

Министерство науки и высшего образования Российской Федерации
Московский Физико-технический институт
(Государственный Университет)
Физтех-школа прикладной математики и информатики
Кафедра технологий цифровой трансформации

Выпускная квалификационная работа
"Развитие инструментов предиктивной аналитики в
целях повышения эффективности мониторинга
проектов в сфере жилищного строительства"

Студента 2-го курса Ефремова Сергея Владимировича

Научный руководитель
кандидат экономических наук, доцент Помулев А. А.

Москва, 2022

Аннотация

Рассматривается задача улучшения инструментов предиктивной аналитики, использующихся при мониторинге проектов в сфере жилищного строительства. Исследованы предложенные ранее схемы решения этой проблемы, на основе изученных материалов разработан подход по прогнозированию даты фактической готовности строительного объекта на основании отчетности, публикуемой в открытом доступе и части данных доступных в Фабрике данных банка. Предложен, реализован и протестирован алгоритм, основанный на алгоритмах машинного обучения. Также в работе представлены способы предобработки данных для улучшения качества оценки модели.

Оглавление

1 Введение	5
1.1 Цели и задачи работы	6
2 Постановка задачи	8
2.1 Мониторинг проектов	8
2.2 Эффективность мониторинга	9
2.3 Предиктивная аналитика	10
3 Обзор действующей практики	12
3.1 Анализ текущего состояния рынка жилищного строительства и тенденции развития	12
3.2 Текущее состояние финансирования в сфере жилищного строительства .	13
3.2.1 Основные источники финансирования	13
3.2.2 Эскроу-счета	14
3.3 Процесс мониторинга проектов и методы оценки коммерческим банком .	17
3.4 Типы моделей предиктивной аналитики и их применение в кредитном процессе	18
3.5 Обзор регрессионных моделей машинного обучения	21
3.5.1 Линейная регрессия	21
3.5.2 Измерение ошибки в задачах регрессии	22
3.5.3 Оценивание качества модели	24
3.5.4 Обучение линейной регрессии	25
3.5.5 Градиентный спуск	26
3.5.6 Градиентный бустинг	26
3.5.7 Нейронные сети	29
3.6 Обзор алгоритмов кластеризации в машинном обучении	31
3.6.1 Метрики качества кластеризации	31
3.6.2 Метод k -средних	32
3.6.3 Иерархическая кластеризация	32
4 Предлагаемые изменения	33
4.1 Формальная постановка задачи	33
4.2 Входные данные	33
4.3 Описание моделей	34
4.3.1 Этап предобработки данных	35
4.3.2 Ядро модели	37

4.4 Результаты работы алгоритмов	38
5 Экономический эффект от внедрения модели	41
6 Заключение	43

1 Введение

Одной из ключевых отраслей экономики России является Строительная отрасль. Доля отрасли составляла от 4 до 8% от ВВП страны в различные годы [5]. Особую роль в экономике государства и каждого отдельного взятого региона России играет жилищное строительство, удовлетворяя потребности населения в технологически качественном и современном жилье. Наличие жилья и его доступность напрямую влияет на такие социальные показатели как темпы прироста населения, рождаемость, оказывает воздействие на экономическую культуру, так как требует значительных вложений. Массовое жилищное строительство важно не только для решения социальных проблем, но и развития экономики в целом. Благодаря ярко выраженному мультиплексивному эффекту, сфера жилищного строительства оказывает сильнейшее влияние на национальную экономику страны. Так, она стимулирует создание новых рабочих мест, обеспечивают комфорт и безопасность жизни, стимулирует и поддерживает благоприятный инвестиционный климат. Трудно преуменьшить значимость положительного влияния отрасли на потенциал развития каждого российского региона. Именно поэтому столь важно осуществлять эффективное прогнозирование тенденций в сфере жилищного строительства, с целью избежания или минимизации рисков возникновения кризисов отрасли. Немаловажное место занимает в экономике прогнозирование. Современные методы анализа данных и прогнозирования позволяют предсказывать динамику развития как всей экономики в целом, так и отдельных рынков, в частности рынка строительства. В данной работе были изучены тенденции развития рынка жилищного строительства, теоретические основы алгоритмов машинного обучения, а также реализации этих алгоритмов применительно к данным в отрасли. В работе использовались открытые данные, доступные на сетевом портале [наш.дом.рф](#), а также часть данных закрытого доступа Фабрики данных ПАО "Сбербанк" без публикации фактических значений. На основе данных, было построено и обучено несколько моделей, предсказывающих дату фактической готовности строительного объекта. В последствии было проведено сравнение результатов работы и точности этих моделей. В работе использовались библиотеки с открытым исходным кодом, включающие реализацию моделей классической линейной регрессии и градиентного бустинга.

1.1 Цели и задачи работы

Цель и задачи исследования. Целью исследования является построение модели предиктивной аналитики, которая позволит повысить эффективность процесса мониторинга проектов коммерческим банком в сфере жилищного строительства и улучшить качество прогнозирования вероятности просрочки платежа по сравнению с существующими моделями. Для реализации этой цели были поставлены следующие задачи:

- изучить определение понятий: «мониторинг», «эффективность мониторинга», «предиктивная аналитика» для использования в настоящем исследовании;
- провести анализ существующих проектов и динамики их развития в сфере жилищного строительства;
- ознакомиться с текущим состоянием финансирования проектов в сфере жилищного строительства и нормативно-правовой базой;
- изучить процесс мониторинга проектов и методы их оценки коммерческим банком;
- исследовать типы моделей предиктивной аналитики и их применение в кредитном процессе;
- выделить основные проблемы процесса мониторинга проектов и определить возможности их решения с использованием инструментария предиктивной модели;
- разработать алгоритм внедрения разработанного инструментария в бизнес-процесс мониторинга;
- рассчитать экономический эффект от внедрения модели.

Научная новизна. Используется подход с использование алгоритмов машинного обучения для определению срока фактической готовности строительного объекта. В работе представлены методы предобработки датасета позволяющие улучшить качество прогнозной модели.

Методы исследования. Алгоритмы реализованы на языке программирования Python с использованием библиотек sklearn и LightGBM.

Практическая ценность. Полученная модель может быть использована в качестве встраиваемого модуля. Например, с её помощью можно:

- оптимизировать работу аналитиков рисков, позволяя им уделять больше внимания наиболее рисковым и проблемным объектам;

- дополнять существующие системы мониторинга объектов строительства показателем даты плановой фактической готовности;
- верхнеуровнево оценивать состояние портфеля финансируемых проектов.

2 Постановка задачи

2.1 Мониторинг проектов

Под мониторингом проекта в специализированной научной литературе понимают систематическое наблюдение и контроль за качественными и количественными показателями выполнения проекта, их оценку, с целью подтверждения соответствия целевым значениям, и распространения полученных данных [7].

Понятие контроль проекта является смежным, но не идентичным понятием. Им называют процесс соотнесения планируемых и фактических значений, с целью выявлению отклонения от целевых значений, оценки тенденций изменения показателей и прогнозирования гипотетических альтернатив, что, как следствие, позволяет скорректировать план реализации проекта и улучшить его прогнозируемые показатели.

Рассмотрим основные цели и задачи мониторинга и контроля инвестиционных проектов. Так, они обеспечивают:

- достижение целей проекта в планируемый срок по согласованной стоимости;
- целевое использование денежных средств, выделенных банком в форме кредита, выполнение других условий договора;
- возможность регулярно и качественно информировать руководителей банка о выявляемых и прогнозируемых рисках и проблемах, возникающих в ходе реализации проекта, а также своевременно разрабатывать меры по их недопущению или снижению;
- достижение показателей социально-экономической эффективности, являющихся одними из основных индикаторов успешности проекта.

