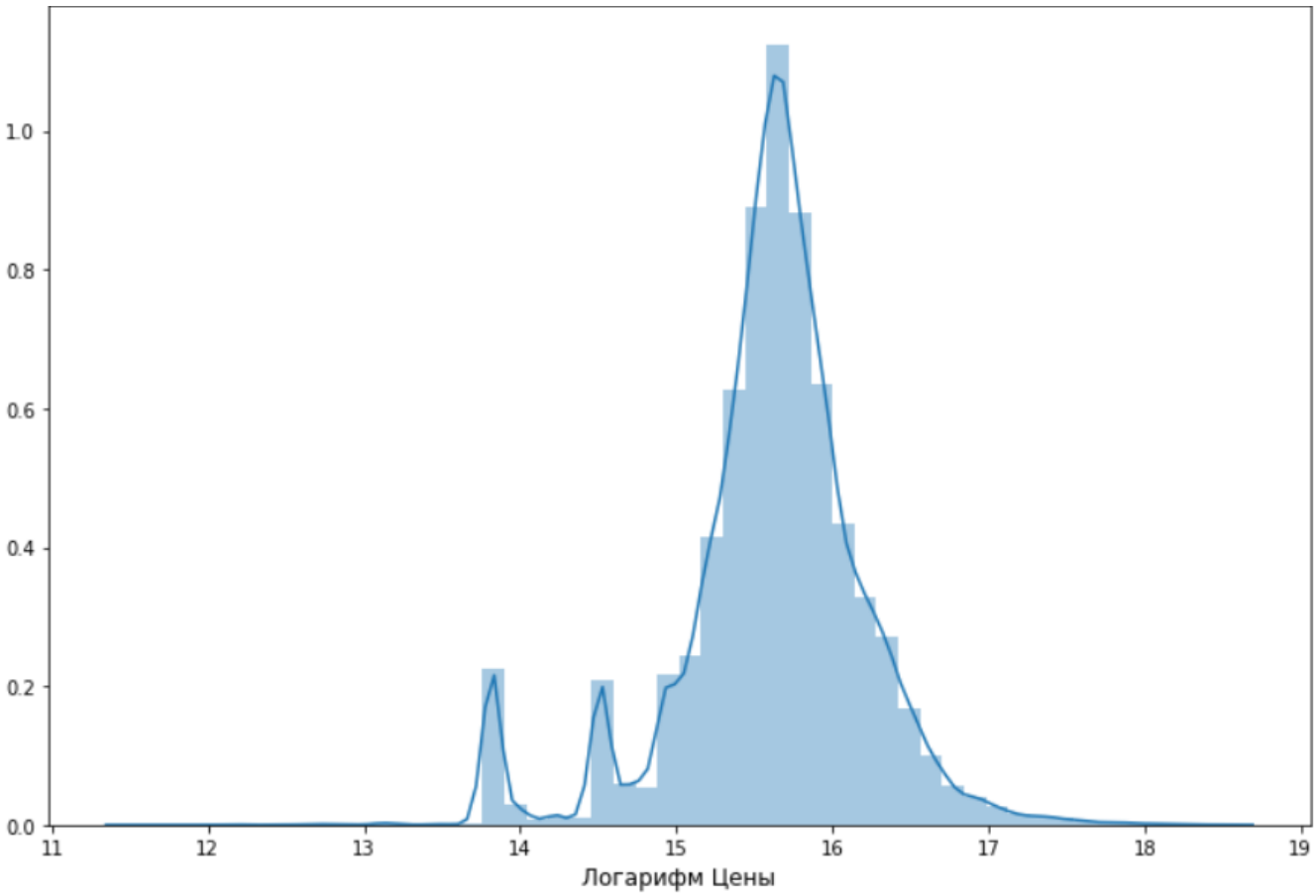

Рисунок 2.5 – Логарифмическое распределение цены в выборке

Продолжая подготовку данных стоит определить, как менялось значение целевой переменной со временем, т.е. как менялась цена на рынке со временем. Результат в виде графика мы видим на рисунке 2.6

Из графика можно сделать вывод что цены реагируют на происходящее на рынке колебаниями цен, но существует общая тенденция к росту с течением времени. Это обусловлено тем, что в хоть наша модель и не имеет таких параметров в выборке, как показатель инфляции, или стоимость валюты, с течением времени эти факторы неизбежно влияют на стоимость. Вместе с тем, что график сам по себе всего лишь отражает распределение средней стоимости жилой недвижимости на рынке Российской Федерации, мы не можем говорить о причинах таких значений.

Также следует отметить, что в нашей выборке могут присутствовать выбросы, т.е. отклонения от нормального распределения в виде завышенной или заниженной цены.

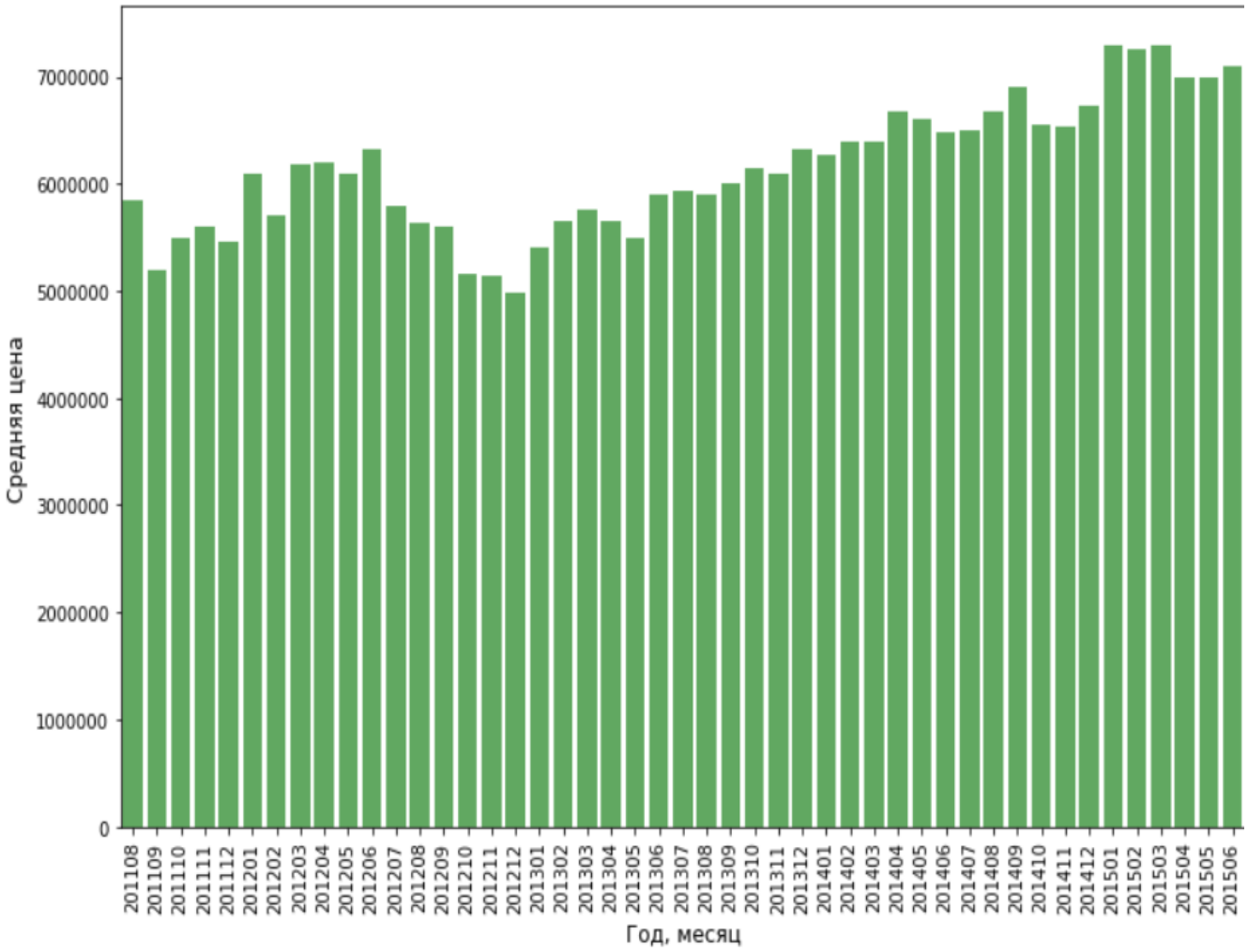

Рисунок 2.6 – Средняя стоимость квартиры за период

Теперь определим, есть ли в нашем наборе пропущенные значения. Для этого построим график пропущенных значений для каждого признака и выведем его на экран. График представлен на рисунке 2.7.

По отдельным признакам видим очень большое количество пропусков, что может привести к неточностям при построении модели XGBoost, и делает невозможной построение модели Random Forest [13], поэтому мы должны избавится от незаполненных значений в нашем наборе, заменив их нулями, что и будет проделано далее. Если наша модель при обучении найдет незаполненное значение, она не сможет оценить его, т.е. классифицировать, что вызовет ошибку выполнения.