Реализация проектов в сфере жилищного строительства предполагает проведение следующих видов мониторинга[7]:

- мониторинг хода реализации, в ходе которого эксперты анализируют соблюдение сроков выполнения работ, отклонения от планового бюджета проекта, расчетную стоимость проекта, сроки выполнения контрольных точек, организацию проверок;
- финансовый мониторинг, в ходе которого внешними экспертами оцениваются платежеспособность лиц, задействованных в реализации проекта, в том числе, заемщика, исполнителей, поручителей; анализируются денежный поток, коэффициенты покрытия и целевое использование денежных средств;

- мониторинг эффективности, направленный на сбор и анализ показателей, включенных в положения об экспертизе проектов банка.

Таким образом, отметим, что в ходе мониторинга осуществляется контроль не только за текущей операционной деятельностью (включая исполнение принятых финансовых обязательств и целевое использование средств), проводимой в ходе реализации проекта, но и за достижением конечных, запланированных результатов проекта, особенно имеющих высокую социально-экономическую значимость.

Также отметим, что мониторинг инвестиционных проектов представляет собой систему, включающую несколько элементов[1]:

- отчетность, предоставляемую заемщиком, в том числе финансовую и техническую;
- оценку независимых экспертов и профильных специалистов банков по направлениям реализации проекта;
- календарные план-графики реализации работ, с указанием сроков выполнения работ и их суммарной стоимости;
- данные, собранные автоматизированными системами мониторинга реализации инвестиционного проекта.

В данной магистерской работе особое внимание будет уделено рассмотрению именно последнего элемента. Но предварительно необходимо выделить ключевые стадии мониторинга реализации проекта. Таких стадий всего три:

- подготовка проекта, начинающаяся с одобрения займа или кредита в финансовой организации и завершающаяся выделением средств;
- инвестиционный этап, в течение которого проводится непосредственное финансирование проекта;
- эксплуатация, длившаяся до момента полного исполнения заемщиком платежных обязательств.

2.2 Эффективность мониторинга

Построением эффективных систем мониторинга занимались многие исследователи: Дж.Филлипс, Р.Фатрелл, Х.Керцнер [8, 9, 21], — каждый из которых внес собственный вклад в формирование современной системы мониторинга. Так, на сегодняшний

день, он представляет собой одну из функций проектного управления, включающую в себя сбор, анализ и трансляцию информации о нюансах хода реализации проекта, позволяющей, при необходимости, разработать комплекс мер по минимизации рисков и предотвращению незапланированных процессов.

Отметим, что таким образом, мониторинг позволяет решать задачи по отслеживанию основных индикаторов проекта, организации процесса обработки собранной информации, в том числе с целью последующего формирования отчетности по проекту, а также интегрировать функции мониторинга в информационную архитектуру реализующего проект предприятия.

Несмотря на то, что процесс формирования и развития системы мониторинга со-пряжен с дополнительными финансовыми, временными и ресурсными издержками, ликвидация потенциальных последствий от финансирования убыточных проектов все же обойдется в значительно более высокую сумму. Что безусловно подчеркивает необходимость формирования эффективной, пусть и достаточно дорогостоящей, системы мониторинга проектов.

Для того, чтобы выбрать наиболее эффективную систему мониторинга, требуется проанализировать основные подходы к изучению эффективности проектов. Так, выделяются следующие подходы[12]:

- целевой подход, направленный на анализ текущего состояния достижения целевых значений основных показателей;
- динамический подход, направленный на анализ динамики изменения основных исследуемых показателей во времени и относительно друг друга;
- затратный подход, направленный на сопоставление затрат на реализацию проекта и достигаемых за счет затрат результатов;
- ресурсный подход, направленный на анализ степени рациональности расходования ресурсов.

2.3 Предиктивная аналитика

Предиктивной в специализированной научной литературе называют некоторые аналитические или статистические методы прогнозирования действий или процессов в будущем. В ее основу закладывают статистические модели, позволяющие выявлять имеющиеся закономерности в данных, с целью определения потенциальных рисков и перспектив развития. Для того, чтобы провести предиктивный анализ требуется подключиться к данным, проанализировать, применив предиктивную модель, получить

оценку и прогноз будущих результатов, а также и графически визуализировать полученные результаты.

В основе предиктивной аналитики лежит выявление связей между данными историческими и прогнозными результатами на их основе. Верхнеуровнево алгоритмы предиктивного анализа можно разделить на контролируемое и неконтролируемое обучение[23].

Контролируемое обучение принято разделять на две ключевые категории: регрессию для количественных ответов и классификацию для определения фактической приналежности ответа к той или иной группе.

Неконтролируемое обучение применяется для получения выводов из входных данных без разметки. Наиболее распространенный вид такого анализа - кластеризация, которую используют для поиска скрытых закономерностей в данных.

3 Обзор действующей практики

3.1 Анализ текущего состояния рынка жилищного строительства и тенденции развития

В "Стратегии строительной отрасли до 2030 года"[2] особое внимание уделяется наращиванию жилищного строительства, а также повышению комфортности жилищных условий. Так, в разделе "Целевые показатели по ипотеке и жилищному строительству" документа сформулирована задача увеличить обеспеченность населения жильем к 2024г. до 28–30 м² на человека в среднем, а к 2030 году превысить этот уровень. Также предполагается повысить долю городов с благоприятной городской средой до 70% к 2030г. Для того, чтобы достичь поставленных целей, необходимо нарастить объем жилищного строительства до не менее чем 120 млн. м² в год (рис. 1).



Рисунок 1 – Целевые показатели стратегии развития строительной отрасли до 2030 года, млн.кв.м в год.

На 21 апреля 2022г. в процессе строительства находится около 95 млн. м² и только строительство 3,5% от суммарной площади финансируется без привлечения средств граждан[3].

Количество действующих договоров проектного финансирования на 1 марта 2022 года - 5273 по всей России[11]. Общая сумма действующих кредитных договоров - 7 813 578,8 млн. рублей. Рост за год по сравнению с данными на 1 марта 2021 года - на 96% по количеству, и на 139% по объему. При этом на Москву и Московскую область приходится около 20% заключенных договоров, сумма которых составляет 57% от всех выданных займов. То есть можно говорить о существенном перевесе в пользу Московского региона, относительно всей России и высоких темпах роста рынка, и в Московском регионе в

частности. Однако, к сожалению, не все объекты строительства одинаково успешны и есть тенденция роста проблемных проектов. Так в реестре проблемных застройщиков числятся – 650 организаций в 66 регионах РФ[3]. И на фоне ограниченности ресурсов, сбоя налаженных логистических поставок и роста ставок по кредитам этот реестр будет только пополняться. В частности поэтому, сейчас, как никогда, банку важно вовремя выявить проблемного застройщика и успеть принять превентивные меры, до того как он перестанет платить по счетам или не сможет далее вести деятельность.

3.2 Текущее состояние финансирования в сфере жилищного строительства

3.2.1 Основные источники финансирования

В общемировой практике недвижимость считается одним из наименее рисковых направлений долгосрочного инвестирования с достаточно высоким уровнем рентабельности. Однако условия доступа, задающиеся высоким уровнем капитальными затрат, ограничивают круг потенциальных инвесторов. Так как крупные девелоперские проекты нуждаются в крупных капиталовложениях, реализация их за счет исключительно собственных средств для большинства компаний оказывается невозможной. Однако учитывая текущую практику, именно внешнее финансирование как нельзя лучше отражает суть девелопмента. В зависимости от стадии реализации девелоперского проекта основные источники финансирования могут быть классифицированы в соответствии со схемой[10] (рис. 2).

Предпроектная стадия			
Взносы девелоперов		Авансы заказчиков	
Оценка местоположения и технико-экономическое обоснование			
Взносы девелоперов		Авансы заказчиков	
Приобретение земельного участка			
Взносы дольщиков (с использованием счетов эскроу / напрямую при условии соблюдения критерий)		Средства инвесторов	
Проектирование и оценка проекта			
Взносы дольщиков	Средства инвесторов	Венчурное инвестирование	
Заключение контрактов и строительство			
Банковские кредиты	Взносы дольщиков по договорам, заключаемым на стадии создания недвижимости	Средства инвесторов	Проектное финансирование
Маркетинг, управление и реализация объектов			
Средства инвесторов	Чистая прибыль / Амортизация	Средства покупателей жилой недвижимости	

Рисунок 2 – Классификация источников финансирования девелоперского проекта в зависимости от стадии его реализации.

На первых этапах, как на самых рисковых, проект в основном финансируется за счет собственных средств. Далее, когда концепция будущего объекта строительства

уже разработана, привлекаются средства дольщиков. Финансирование проекта на стадии приобретения земельного участка носит рисковый характер, поэтому инвестиционные ресурсы оказываются самыми дорогими.

На этапе строительства используются банковские, облигационные займы и эскроу-счета. Этот этап проекта наиболее капиталоемкий и требует значительных объемов инвестиционных ресурсов. На стадии продвижения привлечение заемных средств оказывается невозможным, так как займы выдаются под строительство. В условиях Российского рынка задача усложняется высокими ценами на земельные участки, строительные материалы, значительными транспортными расходами и прочими издержками.

Наиболее распространенные механизмы рыночного финансирования девелоперских проектов в России[6]:

- договоры купли-продажи объектов жилой недвижимости;
- договоры участия в долевом строительстве;
- эскроу-счета.

До 2019 года наиболее распространенным механизмом финансирования было заключение договора участия в долевом строительстве (ДДС). Так на конец 2019 года в России было зарегистрировано 783 тысячи договоров ДДС. Такая популярность обусловлена рядом факторов[3]: для покупателя - покупка на начальных этапах строительства позволяла сэкономить 25-35% от стоимости готового объекта; использование инструментов снижения рисков, в частности государственная регистрация ДДС, страхование; в ряде случаев договоры предусматривали рассрочку платежей. Однако данный механизм имел ряд недостатков, в частности порожденная им проблема обманутых дольщиков. Для решения проблемы был разработан комплексный ряд мер, а именно были внесены правки в Федеральный закон от 30.12.2004г. №214-ФЗ "Об участии в долевом строительстве многоквартирных домов и иных объектов недвижимости и о внесении изменений в некоторые законодательные акты Российской Федерации". Как результат - с 1 июля 2019 года в России осуществляется переход на новую схему финансирования строительства многоквартирных домов через эскроу-счета.

3.2.2 Эскроу-счета

Эскроу-счет предполагает открытие специального счета, на который покупатель перечисляет деньги за свою квартиру, заключая при этом с застройщиком договор участия в долевом строительстве. Далее банк блокирует полученные денежные средства на эскроу-счете до момента окончания срока строительства. И только после того, как

застройщик предоставит в банк документы о вводе здания в эксплуатацию, денежные эскроу-счета размораживаются и передаются застройщику.

Данный подход уже приобрел большую популярность в нашей стране. Так, уже в сентябре 2020 года в российских банках было открыто почти 180 тысяч эскроу-счетов, что суммарно составило более 600 миллиардов рублей[2]. При этом банки одобрили суммарное финансирование на сумму, превышающую 1,8 триллионов рублей. А по состоянию на 1 марта 2022 года количество эскроу-счетов составило более 681 тыс., при объеме аккумулированных на этих счетах средств более 3,39 трлн. рублей. По данным Банка России[11], средняя процентная ставка кредитной линии проектного финансирования составляла 4-6% для договоров, заключенных до марта 2022 года. Средний срок рассмотрения заявки 30-45 дней. Особенность кредитования с применением счетов-эскроу заключается в том, что при значительном размере поступлений денежных средств участников долевого строительства на счета эскроу - ставка по кредиту снижается, в пределе она может быть снижена до 0.01%.



Рисунок 3 – Динамика использования эскроу счетов.

Тем не менее, переход на представленную модель финансирования был сопряжен с целым рядом нюансов. Так, сама по себе данная модель была разработана с целью обеспечения безопасности денежных средств дольщиков, за счет того, что их средства теперь заменяются банковским кредитованием. Согласно расчетам, представленным на официальном сайте банка ДОМ.РФ[3], для того, чтобы осуществить указанный процесс в полной мере, требуется увеличить объем кредитования застройщиков с 0,6 триллионов рублей в 2018 году до 6,4 триллионов рублей к 2024 году. Таким образом, получается, что в горизонте пятилетнего планирования, кредитный портфель застройщиков жилой недвижимости возрастёт более, чем в 10 раз (рис. 3).

Отметим также, что обслуживание ссудной задолженности в рамках такой модели финансирования строительства объектов жилой недвижимости станет реальным только в том случае, если объем продаж квартир на первичном рынке будет ежегодно расти. Для обеспечения должного роста предполагается увеличение числа участников льготного ипотечного жилищного кредитования (далее – ИЖК), субсидируемого государством. Так, по итогам 2019 года, около 60% от всех квартир в новостройках и примерно 50% квартир во вторичном секторе были приобретены с использованием ИЖК[2]. В данном случае, стоит отметить, что при таком способе основным ограничителем спроса становится повышение долговой нагрузки на население, достигаемое в следствие отсутствия индексации доходов населения при повышении платежей по кредиту на фоне роста цен на недвижимость. Все это, как следствие, приводит к снижению благосостояния навыкам и негативным долгосрочным социально-экономическим эффектам. Особенно остро обе эти проблемы ощущались на фоне повышения ключевой ставки ЦБ до 20% в марте, ставки по ИЖК стали неподъемными даже по льготным программам ипотечного кредитования, а цены на недвижимость продолжили рост на фоне повысившихся инфляционных ожиданий национальной валюты.

Реформа отрасли строительства жилой недвижимости оставила девелоперов без возможности использовать бесплатные средства дольщиков. На начальных этапах перехода на проектное финансирование объем жилищного строительства в РФ снизился на 20 млн кв.м. Половина регионов Российской Федерации характеризуется нулевой или отрицательной маржинальностью жилищного строительства, обусловленной низкой платежеспособностью населения. Все это ограничивает застройщика при выборе источников финансирования. Также важен и тот факт, что кредит не покрывает все затраты проекта, а собственных средств застройщика может быть недостаточно для увеличения объема строительства.

Для поддержания темпов роста, застройщики выходят на фондовый рынок. Доверие потенциальных инвесторов к ценным бумагам строительных компаний потенциально могут повысить структурирование девелоперских групп, цифровизация компаний, повышение прозрачности их отчетности и формирование рейтинговой истории. Однако в связи со сложной политической обстановкой, правильно будет ориентироваться на привлечение инвесторов внутри страны, так как доступ внешнего капитала весьма ограничен. Также исключительно важной становится и степень зависимости застройщика от внешнего капитала и импортных комплектующих, материалов и инструментов при строительстве. Чем выше степень зависимости, тем менее устойчивой становится девелопер и конкретные, финансируемые банком проекты, при оказании на них даже косвенного политico-экономического давления.

Все это приводит к необходимости постоянного проектного мониторинга и контроля потенциальных рисков со стороны банка, как у регулятора на новом сложившемся рынке жилищного строительства после введения счетов-эскроу.

3.3 Процесс мониторинга проектов и методы оценки коммерческим банком

Рассмотрим основные этапы мониторинга хода реализации проекта коммерческим банков. Так, в ходе осуществления мониторинга, банки проводят действия по формированию календарного план-графика, с целью отражения на нем основных событий инвестиционного проекта. Данный график помогает упростить контроль за соблюдением плановых сроков проведения работ, а также соотнести плановые затраты с фактически понесенными на выполнение указанных работ. В случае выявления отклонений от плановых значений, коммерческий банк должен провести анализ степени влияния этого отклонения на сроки и бюджет проекта в целом. Если отклонение окажется значительным, то потребуется провести мероприятия по снижению данных отклонений и возврату к целевым, плановым значениям.