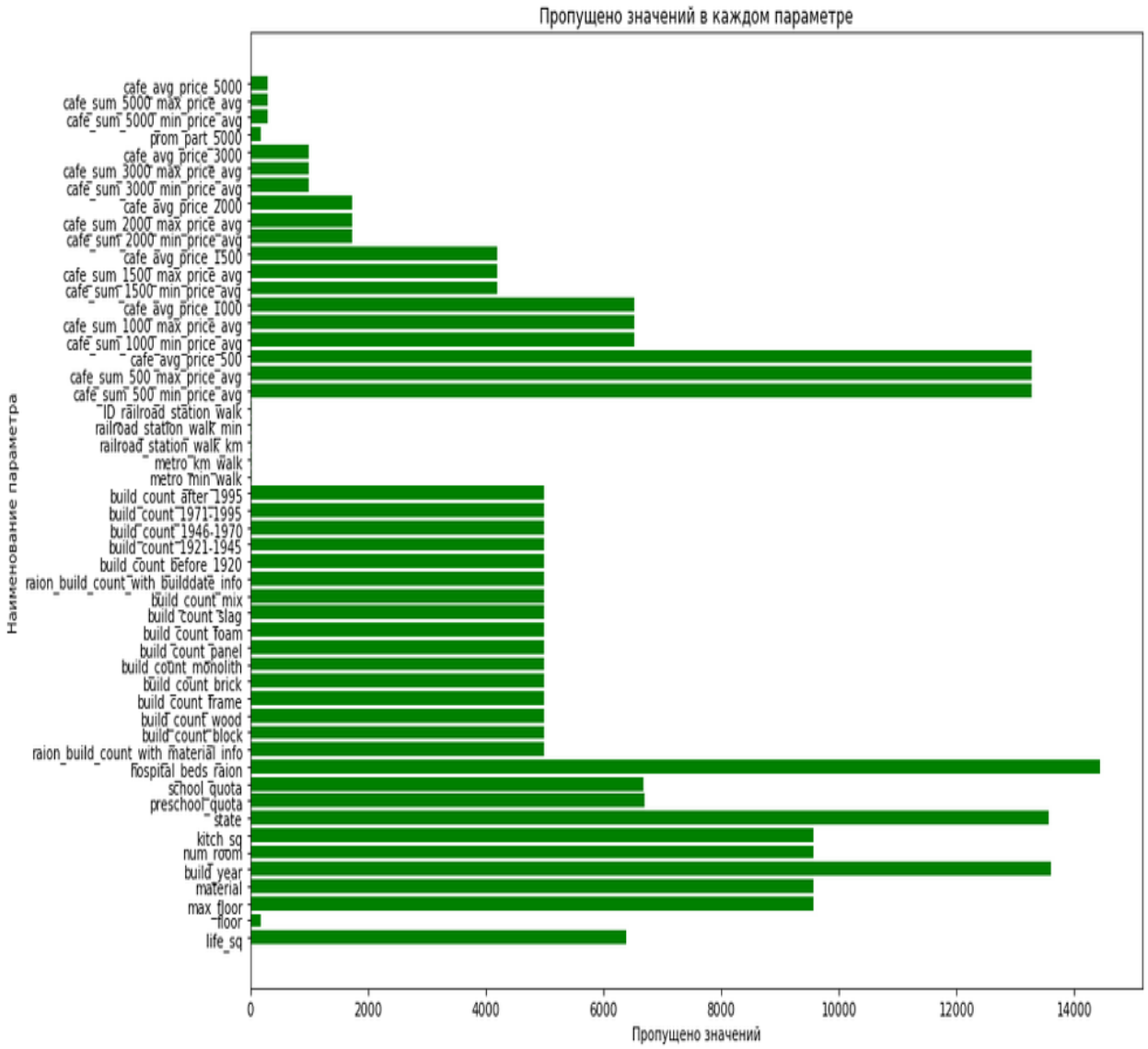

Рисунок 2.7 – Пропущенные значения в данных выборки

Наш набор содержит 16 категориальных признаков. Большинство алгоритмов машинного обучения не могут напрямую работать с категориальными переменными. XGBoost и Random Forest здесь не являются исключениями, так что нам нужно будет преобразовать наши категориальные переменные в численные. Здесь можем выбрать одну из двух стандартных стратегий: кодирование меток (label encoding) или прямое кодирование (one-hot encoding). Какую стратегию использовать – вопрос открытый, и здесь следует рассмотреть несколько факторов:

Прямое кодирование (one-hot encoding) – это базовый способ работы с категориальными признаками. Он выдает разреженную матрицу, где каждый новый столбец представляет одно возможное значение какого-либо одного признака. Так как у нас 16 категориальных переменных мы можем получить разреженную матрицу с огромным количество нулей. Это приведет к более длительному обучению, увеличит затраты памяти и может даже ухудшить итоговые результаты. Другой недостаток прямого кодирования – потеря информации в тех случаях, когда имеет значение порядок категорий [37].

Кодирование меток (label encoding), с другой стороны, просто нормализует столбец входных данных так, что он содержит только значения между 0 и числом классов-1. Для многих алгоритмов регрессии это не слишком хорошая стратегия, однако наши модели могут справляться и справляются с таким преобразованием очень хорошо [37].

И для XGBoost, и для Random Forest мы будем использовать LabelEncoder и нормализуем входные данные.

Еще одним шагом перед построением наших моделей будет разбиение нашего набора на выборку, состоящую из множества вопросов train_X , содержащее основные параметры и множества ответов train_Y. А затем наше множество вопросов train_X поделим на подвыборки Xtrain и Xtest, а множество ответов на подвыборки Ytrain и Ytest. Это делается для разделения общей на выборки на тренировочную (Xtrain и Ytrain) и тестовую (Xtest и Ytest), которая послужит для оценки точности работы наших алгоритмов с помощью функции потерь.

2.2.4 Построение модели Random Forest

Начнем строить нашу первую модель методом Bagging с помощью библиотеки Random Forest.

Random Forest является композицией (ансамблем) множества решающих деревьев. Одними из самых популярных методов решения задач регрессии являются деревья решений - CART (англ. Classification and regression trees —

Классификационные и регрессионные деревья) был пионером методов, его придумали в 1983 четверка известнейших ученых в области анализа данных: Leo Breiman, Jerome Friedman, Richard Olshen and Stone. Алгоритм, представляет собой обычное построение дерева принятия решений [38]. Наглядно этот алгоритм представлен на рисунке 2.8.

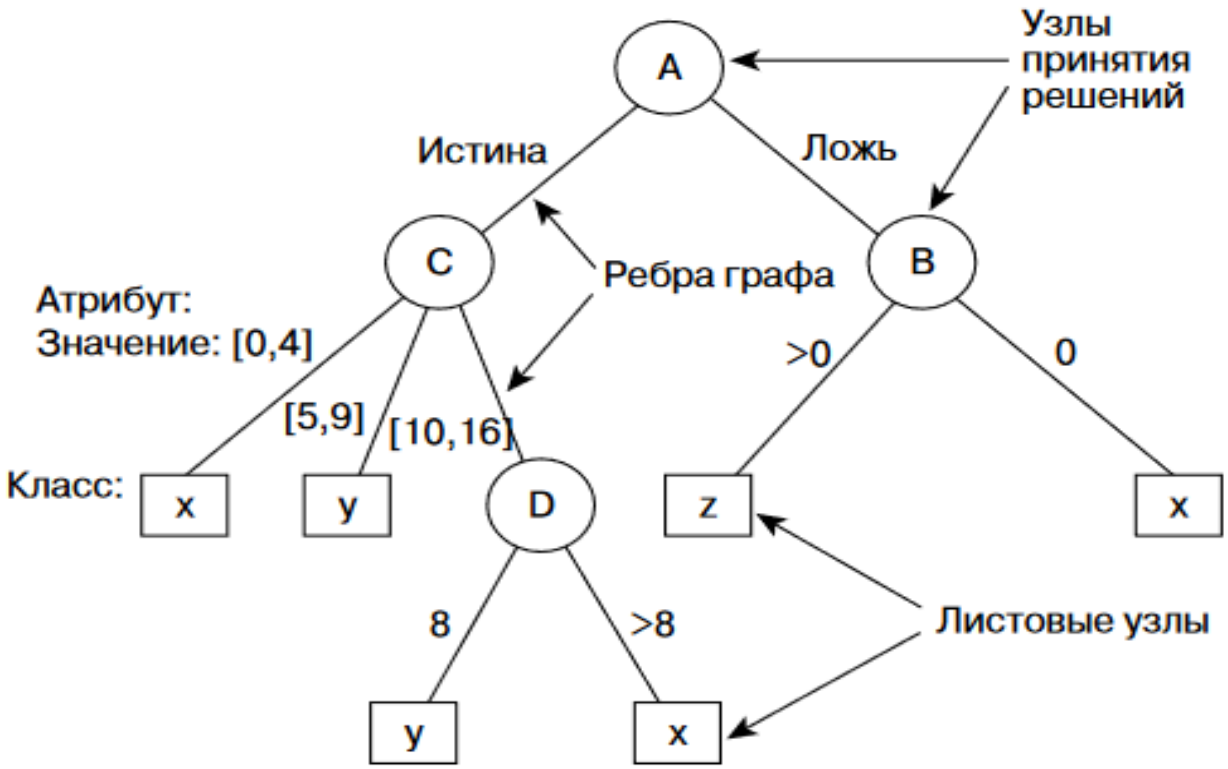

Рисунок 2.8 – Пример дерева решений

На первой итерации мы строим все возможные (в дискретном смысле) гиперплоскости, которые разбивали бы наше пространство на два. Для каждого из разбиений пространства мы считаем количество наблюдений в каждом из подпространств разных классов. В результате выбираем такое разбиение, которое максимально выделило в одном из подпространств наблюдения одного из классов. Очевидно, что разбиение будет представлять корень дерева принятия решений, а листьями будут его два разбиения.

На последующей итерации мы берем один худший (в смысле отношения количества наблюдений разных классов) лист и проводим эти же операции для его разбиения. В итоге наш лист становится одновременно и узлом для последующего

разбиения на два листа. Повторяем до тех пор, пока не достигнем ограничения на количества узлов, или пока от итерации к итерации не перестанет существенно меняться значение ошибки. При этом, если ошибка существенно не меняется, тогда скорее всего, мы столкнулись с переобучением, и не получим номальные результаты на других данных. Чтобы не допустить “переобучения”, используются тестовые выборки (либо кросс-валидация) и, применяется метод обратного анализа (pruning), когда дерево намеренно уменьшается для тестовой выборки. Таков алгоритм для получения простой модели дерева решений. За счет простоты, он считается удобным для первичного анализа данных, например, для проверки на наличие связей между переменными. К плюсам относит быстроту построения модели, а также такую модель легко интерпритировать (достаточно отображения дерева и прослеживания цепочки его решений). Поэтому модели деревьев решений в чистом виде или методы, которые на них основаны, считаются одними из самых популярных методов для работы с данными.

Из недостатков метода можно выделить, что алгоритм часто зависает на локальном решении (выбирает на первом шаге гиперплоскость, не дающую оптимального решения, но дальше этого шага не идет), постоянный контроль и отслеживание переобучения модели и не слишком высокую точность предсказания. Подобных проблем можно избежать если использовать методы ансамблей моделей.