Отметим также, что в ходе мониторинга могут быть выявлены и дополнительные обязательства заемщика, залогодателя или поручителя, в связи с возникновением дополнительных условий или обязательств, предварительно заложенных в договор. В данном случае, банку потребуется провести экспертизу бюджетного проекта, с целью проверки соответствия стоимости параметров проекта текущему состоянию рынка.

Также коммерческие банки в ходе мониторинга могут осуществлять действия по отбору компаний, осуществляющих технический надзор, проводить анализ отчетов и согласовывать виды рисков при страховании работ.

Также ценную информацию возможно получить и при проведении мониторинга эффективности проекта. В ходе данной процедуры возможно узнать достиг ли проект поставленных перед ним целей, и, что особенно важно, удалось ли получить ожидаемые положительные социально-экономические эффекты, а также узнать, соответствует ли он изначальным параметрам.

В ходе такого мониторинга могут проводиться следующие действия:

- анализ достижения запланированных конечных показателей эффективности;
- анализ достижения положительных социально-экономических эффектов;
- анализ эффективности инвестиций акционеров в капитал проектной компании.

Резюмируя, мониторинг результатов развития инвестиционного проекта включает:

- определение целевых показателей на предынвестиционной стадии;
- проведение оценки проектов в процессе их реализации на инвестиционной стадии.

При этом целевые показатели можно разделить на категории:

- финансовые;
- экономические;
- экологические;
- показатели развития частного сектора.

Зачастую указанные выше показатели рассчитываются проектным офисом Банка вручную, для каждого проекта строится индивидуальная модель, сравниваются фактические данные и показатели план-графика и на основе отклонений делается прогноз по каждому объекту. Такой подход экспертной оценки имеет ряд преимуществ, таких как возможность рассмотреть каждый случай досконально и выделить все проблемные зоны. К сожалению, применение методики экспертных оценок является весьма время- и трудо- затратным процессом и невозможно при превышении определенной границы: числа объектов мониторинга в отношении на эксперта. Эта проблема вызвала необходимость разработки автоматизированных систем мониторинга проектов, которые могли бы по расчетным показателям и косвенным признакам подсветить наиболее проблемные объекты. Также стоит отметить, что наиболее эффективно не просто показывать объекты риска, но и индицировать по каким причинам объект считается проблемным.

3.4 Типы моделей предиктивной аналитики и их применение в кредитном процессе

В основе применяемых в Банках систем мониторинга чаще всего лежат методы прямого сравнения с пороговыми показателями и прогнозными сравнениями или алгоритмы решающих деревьев. Причин у такого подхода несколько. В первую очередь, у Банка нет права на ошибку, поэтому система оценки проекта должна быть прозрачной и простой к оценке. Так у контролирующего ее работу эксперта будет возможность доступно интерпретировать результат автоматического расчета и оперативно сделать выводы о его корректности и необходимости корректировок оценки. Также сложные системы проектного мониторинга и скоринговые-системы чаще всего требуют крупных вложений как на уровне разработки или закупки оценочных моделей, так и на этапе разворачивания решений на высокопроизводительных кластерах.

Среди наиболее популярных алгоритмов машинного обучения можно выделить методы [23], представленные в Таблице 1.

Таблица 1 – Наиболее распространенные алгоритмы машинного обучения в банковской сфере

Модель	Краткое описание
Линейная множественная регрессия	Связывает зависимую переменную с линейной функцией независимых переменных. Задача сводится к поиску коэффициентов, при которых точность ответа, полученного в результате подстановки входных значений обучающей выборки в итоговую линейную функцию будет максимальна.
Логистическая регрессия	В основе лежит метод максимального правдоподобия. В целом, логика поиска весов аналогична линейной регрессии, только оценивается вероятность принадлежности к одному из классов, решая задачу бинарной классификации.
Деревья классификации	Зависимость значения результирующей переменной представлена в виде иерархической ступенчатой структуры - дерева.
CHAID (Chi-squared Automatic Interaction Detection)	Для этой модели критерием построения следующих рядов узлов является значимость результата статистического теста. Так, на каждом уровне дерева выявляется переменная оказывающая наибольшее влияние на результат. Также выделяется набор признаков, оказывающий максимальное влияние на результат.
Нейронные сети	Каждый узел / нейрон - простой элемент, который можно промоделировать. Нейросеть позволяет обнаружить сложные и нелинейные зависимости между признаками и выходным результатом. Ключевая проблема данного подхода заключается в сложности интерпретации. Так нет возможности описать в виде простой функции взаимосвязи, задающие структуру нейросети.

Модель	Краткое описание
Случайный лес	Композиция решающих деревьев. Финальная классификация получается методом усреднения результата всех поддеревьев.
Метод опорных векторов	Ключевая идея данного подхода состоит в повышении размерности пространства признаков, и последующем поиске разделяющей гиперплоскости между исходными векторами значений.
Байесовский классификатор	Простой вероятностный классификатор, основанный на предположении о независимости признаков в исходном пространстве и последующем применении теоремы Байеса

Большая часть представленных выше алгоритмов встречаются в кредитном скринге. Задача мониторинга финансируемых проектов со стороны банка выглядит крайне близкой, разница заключается в том, что клиент оценивается не только в момент одобрения кредитной линии, но и на протяжении всего жизненного цикла проекта.

Основные направления кредитного скринга:

- Application-scoring или анализ заявок для определения потенциального риска выдачи кредита. Для задачи ПФ можем считать вероятность наличия просрочки у клиента;
- Fraud-scoring или скринг против мошенничества, оцениваем вероятность того, что клиент является мошенником. Не актуально для задач ПФ;
- Behavioural-scoring или скринг поведения заемщика. Для задачи ПФ, это вероятность что клиент передает несоответствующие действительности данные о ходе выполнения проекта;
- Collection-scoring определяет сиворость мер, которые необходимо применить к заемщику, при просрочке.

Данная работа нацелена, в первую очередь на анализ адаптации Application скринга для задач проектного финансирования. Целевая прогнозируемая переменная имеет числовое значение, поэтому классификационные методы здесь оказываются не уместными. Ключевые модели с такой целевой переменной, о которых пойдет речь далее в

работе, это классическая линейная регрессия, градиентный бустинг и нейронные сети. Также в качестве вспомогательного модуля прогнозной модели будут рассмотрены алгоритмы кластеризации: метод иерархической кластеризации и k-ближайших соседей.

3.5 Обзор регрессионных моделей машинного обучения

Объектом изучения предиктивной модели называется то, для чего планируется производить предсказания, для текущей работы это объект строительства. Далее в работе эти объекты будем обозначать литерой x_i , где i — индекс изучаемого объекта. Множество всех объектов строительства в РФ мы обозначим как \mathbb{X} . Прогнозируемая величина, которую мы хотим определять, например это может быть вероятность просрочки платежа со стороны застройщика, называется целевой переменной, а множество ее значений пространством \mathbb{Y} . В нашем случае это подмножество действительных чисел, задающее вероятностное пространство: $\mathbb{Y} = \mathbb{R}_{[0,1]}$. Отдельный ответ обозначим буквой y .

Значение примера пары объекта и ответа из жизни будем называть обучающим, а всю их совокупность — обучающей выборкой, обозначаемой как $X = \{(x_1, y_1), \dots, (x_l, y_l)\}$, где x_1, \dots, x_l — обучающие объекты, а l — их количество. Особенность обучающей выборки заключается в известности множества ответов y_1, \dots, y_l .

Заметим, что объекты — это некоторые абстрактные сущности, поэтому для анализа будем описывать объекты с помощью некоторого набора признаков. Признаковое описание объекта, которое задается вектором всех признаков, будем отожествлять с самим объектом.

3.5.1 Линейная регрессия

Если в постановке, описанной в предыдущем параграфе, целевая переменная является вещественной и рассматривается вариант обучения с учителем, то задача является задачей регрессии.