Random Forest позволяет снизить проблему переобучения и повысить точность в сравнении с одним деревом. Прогноз получаем путем усреднения ответов множества деревьев. Деревья обучаются независимо друг от друга (на случайных подмножествах), что не просто помогает решить проблемы построений одинаковых деревьев на одном и том же наборе данных, вместе с тем это делает алгоритм весьма удобным для применения в системах распределённых вычислений. Во время классификации конечным результатом будет тот класс, за который проголосовало большинство деревьев, при условии, что одно дерево обладает одним голосом. К примеру, если в задаче бинарной классификации

получена модель из 500 деревьев, из которых 100 указывают на нулевой класс, а остальные 400 на первый класс, то ответом модели будет предсказание первого класса. Данный подход голосования при построения модели Random Forest в задачах регрессии не работает. Используется среднее значение по всем решениям деревьев.

Random Forest (в виду того что независимо строятся глубокие деревья) достаточно ресурсоемок, а если искусственно ограничить глубину, то мы потеряем в точности поскольку именно глубокие деревья хорошо справляются с решением сложных задач. Время построения прямо зависит от количества деревьев. Естественно, увеличение высоты (глубины) деревьев не самым лучшим образом сказывается на производительности, но повышает эффективность этого алгоритма (хотя и вместе с этим повышается склонность к переобучению). Важны оптимально подобранные параметры (гиперпараметры).

Random Forest не может работать с пропущенными значения, поэтому все пропуски в выборках мы заменили нулевыми значениями ранее.

Теперь, когда данные подготовлены, мы можем построить модель Random Forest, определив параметры обучения. Описание параметров:

- n_estimators – число деревьев. Увеличение деревьев улучшит качество, однако вырастет потребление ресурсов и время выполнения;
- max_features – число признаков для выбора расщепления. Если увеличивать значение параметра, то увеличится время построения модели, а деревья станут «более однообразными». По умолчанию он равен sqrt(n) в задачах классификации и n/3 в задачах регрессии. Это важный параметр. Он не может быть больше числа независимых показателей;
- min_samples_leaf – ограничение на число объектов в листьях. Всё, что было описано про min_samples_split, годится и для описания этого параметра. Часто можно оставить значение по умолчанию (1);
- max_depth – максимальная глубина деревьев. При уменьшении глубины, увеличивается скорость работы метода. Если увеличить глубину резко возрастет

качество на обучении, но и на тестовой выборке оно, как правило, увеличивается. Рекомендуют максимально увеличивать значения параметра, кроме случаев, когда время на построение не окупается итоговым качеством. Если использовать неглубокие деревья, то изменение параметров, которые связаны с ограничением числа объектов в листе и для деления, не приводит к значимому эффекту (листья и так получаются «большими»). При решении задач с большим количество шумных данных рекомендуют использовать неглубокие модели.

Из всех параметров для нас наиболее интересны два: число деревьев и максимальная глубина дерева. Для того чтобы построить модель мы определим что мы будем использовать для обучения множества Xtrain, Ytrain. Построим нашу модель по следующим шагам:

- на графике проверим как изменяется результат работы нашей модели в зависимости от значения параметра n_estimators – число деревьев;
- на графике проверим как изменяется результат работы нашей модели в зависимости от значения параметра max_depth – максимальная глубина деревьев;
- определим лучшие параметры в рамках исследования с помощью сводной таблицы.

На первом шаге организуем обучение и предсказание модели random Forest в цикле, который перебирает значения число деревьев и определяет точность каждого варианта модели. Путем вычислений получаем график, представленный на рисунке 2.9.

Рисунок 2.9 – Точность модели Random Forest при изменении числа деревьев

Как можно наблюдать из рисунка 2.9 с увеличением числа деревьев на начальных порах мы можем наблюдать заметный рост точности, однако начиная с 8 деревьев точность начинает то снижаться, то повышаться. Из данных нашего исследования видно, что лучшее количество деревьев для нашей модели – 13. При этом значении погрешность RMSLE для нашей модели составляет 0.237621 на тестовой выборке.

Следующим наблюдаемым параметром станет максимальная глубина деревьев. Поскольку на предыдущем шаге мы определили оптимальное значение количества деревьев, то данная модель будет содержать значение параметра количество деревьев равное 13.

Построим график точности модели в зависимости от изменения значения параметра глубина деревьев. Наглядно это можно наблюдать на рисунке 2.10.

На рисунке наглядно видно, что с увеличением глубины деревьев на начальных порах мы можем наблюдать заметный рост точности, однако начиная со значения параметра 8 точность начинает то снижаться, то повышаться.

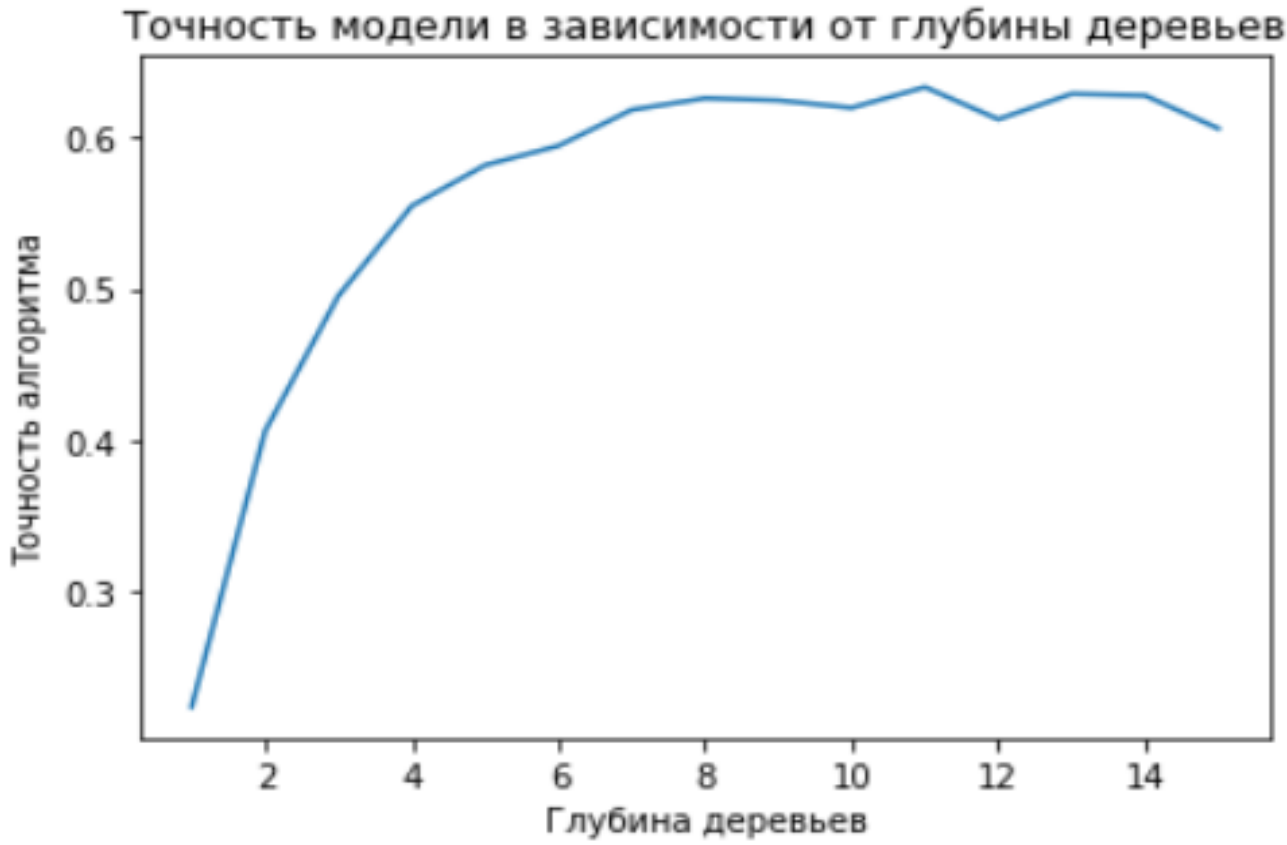

Рисунок 2.10 – Точность модели Random Forest при изменении глубины деревьев

Из данных нашего исследования видно, что лучшее количество деревьев для нашей модели – 11. При этом значении погрешность RMSLE для нашей модели составляет 0,233896 на тестовой выборке. Обратим внимание, что увеличивать глубину деревьев свыше исследуемых значений не рекомендуется, поскольку сильное увеличение данного параметра приводит к сильному переобучению под тренировочную выборку, а также сильно снижает общую точность модели на тестовой выборке в случе, если в тренировочной выборке содержатся частые выбросы, т.е. отклонения от нормально распределенных данных.

Итак, по результатам построения нашей модели алгоритмом Random Forest, лучшими параметрами стали: n_estimators – число деревьев – 13, max_depth – максимальная глубина деревьев – 11. При модели, обученной с данными параметрами, величина функции потерь RMSLE составляет 0,233896.

2.2.5 Построение модели XGBoost

Перейдем к построению модели методом Boosting с помощью библиотеки XGBoost. Это пример ансамблевого мета-алгоритма машинного обучения, применяемого для снижения смещения и дисперсии в обучении с учителем, и семейства алгоритмов машинного обучения, которые превращают слабые модели в более сильные [39]. Изначально, идеи бустинга уходят корнями в поставленный вопрос о том, могут ли «слабые» обучающиеся алгоритмы, которые дают результаты чуть лучше случайного гадания в PAC (вероятно приблизительно правильной) модели, быть “усилены” до “сильного” обучающегося алгоритма произвольной точности. Утвердительный ответ на этот вопрос был дан R.E.Schapire в его статье «Сила слабой обучаемости», которая привела к разработке множества алгоритмов бустинга.

Как видно, основополагающим принципом бустинга является последовательное применение слабых алгоритмов обучения. Каждый последующий слабый алгоритм пытается уменьшить смещение всей модели, объединяя, таким образом, слабые алгоритмы в мощную ансамблевую модель. Можно привести множество различных примеров алгоритмов и методов бустинга, таких как AdaBoost (адаптивный бустинг, который подстраивается под слабые алгоритмы обучения), LPBoost и градиентный бустинг [40].