Регрессионную задачу всегда можно свести к суммированию значений признаков с некоторыми весами:

$$a(x) = \omega_0 + \sum_{j=1}^d \omega_j x_j. \quad (3.1)$$

Параметры модели - коэффициенты ω_i (веса). Вес ω_i также называют свободным коэффициентом или сдвигом (bias в англоязычной литературе). Заметим, что сумма в формуле 3.1 является скалярным произведением вектора признаков на вектор весов, поэтому запись может быть упрощена:

$$a(x) = \omega_0 + \langle \omega, x \rangle, \quad (3.2)$$

где $\omega = (\omega_1, \dots, \omega_d)$ — вектор весов.

Простая форма линейных моделей позволяет сравнительно быстро и легко их обучать, что делает их весьма популярными при работе с большими объемами данных[18]. Также наличие небольшого и интерпретируемого набора параметров позволяет контролировать риск переобучения, учитывать зашумленность данных, а также работать с небольшими выборками.

3.5.2 Измерение ошибки в задачах регрессии

Для обучения любой модели необходимо определить критерий оценки качества предсказаний. Будем обозначать через y значение целевой переменной, а через a — предсказание модели. Методики оценки отклонения $L(y, a)$ прогноза от истинного ответа рассмотрим ниже.

MSE или метод наименьших квадратов заключается в необходимости посчитать квадрат разности:

$$L(y, a) = (a - y)^2. \quad (3.3)$$

Данная функция дифференцируема, поэтому наиболее часто используется в задачах линейной регрессии. Таким образом, для оценки можно использовать функцию:

$$MSE(a, X) = \frac{1}{l} \sum_{i=1}^l (a(x_i) - y_i)^2. \quad (3.4)$$

Однако, величина среднеквадратичного отклонения плохо интерпретируется, так предсказывая цену в рублях, ошибку мы будем получать в квадратах рублей, для решения этой проблемы используют корень из MSE:

$$RMSE(a, X) = \sqrt{\frac{1}{l} \sum_{i=1}^l (a(x_i) - y_i)^2}. \quad (3.5)$$

Среднеквадратичная ошибка хорошо подходит для сравнения двух моделей, а также для контроля качества во время обучения, однако не позволяет сделать вывод о том, насколько качественно конкретная модель решает задачу. Поэтому вместо среднеквадратичной ошибки может быть эффективнее использовать коэффициент детерминации, который определяется как:

$$R^2(a, X) = 1 - \frac{\sum_{i=1}^l (a(x_i) - y_i)^2}{\sum_{i=1}^l (y_i - \bar{y})^2}, \quad (3.6)$$

где $\bar{y} = \frac{1}{l} \sum_{i=1}^l y_i$ — среднее значение целевой переменной. Коэффициент R^2 отражает долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Если

показатель коэффициента близок к единице, то модель хорошо объясняет данные, если же близок к нулю, то модель нерепрезентативна.

MAE. Если заменить в формуле оценки отклонения квадрат отклонения на модуль:

$$L(y, a) = |a - y|, \quad (3.7)$$

соответствующая функция ошибки называется **средним абсолютным отклонением**:

$$MAE(a, X) = \frac{1}{l} \sum_{i=1}^l |a(x_i) - y_i|. \quad (3.8)$$

Модуль отклонения менее чувствителен к выбросам, но при этом не является дифференцируемым.

Важная проблема абсолютной функции потерь скрыта в ее производной, рассмотрим производные для нее и квадратичной функции:

$$\frac{\partial}{\partial a} |a - y| = sign(a - y), a \neq y; \quad (3.9)$$

$$\frac{\partial}{\partial a} (a - y)^2 = 2(a - y). \quad (3.10)$$

При использовании градиентных методов обучения, параметры модели постепенно меняются на основе функции потерь, при этом производная абсолютной функции потерь никак не зависит от близости прогноза к правильному ответу, поэтому ее использование приводит к более долгой и сложной процедуре обучения.

MAPE. **Средняя абсолютная ошибка** очень похожа на абсолютную ошибку, однако здесь исследуется отклонение не абсолютного значения целевой переменной а относительной величины:

$$MAPE(a, X) = 100\% \times \frac{1}{l} \sum_{i=1}^l \frac{|a(x_i) - y_i|}{|y_i|} \quad (3.11)$$

Данный коэффициент не имеет размерности и очень просто интерпретируем. Он означает на какой процент отличается фактическое значение от прогнозного. Основная проблема этой ошибки — ее нестабильность.

Huber loss. Одним из вариантов объединения абсолютной и квадратичной функции потерь, учитывающим преимущества как одной, так и второй модели является функция потерь Хубера:

$$L_\delta(y, a) = \begin{cases} \frac{1}{2}(y - a)^2, & \text{если } |y - a| < \delta \\ \delta(|y - a| - \frac{1}{2}\delta), & \text{если } |y - a| \geq \delta \end{cases} \quad (3.12)$$

С помощью параметра δ можно регулировать выбросы. Если параметр небольшой, то функция ведет себя квадратично только в окрестности нуля. Если же наоборот его увеличивать, то даже для значительных отклонений $(a - y)$ штраф будет вести себя квадратично, и при обучении будет большой акцент на их уменьшение.

Log-Cosh. Для некоторых моделей необходима непрерывная вторая производная функции ошибки. Для функции потерь Хубера это не так, данный недостаток отсутствует у функции потерь log-cosh:

$$L(y, a) = \log \cosh(a - y). \quad (3.13)$$

Предельное поведение этой функции аналогично функции потерь Хубера, для небольших отклонений оно квадратично, а для больших - линейное.

Описанные выше функции потерь изображены на (рис. 4).

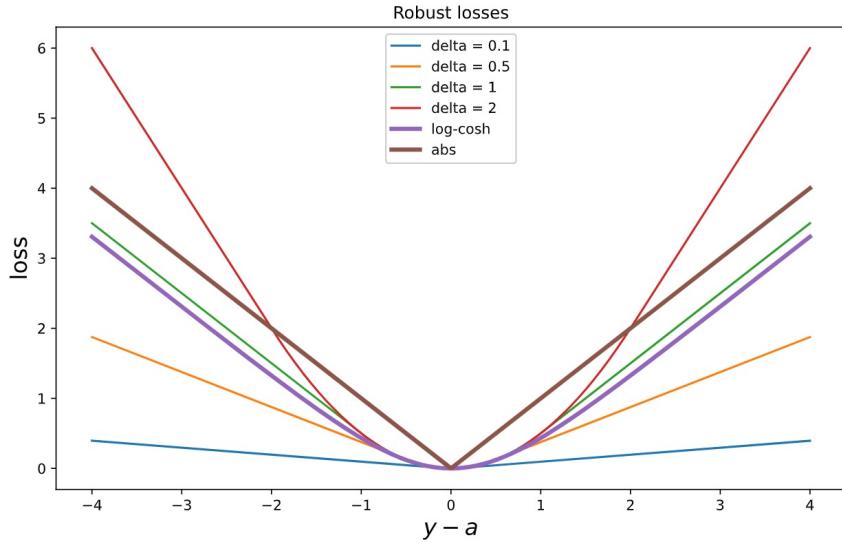


Рисунок 4 – Визуализация функций потерь

3.5.3 Оценивание качества модели

Для оценки качества исследуемой модели размеченные данные разделяются на две части: обучающую и контрольную выборки. На первой будет происходить обучение модели, на второй будем производить тестирование. Так как результат существенно зависит от разбиения данных, размеченные данные разбиваются на k блоков X_1, \dots, X_k примерно одинакового размера. После чего обучается k моделей $a_1(x), \dots, a_k(x)$, причем i -я модель обучается на всех объектах кроме блока i . После этого качество модели оценивается по тому блоку, который не участвовал в обучении, а результат усредняется:

$$CV = \frac{1}{k} \sum_{i=1}^k Q(a_i(x), X_i). \quad (3.14)$$

Заметим, что в нашей задаче данные имеют структуру временного ряда, поэтому случайно их перемешивать будет некорректно. Решением этой проблемы может быть подход **time series cross-validation**.