XGBoost, в частности, является библиотекой, реализующей схему градиентного бустинга. Модели градиентного бустинга строятся поэтапно, точно так же, как и при использовании других методов бустинга. Этот метод бустинга обобщает слабые обучающиеся алгоритмы, допуская оптимизацию произвольной дифференцируемой функции потерь (функции потерь с вычислимым градиентом).

XGBoost, как разновидность бустинга, включает оригинальный основанный на решающих деревьях алгоритм машинного обучения, пригодный для работы с разреженными данными; теоретически обоснованная процедура позволяет работать с весами различных элементов в обучении деревьев. Можно привести ряд достоинств алгоритма XGBoost:

− регуляризация. XGBoost предоставляет очень надежные готовые к использованию средства регуляризации вместе с набором параметров для настройки этого процесса. Перечень этих параметров включает: gamma (минимальное уменьшение функции потерь, необходимое для дальнейшего деления дерева), alpha (вес для L1-регуляризации), lambda (вес для L2-регуляризации), max_depth (максимальная глубина дерева), min_child_weight (минимальная сумма весов всех наблюдений, требующаяся для дочернего объекта);

− реализация параллельных и распределенных вычислений. В отличие от многих других алгоритмов бустинга, обучение здесь может производиться параллельно, тем самым сокращая время обучения. XGBoost работает действительно быстро. По утверждению авторов вышеупомянутой статьи, «система работает более чем в 10 раз быстрее существующих популярных решений уже на одном компьютере и может быть масштабирована на миллионы экземпляров в распределенном или ограниченном по памяти окружении».

Для того чтобы построить модель мы определим, что множества Xtrain, Ytrain мы будем использовать для обучения, а множество Xtest для тестирования .

Модель XGBoost Имеет следующие параметры:.

− max.depth – максимальная глубина дерева. Критически важный параметр: выбор слишком глубоких деревьев приводит к переобучению, а слишком маленькие деревья не позволят эффективно восстановить искомую зависимость.

− eta – скорость обучения модели. Критически важный параметр, контролирующий, с каким весом предсказания каждой следующей модели суммируются с предсказаниями ансамбля. Значение по умолчанию (0.3) является слишком большим, обычно хорошо работают значения меньше 0.1. Слишком маленьким его сделать тяжело, уменьшение eta компенсируется увеличением количества итераций;

− gamma – минимальное уменьшения значения функции потерь;

− subsample – доля объектов обучающей выборки, используемых на каждой итерации;

– colsample_bytree – доля переменных, используемых на каждой итерации; реализует преимущества алгоритма Random Forest (на каждой итерации используется только часть наблюдений и предикторов;

– min_child_weight – минимальное количество наблюдений в листе дерева.

В качестве слабой модели для нашего ансамбля послужит метод линейной регрессии. Математическое уравнение, которое оценивает линию простой (парной) линейной регрессии видно на рисунке 2.11:

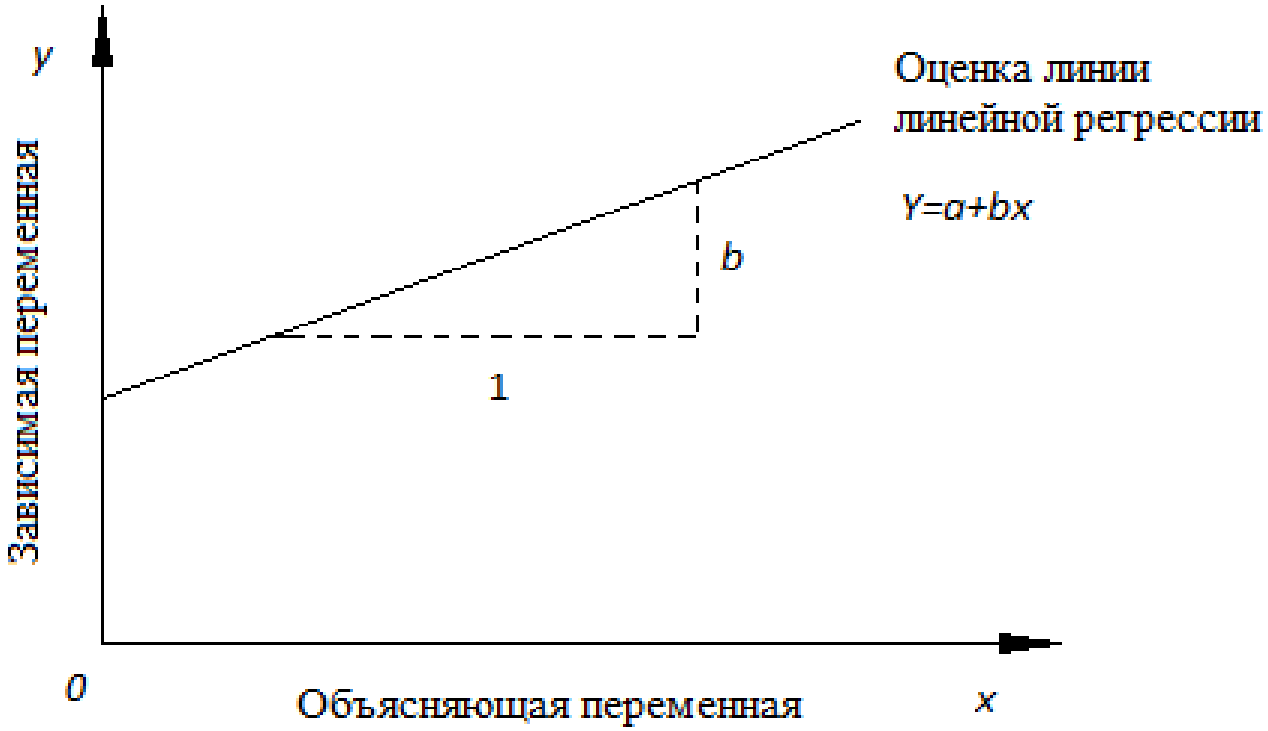

Рисунок 2.11 – Пример линейной регрессии

Y=a+bx,

где x называется независимой переменной или предиктором, Y – зависимая переменная или переменная отклика. Это значение, которое мы ожидаем для y (в среднем), если мы знаем величину x, т.е. это «предсказанное значение y».

a – свободный член (пересечение) линии оценки; это значение Y, когда x=0

b – угловой коэффициент или градиент оценённой линии; она представляет собой величину, на которую Y увеличивается в среднем, если мы увеличиваем x на одну единицу.

a и b называют коэффициентами регрессии оценённой линии, хотя этот термин часто используют только для b.

Можно применять регрессионную линию для прогнозирования y значения по значению x в пределе наблюдаемого диапазона Мы предсказываем среднюю величину y для наблюдаемых, которые имеют определенное значение x, путем подстановки этого значения x в уравнение линии регрессии.

Используем эту предсказанную величину и ее стандартную ошибку, чтобы оценить доверительный интервал для истинной средней величины. Повторение этой процедуры для различных величин x позволяет построить доверительные границы для этой линии. Это полоса или область, которая содержит истинную линию, например, с 95% доверительной вероятностью. Подобным образом можно рассчитать более широкую область, внутри которой, как мы ожидаем, лежит наибольшее число (обычно 95%) наблюдений. В качестве функции потерь для слабой модели будет использована среднеквадратичная ошибка.

Работа самого XGBoost будет заключаться в построении ансамбля решений слабых моделей, строя из них деревья решений.

Построим модель XGBoost использую следующий алгоритм:

- на графике проверим как изменяется результат работы нашей модели в зависимости от значения параметра eta – скорость обучения;
- на графике проверим как изменяется результат работы нашей модели в зависимости от значения параметра max_depth – максимальная глубина деревьев;
- определим лучшие параметры в рамках исследования.

На первом шаге организуем обучение модели XGBoost, на котором определяем значение погрешности алгоритма RMSLE при изменении параметра eta, и построим график, наблюдаемый на рисунке 2.12.

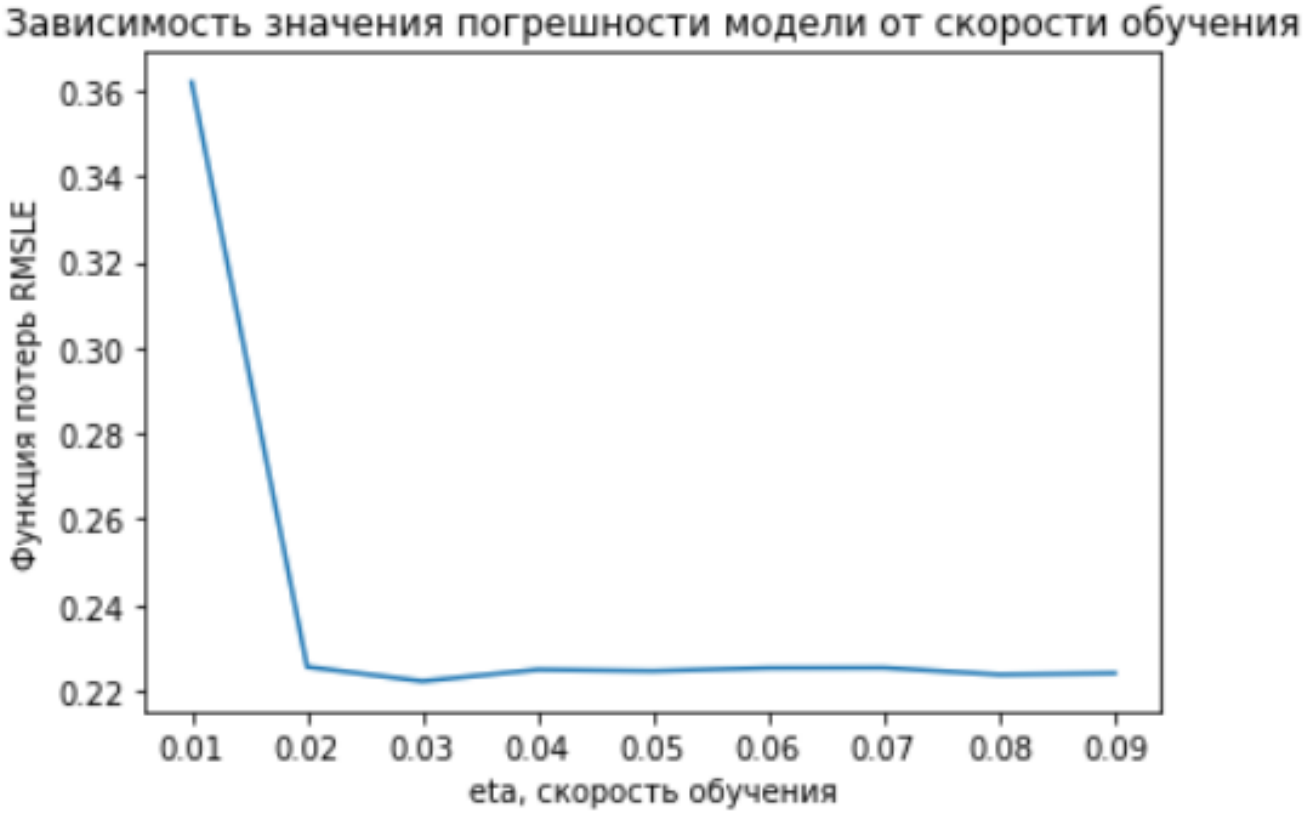

Рисунок 2.12 – Зависимость погрешности модели XGBoost от скорости обучения

Из рисунка 2.12 мы можем наблюдать, что не всегда при увеличении скорости обучения снижается значение погрешности, слишком высокое значение скорости обучения может привести к некорректным ответам на тестовой выборке. Однако есть риск что на нашей не слишком большой выборке алгоритм просто не успеет обучиться должным образом. Однако примем за лучшее значение eta – 0.03, при котором значение функции потерь RMSLE составляет – 0.222148.

Следующим нашим шагом будет исследование зависимости параметра max_depth – глубина дерева. Построим график зависимости значения функции потерь при постоянном значении eta – 0.03, и меняющемся значении max_depth рисунок 2.13

На графике из рисунка 2.13 мы можем наблюдать, что значение погрешности снижается пока значение max_depth не достигнет 10, а затем начинает колебаться. Самый низкий параметр погрешности модель имеет при глубине дерева – 10.

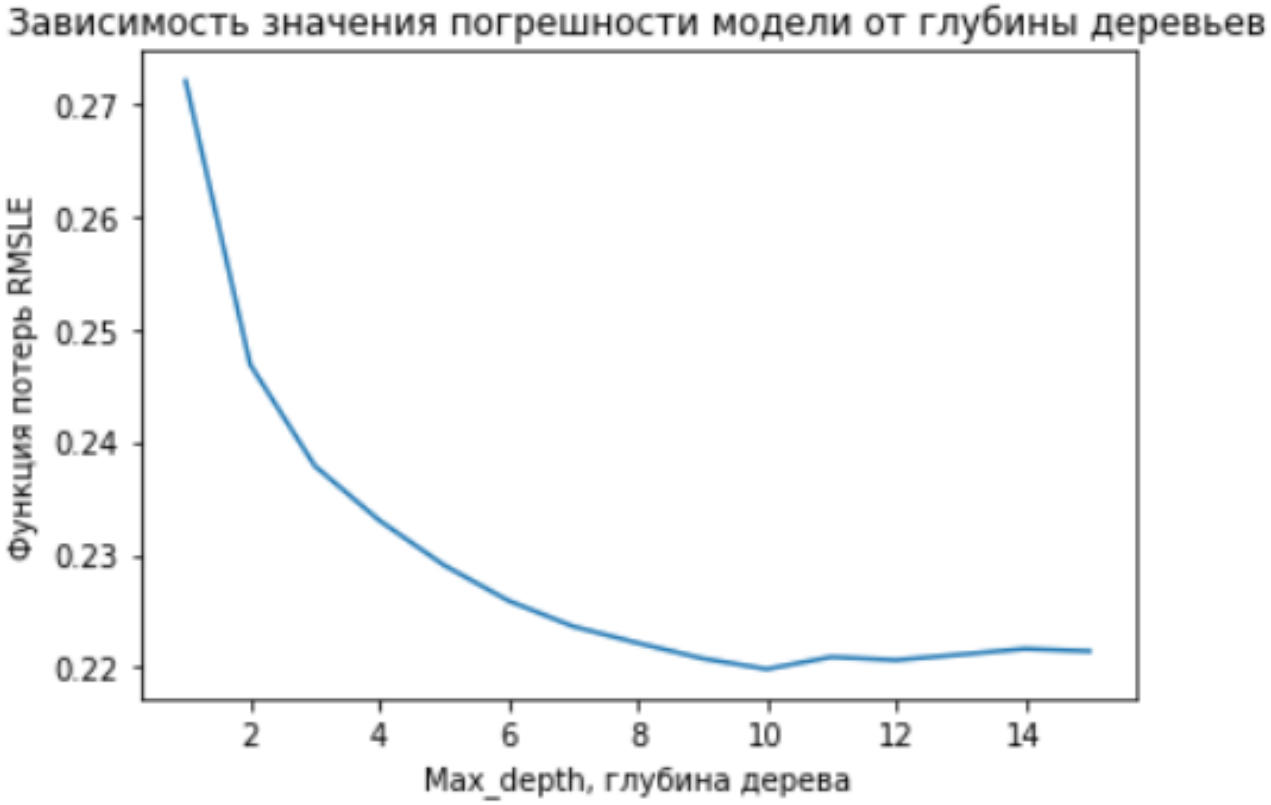

Рисунок 2.13 – Зависимость погрешности модели XGBoost от глубины деревьев

Значение функции потерь при это составляет - 0.220777.

Лучшие параметры для модели составили: eta, скорость обучения – 0.03, , max_depth, глубина дерева – 10. Минимальное значение RMSLE при этих параметрах – 0.220777

Обе модели построены, сравним показатели точности обеих моделей, в данном случае под точностью мы понимаем минимальное значение функции потерь RMSLE.

Под под моделью 1 мы понимаем модель, построенную методом Bagging с помощью алгоритма Random Forest. Под моделью 2 мы понимаем модель, построенную методом Boosting с помощью алгоритма XGBoost.

Модель 1: n_estimators – число деревьев – 13, max_depth – максимальная глубина деревьев – 11. обученная с данными параметрами, минимальное значение функции потерь RMSLE составляет 0,233896.

Модель 2:eta, скорость обучения – 0.03, , max_depth, глубина дерева – 10. Минимальное значение функции потерь RMSLE при этих параметрах – 0.220777.

Из проведенного выше исследования можно сделать вывод, что на нашем наборе данных с задачей прогнозирования лучше справилась модель реализованная с помощью метода Boosting, она оказалась точнее, и в качестве модели прогнозирующей стоимость жилой недвижимости на рынке Российской Федерации мы будем использовать именно данный алгоритм.

Выводы по второй главе:

В результате исследования были решены следующие задачи:

- проведен анализ инструментов разработки;
- подготовлены данные для построения;
- выбрана функция потерь для оценки значения погрешности модели;
- исследованы, распределены и изменены данные для более корректного построения моделей;
- обучены выбранные алгоритмы – XGBoost и Random Forest;
- осуществлена настройка алгоритмов, используя параметры каждого из них;
- произведено сравнение полученных результатов.

Также можно сделать вывод, что при имеющейся выборке, а также при получении сторонних данных мы можем успешно оценивать стоимость жилой недвижимости на рынке Российской Федерации с помощью методов машинного обучения, а именно с помощью построенной нами модели.

3.КОММЕРЦИАЛИЗАЦИЯ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

3.1 Тренды на рынке услуг и методы коммерциализации

Обобщим важные для коммерциализации вопросы, проработанные в ходе настоящей работы.

В главе первой была доказана потребность в определении стоимости объектов недвижимости в ситуациях:

- для страхования объектов недвижимости;
- при операциях купли-продажи или сдачи в аренду;
- при кредитовании под залог объектов недвижимости;
- при акционировании предприятий и перераспределении имущественных долей;
- при исполнении права наследования, судебного приговора;
- при ликвидации объектов недвижимости;
- при кадастровой оценке для целей налогообложения объектов недвижимости: зданий и земельных участков;
- при других операциях, связанных с реализацией имущественных прав на объекты недвижимости.

Сторонами, имеющими такую потребность, являются покупатели и собственники, в т.ч. в роли продавцов объектов жилой недвижимости, банки, страховые компании, государственные органы.

При этом объем рынка (ситуаций, в которых требуется оценка объекта жилой недвижимости) растет. Так, за 11 месяцев 2017 года в Москве зарегистрировано 47 518 договоров долевого участия (ДДУ – сделок в новостройках), что уже на 54% больше, чем за весь 2016-й, и в два раза больше, чем в докризисном 2014-м.

При этом объем сделок с недвижимостью на вторичном рынке Москвы последние три года сохраняет относительную стабильность. По итогам 2017 года в столице зарегистрировано 123 894 прав на основании договоров купли-продажи (мены) жилья. В 2016 году таковых было 126 045, в 2015 –113 769.

По итогам 2017 года в России ожидается выдача рекордного количества ипотечных кредитов – более 1 млн на сумму 1,9–2 трлн руб. В 2014 году исторический рекорд по объемам выдачи ипотеки составил 1,7 трлн руб. Задачи определения и прогнозирования цены решаются достаточно трудоемким способом, требующим обработки большого количества данных обусловленная большим количеством влияющих признаков, в том числе и происходящих не только в Российской Федерации, а также большим количеством наблюдаемых объектов.

В связи с чем была определена проблема в решении задачи удовлетворения данной потребности, связанная с недостатками существующих методов оценки объектов жилой недвижимости:

- метод сравнительного анализа продаж;
- метод капитализации доходов;
- затратный метод;
- экспертные методы прогнозирования.

В случае, если они применяются классическими способами обработки имеющихся данных – высокие сроки получения оценки, совершения ряда рутинных операций, и относительно высокую стоимость подобной услуги, поскольку в цену заложены все вышеперечисленные действия.

Как способ автоматизации, т.е. решения данной проблемы, было предложено использование машинного обучения, но с учетом существующей сегодня специфики и этого рынка.

Российский рынок искусственного интеллекта и машинного обучения только начинает развиваться, демонстрируя отставание от зарубежных рынков: технологии машинного обучения появились всего несколько лет назад, закрытость инноваций в данной области, недостаток вычислительных мощностей и невысокий уровень автоматизации.