Суть метода заключается в том, что мы начинаем обучать модель на небольшом отрезке временного ряда от начала до некоторого t , затем делаем прогноз на $t + n$ и считаем ошибку. На следующем шаге расширяем обучающую выборку до $t + n$ и прогнозируем с $t + n$ до $t + 2n$. Так мы продолжаем сдвигать тестовый отрезок ряда до тех пор пока не достигнем последнее доступное наблюдение. Схематически логику можно представить так (рис. 5).

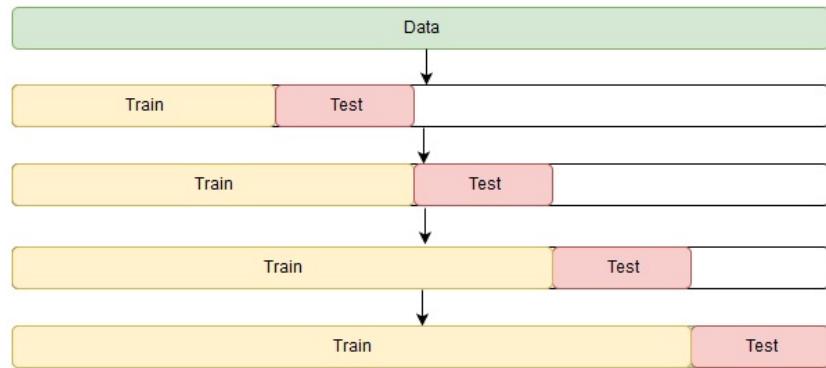


Рисунок 5 – Кросс-валидация временного ряда

3.5.4 Обучение линейной регрессии

Довольно распространенный способ обучения линейной регрессии с использованием среднеквадратичной ошибки. В этом случае получаем задачу оптимизации:

$$\frac{1}{l} \sum_{i=1}^l (\langle \omega, x_i \rangle - y_i)^2 \rightarrow \min_{\omega}. \quad (3.15)$$

Пусть X — это матрица "объекты-признаки" y — вектор ответов, ω — вектор параметров, тогда задачу можно переписать в матричном виде:

$$\frac{1}{l} \|X\omega - y\|^2 \rightarrow \min_{\omega}, \quad (3.16)$$

здесь используется обычная L_2 -норма. Если найти экстремум данного функционала, получим:

$$\omega = (X^T X)^{-1} X^T y. \quad (3.17)$$

Безусловно наличие явной формулы для оптимального веса векторов является большим преимуществом линейной регрессии с квадратичным функционалом. Но данная форму-

ла не всегда применима. Так обращение матрицы X — это сложная операция с кубической сложностью от количества признаков, данная проблема решается использованием линейных методов оптимизации. Но также матрица $X^T X$ может быть вырожденной или плохо обусловленной, данную проблему приходится решать регуляризацией.

Вывод аналитических формул в машинном обучении редко выполним, поэтому был разработан подход, в рамках которого модель можно обучать для широкого класса функционалов.

3.5.5 Градиентный спуск

Оптимизационные задачи вида 3.16 можно решать итерационно с помощью градиентных методов. Основное свойство антиградиента — указывать направление наискорейшего убывания функции в точке. Пусть $\omega^{(0)}$ — начальный набор параметров, который можно задать нулевым или сгенерировать случайным образом. Градиентный спуск заключается в повторении следующих шагов до сходимости:

$$\omega^{(k)} = \omega^{(k-1)} - \eta_k \nabla Q(\omega^{(k-1)}), \quad (3.18)$$

где $Q(\omega)$ — значение функционала ошибки для набора параметров ω , η_k — длина шага, которая позволяет контролировать скорость движения.

Останавливать схождение итерационного процесса можно, например, при близости градиента к нулю или при слишком малом изменении вектора весов. Для градиентного спуска имеет место следующая оценка сходимости:

$$Q(\omega^{(k)}) - Q(\omega^*) = O\left(\frac{1}{k}\right). \quad (3.19)$$

3.5.6 Градиентный бустинг

В основе этого подхода лежит идея комбинирования нескольких моделей для получения более точного результата. Ансамблем называется набор предсказателей, которые вместе дают результат. Ансамбль, в котором предсказатели выстроены последовательно, называется бустинг. Градиентный бустинг — это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля базовых предсказывающих моделей, чаще всего решающих деревьев[13–16].

Для начала рассмотрим **применение бустинга в задаче регрессии**, а именно обратимся к задаче минимизации квадратичного функционала:

$$\frac{1}{2} \sum_{i=1}^l (a(x_i) - y_i)^2 \rightarrow \min_a \quad (3.20)$$

Итоговый алгоритм будем искать в виде суммы базовых моделей $b_n(x)$:

$$a_N(x) = \sum_{i=1}^N b_n(x), \quad b_n \in \mathcal{A}. \quad (3.21)$$

Построим первый базовый алгоритм:

$$b_1(x) := \arg \min_{b \in \mathcal{A}} \frac{1}{2} \sum_{i=1}^l (b(x_i) - y_i)^2. \quad (3.22)$$

Для большинства семейств алгоритмов эта задача решается просто. Далее необходимо рассчитать остатки на каждом объекте — расстояния от ответа алгоритма до истинного ответа:

$$s_i^{(1)} = y_i - b_1(x_i). \quad (3.23)$$

Если прибавить эти остатки к ответам на обучающей выборке, то алгоритм не будет допускать ошибок на обучающей выборке. Поэтому будет разумно построить второй алгоритм, таким образом, чтобы его ответы были как можно ближе к остаткам:

$$b_2(x) := \arg \min_{b \in \mathcal{A}} \frac{1}{2} \sum_{i=1}^l (b(x_i) - s_i^{(1)})^2. \quad (3.24)$$

И так далее каждый следующий алгоритм будем настраивать на остатки предыдущих:

$$s_i^{(N)} = y_i - \sum_{n=1}^{N-1} b_n(x_i) = y_i - a_{N-1}(x_i), \quad i = 1, \dots, l; \quad (3.25)$$

$$b_N(x) := \arg \min_{b \in \mathcal{A}} \frac{1}{2} \sum_{i=1}^l (b(x_i) - s_i^{(N)})^2. \quad (3.26)$$

Стоит отметить и тот факт, что остатки могут быть найдены как антиградиент функции потерь по ответу модели, посчитанный в точке ответа уже построенной композиции:

$$s_i^{(N)} = y_i - a_{N-1}(x_i) = -\frac{\partial}{\partial z} \frac{1}{2} (z - y_i)^2 \Big|_{z=a_{N-1}(x_i)} \quad (3.27)$$

Получается, что выбирается такой базовый алгоритм, который как можно сильнее уменьшит ошибку композиции, так как он близок к антиградиенту функционала на обучающей выборке. Рассмотрим это свойство подробнее, а также обобщим его на другие функции потерь.