На данный момент в распоряжении крупных компаний и банков находятся огромные массивы данных в однотипных, повторяющихся действиях, что позволяет начать работу по созданию онлайн-сервисов для рынка недвижимости, основанных на работе компьютеров.

Ранее в нашем исследовании мы пришли к выводу, что наша модель при наличии подобных данных отлично справится с актуальной проблемой оценки жилой недвижимости, и займет пустующую нишу на рынке услуг.

Таким образом, как решение проблемы мы можем предложить возможность получения независимой, своевременной и точной информации о стоимости жилой недвижимости без привлечения значительных средств и временных затрат.

Поскольку данное предложение представляет собой информационную услугу, то необходим учет и особенностей рынка информационных услуг в настоящее время.

Рынок информационных услуг в настоящее время представляет собой совокупность правовых, организационных и экономических отношений по продаже и покупке информационных услуг, складывающихся между их потребителями и поставщиками. Информационные услуги в связи с их специфичностью относят по целому ряду признаков к нетрадиционным, нетипичным. Основанием для этого являются следующие особенности информационных услуг.

Во-первых, потребление информационной услуги само по себе подразумевает возможность хранения и транспортировки.

Во-вторых, информационная услуга может быть оказана без прямого личного контакта потребителя и производителя

В-третьих, [41 результат оказания информационной услуги овеществлен в хранимой на материальных носителях документальной, машиночитаемой формах, Эти документы содержатся в виде бумажных копий и на машиночитаемых носителях (дискетах, жестких дисках, микрофильмах и т. д.), пригодных к хранению (в видео-, аудио-, библио-, архивах, фильмотеках) и транспортировке (с помощью телекоммуникационных средств связи или на носителях).

Работать с информацией в глобальном масштабе позволяет использование современных телекоммуникационных средств. Основой функционирования современного рынка информационных услуг и соответствующего развития целой

отрасли, обеспечивающей доступ к удаленным базам данных, служит возможность для пользователя получить услуги на расстоянии[42].

Одной из важных особенностей информационных услуг является короткий производственный цикл и как следствие высокая скорость оборота капитала, что дает преимущества бизнесу в этой сфере. Наиболее перспективным выступает сектор электронной коммерции, так как увеличение скорости оборачиваемости капитала вызвано снижением затрат на содержание складских и торговых площадей в сочетании с более быстрым движением денежных средств от покупателя к продавцу.

Информационные услуги характеризуются и специфичностью организации производства. Поставщики услуг выступают чаще малые и средние предприятия различных профилей деятельности. Их высокая мобильность и небольшой размер предоставляют эффективные в условиях локального рынка возможности для гибкой реакции на изменения рыночной конъюнктуры. Однако особое положение на рынке в условиях глобализации занимают центры-поставщики информации, к конкурентное преимущество которых обеспечивает их крупный размер, а быстро получать и обрабатывать большие объемы информации, используя для этого современное компьютерное и телекоммуникационное оборудование, которое требует значительных инвестиций, позволяет эффект масштаба.

Снижение роли субъективного фактора оказания услуги и неопределенность результата осуществляется в связи с отсутствием личного контакта поставщика и потребителя при предоставлении информационных услуг, а также вследствие удаленного их оказания при использовании коммуникационных средств и овеществления на носителях результата в виде баз данных, справок, отчетов, реферативных документов.

Обусловленность персонификацией и индивидуализацией спроса, а также появление новых, нестандартных услуг создает высокую степень дифференциации информационных услуг.

Современный этап развития рынка информационных услуг характеризуется быстрыми темпами роста и основан на телекоммуникационной инфраструктуре и сети интернет. Наблюдается увеличение аудитории интернет-пользователей, быстрый рост мобильной телефонии. Рост оснащенности компьютерами, в том числе персональными средствами связи, послужил одним из основных условий доступа к интернету и соответственно возможности пользоваться информационными услугами.

Быстро увеличивающееся число постоянно подключенных к Интернету мобильных устройств, а также широкое использование социальных сетей и развитая облачная инфраструктура, применяемая для решения сложных аналитических задач, все это характеристики так называемой «третьей платформы». Контент, приложения и услуги, построенные на базе технологий третьей платформы становятся доступны сотням миллионов пользователей. И технологии третьей платформы (большие данные, облачные вычисления, социальные и мобильные технологии) способствуют взаимному развитию. Количество активных пользователей социальных сетей все время увеличивается за счет роста числа мобильных устройств и эти пользователи производят все больше контента, который хранят в облаках. С помощью технологий больших данных накапливаемый таким образом контент становится важным источником для анализа и извлечения ценной информации.

Использование приложения с мобильного устройства можем привести как типичный пример решения, в основе которого используются технологии Третьей платформы.

Приложение может предоставить как доступ к корпоративной информации или информации, находящейся в социальных сетях, так и проанализировать эти данные в режиме реального времени и в ссответствие с полученным результатом предложить выстраивание деятельности. При этом как для приложения, так и данных используются различных частные или публичные облака.

Четыре элемента, лежащие в основе Третьей платформы: мобильные устройства, большие данные, облачные сервисы и социальные технологии.

Таким образом, мы предлагаем следующий продукт для удовлетворения потребности в получения независимой, своевременной и точной информации о стоимости жилой недвижимости без привлечения значительных средств и временных затрат – мобильное приложение

Поскольку наше предложение касается создания нового продукта, то мы имеем дело с инновацией в соответствии с одним из первых определений инновации, как одного из пяти следующих явлений:

- внедрение нового продукта;
- внедрение нового метода производства;
- открытие нового рынка;
- доступ («завоевание») к новым источникам сырья;
- внедрение новых форм организации.

А соответственно и управление процессом его создания будет управлением инновацией.

Говоря о коммерциализации проекта, мы имеем в виду привлечение инвесторов для финансирования деятельности по реализации этой инновации с целью участия в будущей прибыли в случае успеха. В тоже время процесс выведения инновации на рынок является одним из сложных и ключевых этапов инновационного процесса.

Именно этот этап определяет, происходит ли возмещение затрат разработчика (или владельца) инновационного продукта и получение им прибыли от своей деятельности.

Предпринимательство в стране развивается благодаря инновационной деятельности. При отсутствии таковой мы наблюдаем застойные явления во всех других сферах деятельности: оптимизация и рационализация производства, предоставление услуг, осуществление управленческих процессов[43].

Мы можем сделать вывод о том, что в основе инновационной деятельности в целом лежит ее коммерциализация. Таким образом, можно рассматривать коммерциализацию инновационной продукции в качестве процесса вывода инновационного продукта (изобретения, промышленного образца, рационализаторского предложения, полезной модели, ноу-хау) на рынок с целью получения экономической выгоды.

Для коммерциализации инновационной продукции необходимо выбрать соответствующую форму и метод ее осуществления в зависимости от целей и возможностей каждого предприятия.

Основными формами коммерциализации инновационной продукции является:

- коммерциализация самостоятельно предприятием – осуществляется собственными силами за счет своих средств и ресурсов, или ответственности за привлеченные
- смешанная форма, т.е. несколькими участниками.

На предприятии создается временная, например, проектная команда комерциализаторов или функционирует постоянно действующее подразделение, входящее в структуру предприятия. Все зависит от распространенности инновационной деятельности на предприятии, в т.ч. коммерциализации инновации. Когда деятельность предприятия ориентирована в большей части на создание инновационной продукции, а не на производство серийной, то целесообразно создать структурное подразделение, когда инновационная деятельность носит фрагментарный характер и не относится к основным для предприятия уместно создать временную команду сотрудников, которые будут осуществлять процесс коммерциализации.

Когда осуществление коммерциализации невозможно или нецелесообразно, такая деятельность может осуществляться внешним предприятием-коммерциализатором, в т.ч. являющимся профессионалом в этой сфере деятельности. Тогда проведение этапов инновационного процесса

(коммерциализации) возможно без существенного привлечения предприятия-разработчика.

При смешанной форме коммерциализация инноваций может реализовываться в какой-то части самостоятельно предприятием-разработчиком инновационной продукции, в какой-то части внешней организацией. Такую форму можно использовать при коммерциализации инновационной продукции одновременно как для внутреннего, так и для внешнего рынка.

Методы коммерциализации - одним из основных является использование инновационной продукции собственным предприятием для продажи продукции, производимой предприятием, или использование в соответствующих процессах предприятия (производственных, управленческих и др.), а также путем создания дочернего предприятия, которое будет поручено полная реализация производства, сбыта и продвижения инновационной продукции.

Другой метод коммерциализации – это совместное использование, которое может производиться как в виде промышленной кооперации, так и в виде совместного предприятия. При этом, промышленная кооперация предприятий объединяет их на принципах взаимовыгодных отношений по обмену инновационной продукцией. Создание совместных предприятий производится для целей объединения активов.

Диверсификация рисков, имеющая место при совместной деятельности, также благоприятствует коммерциализации инновационной продукции.

Следующий метод коммерциализации инновационной продукции представляет собой частичную передачу прав собственности на использование технологии производства продукции в различных формах – лицензирование, инжиниринг, лизинг, франчайзинг. Где под лицензированием понимается передача прав на использование или продажу инновационной продукции с соблюдением определенных условий. Предоставлением определенных инженерно-технических услуг консультационного характера, связанных с внедрением, эксплуатацией и использованием инновационной продукции. Передачей для пользования

инновационной продукции при соблюдении определенных условий называется лизинг. Сотрудничество, при котором по договору франшизы предоставляется право на использование какой-либо инновации в условиях соблюдения корпоративных стандарта и стиля, других требований франчайзи, именуется моделью франчайзинга[44].

Предприятия полностью передают свои права на инновационную продукцию в случае продажи патента. Этот метод также применяется, когда для предприятия невозможно или нецелесообразно самостоятельно осуществить коммерциализацию.

Коммерциализацией нашей модели станет разработка мобильного приложения, которое позволит каждому пользователю за небольшую плату произвести оценку стоимости жилого имущества на рынке недвижимости Российской федерации. Расчет будет производится на основании построенной нами модели.

3.2 Дорожная карта коммерциализации проекта.

Метод технологического дорожного картирования (Technology Roadmap) используется с 70-х годов прошлого столетия и был разработан компанией Motorola. При его помощи предприятия формируют долгосрочную стратегию развития технологии в отрасли, крупной компании или ряда компаний. К организации стратегического планирования привлекаются эксперты в различных функциональных областях, таких как маркетинг, финансы, технологии, операции и т.д.

Дорожной картой принято называть пошаговое представления сценария достижения стратегических целей от внедрения разработки, сценарий выхода на эффективный уровень, а также способность поддерживать стабильное состояние процесса коммерциализации разработки [45]. Благодаря принципам, разработанным Демингом, а именно его циклом PDCA, который расшифровывают, как «Plan-Do-Check-Act, т.е. «Планируй-Делай-Проверяй-Корректируй/Действуй», показанной на рисунке 3.1.

Остановимся подробнее на каждом из принципов:

– планируй: на данном этапе необходимо определить основную стратегическую цель и сформировать целеполагание;

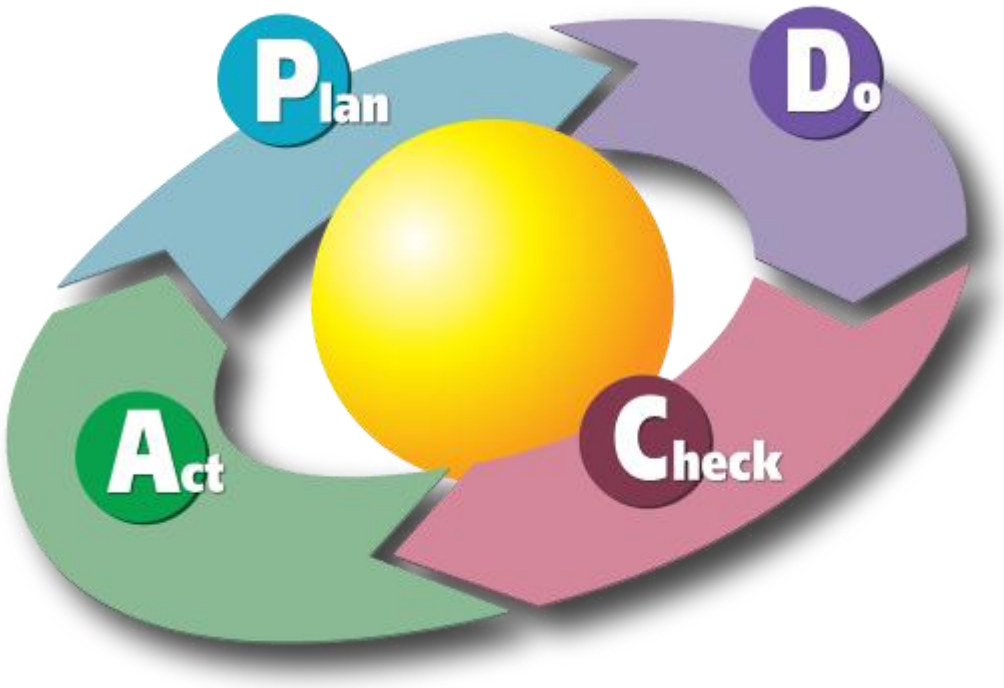

Рисунок 3.1 — Цикл PCDA

– делай: данный этап подразумевает конкретную формулировку основных шагов, необходимых для достижения поставленной цели;

– проверяй: этап применения теории «Бережливого производства» для осуществления непрерывного контроля за проведенными действиями, а также принятием управленческих решений с целью оценки эффективности выбранных решений;

– действуй/корректируй: этап проводящий контроль, аудит и исправления всех проделанных действий.

Дорожная карта является связующим звеном между ожиданием, идеей, стратегией и планом развития процесса, а также выстраивание временное данных шагов по принципу «прошлое – настоящее – будущее». Правильное построение дорожной карты позволяет не только просмотреть вероятные пути развития

разработки, а также ее рентабельность, что позволяет произвести выбор наиболее оптимальных путей достижения поставленных экономических целей и рентабельности разработки.

Во многих зарубежных странах дорожное картирование признано одним из самых эффективных методов планирования, прогнозирования и управления в сфере экономики. При этом данный термин охватывает достаточно широкий круг различных аналитических методик.

3.2 Разработка дорожной карты коммерциализации проекта

Основной целью разработки Дорожной карты является создание стратегии разработки, ввода в эксплуатацию и продвижения мобильного приложения на рынке в 2018/2019 гг. Согласно установленного календарного плана. Также немаловажными являются следующие цели:

- определение стратегических возможностей для роста;
- решение возникших проблем и вопросов.

Дорожная карта данного проекта будет состоять из трех этапов:

1. Анализ ситуации на рынке и оценка потребностей в данном продукте.

- провести анализ существующих на рынке приложений и сайтов для прогнозирования и оценки стоимости жилой недвижимости;
- провести анализ сред разработки, изучить рынок программного обеспечения, проанализировать всевозможные риски использования программного обеспечения;
- проанализировать государственную политику в области внедрения инновационных продуктов в научную среду, оценить уровень участия частного сектора экономики, макроэкономической ситуации;
- выявить информационные пробелы и оценить возможности продвижения разработанного мобильного приложения на рынок Российской федерации.

2. Разработка проекта.

Выполнение данного этапа будет происходить посредством оценки результатов проведения первого этапа «Анализа ситуации и оценка потребности».

Основными пунктами данного этапа являются:

− разработка мобильного приложения в выбранной среде разработки;

− тестирование разработанного мобильного приложения;

− анализ возникших проблем в ходе работы приложения, анализ работы модели, решение возникших проблем;

− введение мобильного приложения в эксплуатацию;

− разработка стратегии маркетинга и выхода на рынок;

− прохождение лицензирования и проведения защиты интеллектуальной собственной разработки.

3. Выполнение, мониторинг и оценка

− три этапа реализации: безотлагательные/краткосрочные действия (приносящие быстрые результаты), среднесрочные действия (зависящие от финансирования) и долгосрочные действия, основывающиеся на общих целях;

− система мониторинга будет введена в действие для измерения хода выполнения каждого проекта в плане его ключевых показателей эффективности;

Оценка будет составной частью этого процесса в целях проведения анализа эффективности и внесения улучшений.

Представление поэтапной картины реализации проекта в таблице 1.

Таблица 1 – Поэтапная картина реализации проекта

Цель	2018/2019			
	IV квартал 2018	I квартал 2019	II квартал 2019	III квартал 2019
Цель	Анализ методов машинного обучения	Мобильное приложение	Привлечение клиентов	Увеличение клиентов
Действия	Анализ существующих методов, построение моделей, выбор	Анализ существующих ресурсов, написание ТЗ, разработка приложения	сбор информации об ошибках, продвижение в социальных сетях, App	Расширение сервиса, пополнение данных о рынке недвижимости,

	модели для реализации		Store и Google Play,	удерживание клиентов

Продолжение Таблицы 1 – Поэтапная картина реализации проекта

Инструменты	Python Jupyter Notebook	Swift Java C#	Социальные сети Google Direct	Публикации в крупных интернет изданиях
Результаты	Готовая модель, прогнозирующая стоимость жилой недвижимости на рынке Российской федерации	Работающее мобильное приложение	Клиенты	Расширенный функционал, увеличенная клиентская база, актуальная база данных по жилой недвижимости

ЗАКЛЮЧЕНИЕ

В процессе исследования была сформирована созданию модели прогнозирования стоимости жилой недвижимости с использованием методов машинного обучения, с последующей коммерциализацией проекта в виде мобильного приложения, предоставляющего возможность расчета стоимости жилой недвижимости.

Была дана характеристика рынка жилой недвижимости и сформирована описание проблемы, а также необходимость применения методов машинного обучения.

Был проведен анализ методов и инструментов для прогнозирования и оценки стоимости жилой недвижимости, в результате, дано описание методов оценки недвижимости, в том числе методов машинного обучения, таких как ансамбли моделей.

Сформулирована задача для построения моделей, а также способы построения алгоритмов машинного обучения.

В результате анализа инструментов построения моделей был выбран язык программирования Python, а также набор необходимых библиотек, выбранными алгоритмами стали Random Forest и XGBoost.

Была сформирована выборка данных и подготовлена для построения выбранными алгоритмами машинного обучения в результате модели были построены, и был проведен анализ точности моделей на основании значения выбранной функции потерь.

Было проведено сравнение лучших по результатам исследования значений параметров каждой модели, и проведено сравнение для определения наиболее точной модели, которой оказалась XGBoost.

Была произведена коммерциализация проекта, на основе которой будет разработано мобильное приложение для прогнозирования стоимости жилой недвижимости с помощью построенной модели.

В виду всего вышеперечисленного можно подвести итоги данной магистерской работы. Все цели поставленные перед началом исследования можно считать достигнутыми, задачи выполненными, а результатом исследования является построенная модель машинного обучения с помощью которой можно прогнозировать стоимость жилой недвижимости. Данная модель была построена с помощью алгоритма XGBoost методом Boosting на языке программирования Pytnoh.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1) Грабовский, П.Г. Экономика и управление недвижимостью: Учебное пособие – Смоленск: «Смолин Плюс»; М.: «АСВ», 2011. – 498 с.

2) Марченко А.В.: Экономика и управление недвижимостью. - Ростов н/Д: Феникс, 2010. – 298 с.

3) Иваницкая, И.П. Введение в экономику недвижимости: учебное пособие/ И.П. Иваницкая, А.Е. Яковлев. - М.:КНОРУС, 2010. – 325 с.

4) Рахман, И.А. Развитие рынка недвижимости в России: теория, проблемы, практика. - М.: Экономика, 2010. – 327 с.

5) Бабкин, С.А. Основные начала организации оборота недвижимости. М.: ЮрИнфоР, 2011. С. 9

6) Экономика недвижимости: Учебное пособие – Владим. гос. ун-т; Сост.: Д.В. Виноградов, Владимир, 2012. – 136 с.

7) Озеров Е.С. Экономический анализ и оценка недвижимости. СПб., 2012. – 412 с.