Градиентный бустинг. Рассмотрим некоторую дифференцируемую функцию потерь $L(y, z)$. Будем строить взвешенную сумму базовых алгоритмов:

$$a_N(x) = \sum_{n=0}^N \gamma_n b_n(x), \quad (3.28)$$

где $b_0(x)$ — начальный базовый алгоритм, как правило коэффициент при нем берут равным единице, а сам алгоритм выбирают максимально простым, для задач регрессии это обычно средний ответ:

$$b_0(x) = \frac{1}{l} \sum_{i=1}^l y_i. \quad (3.29)$$

Применим индукционный подход и предположим, что мы построили композицию $a_{N-1}(x)$ из $N - 1$ алгоритма, и хотим теперь выбрать следующий базовый алгоритм $b_N(x)$ таким образом, чтобы как можно сильнее уменьшить ошибку:

$$\sum_{i=1}^l L(y_i, a_{N-1}(x_i) + \gamma_N b_N(x_i)) \rightarrow \min_{b_N, \gamma_N}. \quad (3.30)$$

Для решения этой задачи предположим, что в качестве алгоритма $b_N(x)$ мы могли бы рассмотреть произвольную функцию, тогда задача сводится к необходимости понять какие значения s_i она должна принимать на объектах обучающей выборки:

$$\sum_{i=1}^l L(y_i, a_{N-1}(x_i) + s_i) \rightarrow \min_{s_1, \dots, s_l}. \quad (3.31)$$

Конечно, можно потребовать $s_i = y_i - a_{N-1}(x_i)$, но такой подход никак не учитывает особенности функции потерь $L(y, z)$, поэтому правильнее предложить s_i противоположным производной функции потерь в точке $z = a_{N-1}(x_i)$:

$$s_i = -\left. \frac{\partial L}{\partial z} \right|_{z=a_{N-1}(x_i)}. \quad (3.32)$$

При таком подходе мы сдвинемся в сторону скорейшего убывания функции потерь. При этом интересным свойством оказывается совпадение вектора сдвигов $s = (s_1, \dots, s_l)$ с антиградиентом:

$$\left(-\left. \frac{\partial L}{\partial z} \right|_{z=a_{N-1}(x_i)} \right)_{i=1}^l = -\nabla_z \sum_{i=1}^l L(y_i, z_i)|_{z_i=a_{N-1}(x_i)}. \quad (3.33)$$

При таком выборе сдвигов по сути происходит один шаг градиентного спуска, при движении в сторону наискорейшего убывания ошибки на обучающей выборке. Речь идет о градиентном спуске в l -мерном пространстве предсказаний алгоритма на объектах обучающей выборки.

Теперь, когда есть понимание, какие значения должен принимать новый алгоритм на обучающей выборке необходимо по данным значениям, заданным в конечном числе точек построить функцию определенную на всем пространстве объектов. По сути это

решение классической задачи обучения с учителем, воспользуемся простым функционалом, а именно среднеквадратичной ошибкой, для построения базового алгоритма, приближающего градиент функции потерь на обучающей выборке:

$$b_N(x) = \arg \min_{b \in \mathcal{A}} \sum_{i=1}^l (b(x_i) - s_i)^2. \quad (3.34)$$

Важно, что на этом шаге мы оптимизируем функцию потерь независимо от функционала исходной задачи, так как вся информация о функции потерь L находится в антиградиенте s_i и на данном шаге решается лишь задача аппроксимации функции по l точкам. Естественно можно использовать и другие функционалы, но среднеквадратичной ошибки обычно оказывается достаточно.

После того, как новый базовый алгоритм найден, можно подобрать коэффициент при нем по аналогии с наискорейшим градиентным спуском:

$$\gamma_N = \arg \min_{\gamma \in \mathbb{R}} \sum_{i=1}^l L(y_i, a_{N-1}(x_i) + \gamma b_N(x_i)). \quad (3.35)$$

Подход с аппроксимацией антиградиента базовыми алгоритмами, описанный выше, и называется градиентным бустингом. Данный метод представляет собой поиск лучшей функции, восстанавливающей зависимость ответов от объектов, в пространстве всех возможных функций. Поиск этой функции происходит с помощью псевдоградиентного спуска, каждый шаг делается вдоль направления, задаваемого некоторым базовым алгоритмом. При этом сам базовый алгоритм выбирается так, чтобы как можно лучше приближать антиградиент ошибки на обучающей выборке.

3.5.7 Нейронные сети

Для начала необходимо определить понятие нейрона в машинном обучении. Пусть на вход подается n величин x_1, \dots, x_n бинарных признаков, описывающих объект x . Эти признаки будем трактовать как импульсы, поступающие на вход нейрона через n входных каналов. Будем считать, что импульсы попадая в нейрон складываются с весами $\omega_1, \dots, \omega_n$. Если суммарный импульс превышает заданный порог активации ω_0 , то нейрон возбуждается и выдает на выходе один результат, иначе другой, в основе выбора лежит функция активации. Таким образом работу нейрона можно представить как вычисление функции:

$$a(x, \omega) = \sigma(\langle \omega, x \rangle) = \sigma \left(\sum_{i=1}^n \omega_i f_i(x) - \omega_0 \right) \quad (3.36)$$

где $\sigma(z)$ — функция активации, схематично принцип работы нейрона изображен на (рис. 6).

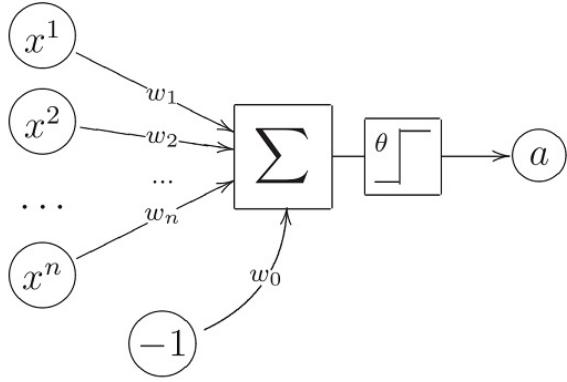


Рисунок 6 – Нейрон

Нейронной сетью называется композиция таких нейронов и визуально может быть представлена следующим образом[19] (рис. 7).

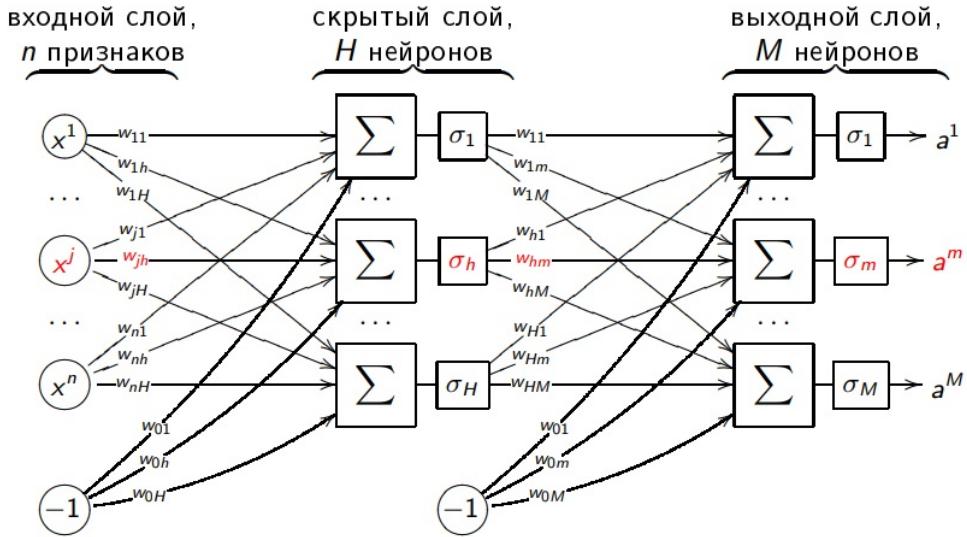


Рисунок 7 – Нейронная сеть

Обучение нейронной сети в задачах регрессии схоже с градиентным спуском. Пусть $\{x_i, y_i\}_{i=1}^N$ — обучающая выборка, $a(x, \omega)$ — исследуемая нейросетевая модель, $L(y, z)$ — функция потерь, тогда задача обучения может быть представлена, как:

$$Q(\omega) = \sum_{i=1}^N L(y_i, a(x_i, \omega)) \rightarrow \min_{\omega}. \quad (3.37)$$

Важным преимуществами нейросетевых алгоритмов является возможно наделять архитектуру сети нужными свойствами, принимать на входе и генерировать на выходе сложные структурированные данные. Однако, архитектуру необходимо подбирать вручную и моделям свойственно сильное переобучение. Нейронные сети хорошо подходят для решения задач на однородных входных данных, например изображениях. При регрессии на неоднородных табличных данных при применении классических моделей показывают не лучший результат согласно исследованиям ??, исследование работы

нейросетевых моделей в рамках данной работы не проводилось.