8) Аналитический центр Tadviser http://www.tadviser.ru/index.php (дата обращения – 10.10.2018)

9) Профессиональный информационный ресурс, посвященный машинному обучению http://www.machinelearning.ru/wiki/index.php (дата обращения – 12.10.2018)

10) Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин)– курс лекций для студентов. http://www.machinelearning.ru/wiki/index.php (дата обращения – 15.10.2018)

11) Шитиков В.К., Мастицкий С.Э. Классификация, регрессия и другие алгоритмы Data mining с использованием R. 2017 https://ranalytics.github.io/data-mining/ (дата обращения – 20.10.2018)

12) Описание ансамблей решения https://ru.wikipedia.org/wiki/АнсамблиРешений (дата обращения – 02.11.2018)
13) Leo Breiman, “Bagging Predictors”, 1994. – 453 с.
14) Ron Meir, “Boosting Tutorial”, 2002. – 233 с.
15) Ritchie Ng, Gul Md Ershad, «Machine Learning for a London Housing Price Prediction Mobile» 2015. – 46 с
16) Neha Chanu, Fatima Hamdan, Lainey Liu «Boston Home Prices Prediction and Evaluation», 2017. – 59 с.
17) Andrew Caplin, Sumit Chopra, John Leahy, Yann LeCun, Trivikrmaman Thampy, «Machine Learning and the Spatial Structure of House Prices» 2016, – 163 с.
18) Свирчков Д.В., Стерник Г.М., Стерник С.Г., Свиридов А.В., Методология прогнозирования российского рынка недвижимости» 2014. – 152 с.
19) Левченко В.В. «Модель определения привлекательности инвестирования в сфере недвижимости с помощью методов машинного обучения» 2014. – 120с.
20) https://www.purplebricks.com/ (дата обращения – 15.10.2018)
21) https://homeapp.ru/ (дата обращения – 15.10.2018)
22) https://www.skyline.ai/about.html (дата обращения – 15.10.2018)
23) https://www.boweryvaluation.com (дата обращения – 15.10.2018)
24) https://ocenka.mobi/ (дата обращения – 15.10.2018)
25) https://www.cian.ru/ (дата обращения – 15.10.2018)
26) Леонид Жуков. Профессия Data scientist, 2013. – 161 с.
27) https://www.r-project.org/ 03.11.2018
28) Марк Лутц. Изучаем Python 4-е издание СПб.: Символ-Плюс, 2011. – 600с
29) http://scikit-learn.org (дата обращения – 04.11.2018)
30) https://docs.scipy.org/doc/numpy/user/quickstart.html (дата обращения – 04.11.2018)

31) https://www.scipy.org/ (дата обращения – 04.11.2018)
32) https://matplotlib.org/ (дата обращения – 05.11.2018)
33) https://jupyter.org/ (дата обращения – 05.11.2018)
34) https://pandas.pydata.org/ (дата обращения – 05.11.2018)
35) https://www.kaggle.com/ (дата обращения – 07.11.2018)
36) https://ru.wikipedia.org/wiki/ФункцияПотерь (дата обращения – 07.11.2018)
37) Воронцов К. В. Введение в машинное обучение– курс лекций для студентов https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie (дата обращения – 10.11.2018)
38) https://ru.wikipedia.org/wiki/CART (дата обращения – 10.11.2018)
39) https://xgboost.readthedocs.io/en/latest/ (дата обращения – 28.11.2018)
40) Robert E. Schapire, “The strength of weak learnability”, Kluwer Academic Publishers, Boston, 1990. – 728 с.
41) Т.В.Савина Экономика и управление: новые вызовы и перспективы., 2010. – 480 с.
42) Васильев, К.А. Влияние развития рынка информационных продуктов и услуг на конкурентоспособность субъектов хозяйствования/ К.А.Васильев//Управленец. - 2013. – 324 с.
43) Жданова О.А. Роль инноваций в современной экономике, Пермь: Меркурий, 2011. – 280 с.
44) Мирошниченко Б. О. Франчайзинг, Economic – definition, 2017 http://economic-definition.com/Business/Franchayzing_Franchising__eto.html (дата обращения – 26.12.2018)
45) https://ru.wikipedia.org/wiki/ТехнологическаяДорожнаяКарта (дата обращения – 26.12.2018)
46) Иванова, Е.Н. Оценка недвижимости: учебное пособи е/ Е.Н. Иванова; под ред. д-ра экон. наук, проф. М.А.Федотовой. – М.:КНОРУС, 2011. – 344с.
47) Золотухина К. Н. Прогнозные оценки цен на региональном рынке жилой недвижимости/К. Н. Золотухина// Современные проблемы гуманитарных и

естественных наук: материалы международной научно-практической конференции. – Москва: Научно-информационный издательский центр «Институт стратегических исследований». Издательство «Спецкнига», 2012. – 56 с.

48) Тургель И. Д. Российский рынок труда: тенденции и угрозы развития в условиях экономического кризиса // Национальные интересы: приоритеты и безопасность. 2010. – 18 с

49) Xiangliang Zhang, “Classification Ensemble Methods”, 2010. – 210 с.

ПРИЛОЖЕНИЕ А

МОДУЛЬ ПОСТРОЕНИЯ МОДЕЛЕЙ RANDON FOREST и XBOOST

```
#Загружаем библиотеки
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import model_selection, preprocessing
import xgboost as xgb
import datetime
from sklearn.ensemble import RandomForestRegressor
color = sns.color_palette()
%matplotlib inline
train_df = pd.read_csv(r'C:\DATA\INPUT\train.csv')
#Нормализация
for f in train_df.columns:
    if train_df[f].dtype=='object':
        lbl = preprocessing.LabelEncoder()
        lbl.fit(list(train_df[f].values))
        train_df[f] = lbl.transform(list(train_df[f].values))
train_X = train_df.drop(["id", "timestamp", "price_doc"], axis=1)
train_Y = train_df["price_doc"]
#Загружаем метод разделения выборки
from sklearn.model_selection import train_test_split
#Готовим подвыборки
Xtrain,Xtest,Ytrain,Ytest = train_test_split(train_X, train_Y, test_size=0.4)
#Подгружаем RMSLE
from sklearn.metrics import mean_squared_log_error
#Замена Nan на 0
```

```
Xtrain = Xtrain.fillna(0)
Ytrain = Ytrain.fillna(0)
Xtest=Xtest.fillna(0)
Ytest=Ytest.fillna(0)
#Организуем обучение в цикле для Random Forest
gr=np.arange(1,15,1)
facc=[]
acc=0
for i in gr:
    scc=0
    model = RandomForestRegressor(n_estimators=i, max_depth = 8)
    model.fit(Xtrain,Ytrain)
    y_predicted = model.predict(Xtest)
    scc=model.score(Xtest,Ytest)
    facc.append(scc)
    if scc > acc:
        acc=scc
        mf=i
        print("Random Forest: , n_estimators", i,   " Точность", scc)
plt.plot(gr,facc)
plt.title("Точность модели в зависимости от числа деревьев")
plt.xlabel("Число деревьев")
plt.ylabel("Точность алгоритма")
print("best n_estimators", mf,  "Наилучшая точность", acc  )
scc=mean_squared_log_error(Ytest, y_predicted)
print("Error RMSLE", scc)
gmd=np.arange(1,16,1)
facc_md=[]
acc=0
```

```
for i in gmd:
    scc=0
    model = RandomForestRegressor(n_estimators=14, max_depth = i)
    model.fit(Xtrain,Ytrain)
    y_predicted = model.predict(Xtest)
    scc=model.score(Xtest,Ytest)
    facc_md.append(scc)
    if scc > acc:
        acc=scc
        mf=i
        print("Random Forest: , Max_depth", i,   " Точность", scc)
plt.plot (gmd,facc_md)
plt.title("Точность модели в зависимости от глубины деревьев")
plt.xlabel("Глубина деревьев")
plt.ylabel("Точность алгоритма")
#Определение обучающей матрицы
dtrain = xgb.DMatrix(Xtrain, Ytrain)
#Определение тестовой матрицы
dtest = xgb.DMatrix(Xtest)
#Цикл для XGB eta
i=0.01
acc_xgb = 0
gr=np.arange(0.01,0.1,0.01)
err=[]
for et in gr:
    xgb_params = {
        'eta': et,
        'max_depth': 8,
        'subsample': 0.7,
```

```
        'colsample_bytree': 0.7,
        'objective': 'reg:linear',
        'eval_metric': 'rmse',
        'verbosity': 1
        }
    model = xgb.train(dict(xgb_params, silent=1), dtrain, num_boost_round=100)
    tt=model.predict(dtest)
    scc_xgb=mean_squared_log_error(Ytest, tt)
    print("et",et,"i",i,"scc_xgb",scc_xgb)
    if scc_xgb>acc_xgb:
        acc_xgb=scc_xgb
        best_i=i
    err.append(scc_xgb)
plt.plot(gr,err)
plt.title("Зависимость значения погрешности модели от скорости обучения")
plt.xlabel("eta, скорость обучения")
plt.ylabel("Функция потерь RMSLE")
acc_xgb = 0
gr=np.arange(0.01,0.1,0.01)
gmd= np.arange(1,16,1)
err=[]
err_x=[]
for imd in gmd:
    print("imd",imd,"i",i,"scc_xgb",scc_xgb)
    xgb_params = {
        'eta': 0.03,
        'max_depth': imd,
        'subsample': 0.7,
        'colsample_bytree': 0.7,
```

```
        'objective': 'reg:linear',
        'eval_metric': 'rmse',
        'verbosity': 1
        }
    model = xgb.train(dict(xgb_params, silent=1), dtrain, num_boost_round=100)
    tt=model.predict(dtest)
    scc_xgb=mean_squared_log_error(Ytest, tt)
    if scc_xgb>acc_xgb:
        acc_xgb=scc_xgb
        best_i=i
    err_x.append(scc_xgb)
plt.plot (gmd,err_x)
plt.title("Зависимость значения погрешности модели от глубины деревьев")
plt.xlabel("Max_depth, глубина дерева")
plt.ylabel("Функция потерь RMSLE")
#Проверка точности с помощью RMSLE
scc_xgb=mean_squared_log_error(Ytest, tt)
print("Погрешность модели", scc_xgb)
print("Модель 1, точность",scc_xgb,"Модель 2, точность", scc, "Разница",diffscc)
```