3.6 Обзор алгоритмов кластеризации в машинном обучении

Пусть дана выборка объектов $X = (x_i)_{i=1}^l$, $x_i \in \mathbb{X}$. Задача кластеризации заключается в том, чтобы необходимо выявить в данных K кластеров — таких областей, что объекты внутри одного кластера похожи друг на друга, а из разных кластеров нет. Если определять задачу формально, то требуется построить алгоритм $a : \mathbb{X} \rightarrow 1, \dots, K$, определяющий для каждого объекта номер его кластера.

3.6.1 Метрики качества кластеризации

Подходы к оценке качества кластеризации можно разделить на два ключевых: внутренний и внешний. В основе внутреннего лежат некоторые свойства выборки и кластеров, а внешнего на некоторых дополнительных данных, например информации об истинных кластерах.

Если известно истинное распределение объектов по кластерам, то задачу можно рассматривать как многоклассовую классификацию. Так как в текущем исследовании эталонных данных о распределении нет, то этот подход далее рассмотрен не будет.

Далее будут рассмотрены внутренние метрики качества.

Внутриклассовое расстояние:

$$\sum_{k=1}^K \sum_{i=1}^l [a(x_i) = k] \rho(x_i, c_k), \quad (3.38)$$

где c_k — центр кластера, $\rho(x, z)$ — некоторая функция расстояния. Этот функционал требуется минимизировать.

Межклассовое расстояние:

$$\sum_{i,j=1}^l [a(x_i) \neq a(x_j)] \rho(x_i, x_j). \quad (3.39)$$

Этот функционал необходимо максимизировать.

Индекс Данна:

$$\frac{\min_{1 \leq k \leq k' \leq K} d(k, k')}{\max_{1 \leq k \leq K} d(k)}, \quad (3.40)$$

где $d(k, k')$ — расстояние между кластерами k и k' , например можно взять евклидово расстояние между центрами, а $d(k)$ — внутрикластерное расстояние для k -го кластера. Данный индекс необходимо максимизировать.

3.6.2 Метод k -средних

Одним из наиболее популярных методов кластеризации является $K - Means$, который оптимизирует внутрикластерное расстояние 3.38, используя квадрат евклидовой метрики. В данном функционале две степени свободы: центры кластеров c_k и распределение объектов по кластерам $a(x_i)$. Выберем для этих величин произвольные начальные приближения, а затем пошагово будем их оптимизировать.

Фиксируем центры кластеров, внутрикластерное расстояние будет минимальным при условии, что каждый объект будет относиться к тому кластеру, чей центр является ближайшим:

$$a(x_i) = \arg \min_{1 \leq k \leq K} \rho(x_i, c_k). \quad (3.41)$$

Фиксируем распределение объектов по кластерам. После чего внутрикластерное расстояние с квадратом евклидовой можно продифференцировать по центрам кластеров и вывести аналогичные формулы для них:

$$c_k = \frac{1}{\sum_{i=1}^l [a(x_i) = k]} \sum_{i=1}^l [a(x_i) = k] x_i. \quad (3.42)$$

Повторяя эти шаги до сходимости, алгоритм выдает распределение объектов по кластерам. Стоит отметить, что результат работы $k - means$ существенно зависит от начального приближения

3.6.3 Иерархическая кластеризация

Описанный выше метод кластеризации имеет плоскую структуру кластеров. Для части задач бывает необходимо построить иерархию вложенных кластеров, в которой верхним уровнем является один большой кластер, а на нижнем уровне l кластеров, состоящих из 1 объекта каждый. Одним из таких подходов является восходящая кластеризация, на нижнем уровне все объекты принадлежат к отдельным кластерам: $C^l = \{\{x_1\}, \dots, \{x_l\}\}$. Каждый следующий уровень C^j получается путем объединения двух наиболее похожих кластеров с предыдущего уровня $C^{j+1} = \{X_1, \dots, X_{j+1}\}$. При этом схожесть кластеров можно определять с помощью некоторой функции $d(X_m, X_n)$, например это может быть расстояние между центрами кластеров.

4 Предлагаемые изменения

На фоне обостренной быстро меняющейся экономической ситуации в стране, а также крупных логистических проблем в строительной отрасли, мониторинг финансируемых объектов стройки, по которым банк является гарантом, важен как никогда. Поэтому для понимания необходимости заложения резервов по строительным объектам и прогноза соответствия фактических работ по строительному объекту и план-графиков, необходимо оценивать фактическую степень готовности объекта. Как унифицированный параметр-индикатор в данном исследовании предлагается дата полной готовности объекта.

4.1 Формальная постановка задачи

Необходимо спрогнозировать фактическую дату готовности строительного объекта на отчетную дату на основании актуальных данных об объекте строительства и девелопере. Так как прогноз значения с календарным типом данных является затруднительным, будем прогнозировать число дней, которое пройдет с отчетной даты до даты готовности. Прогноз будет реализован с применением алгоритмов машинного обучения.

В качестве критериев оценки качества работы алгоритмов предлагается сравнение ошибок: MSE, R^2 , MAE и MAPE.

4.2 Входные данные

В роли основного источника данных выступает закрытое внутреннее корпоративное хранилище данных "ПАО Сбербанк" (рис. 8) Sber Data Factory или Фабрика данных. Данные, интересующие нас в рамках решаемой задачи, реплицируются из трех основных источников: Единой информационной системы жилищного строительства (ЕИСЖС), Единого ресурса застройщиков (ЕРЗ)[4] и внутрибанковских сервисов. Из последних тяготеет количественная аналитика по объектам ипотеки, эскроу счетов и проектному финансированию. Единый ресурс застройщиков предлагает агрегированные данные продаж, собранные на базе статистики агентств и самих девелоперов. ДОМ.РФ же при подключении к API ЕИСЖС предлагает некоторые аналитические данные собранные на базе собственных моделей и статистики дочернего банка Банк ДОМ.РФ, а также набор данных, представленных на открытом портале [наш.дом.рф](#)[3]. Последние в обязательном порядке поставляются застройщиками в виде соответствующих документов: проектных деклараций, разрешительной документации, проектной документации и отчетности застройщика. Здесь важно заметить, что в проектной декларации, публикуемой ежемесячно, фигурирует дата планового окончания строительства, прогно-

зируемая застройщиком. Эта data будет выбрана в качестве значений для обучающей выборки.

Обработка данных выполнена при помощи Jupyter Notebook на языке программирования python.

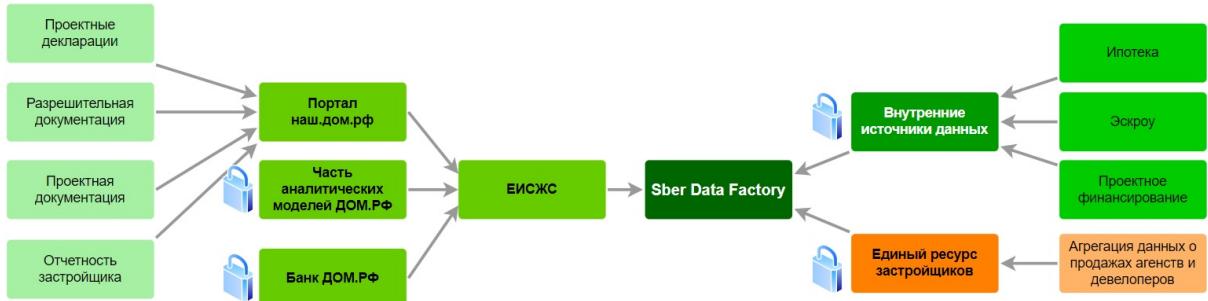


Рисунок 8 – Схема источников данных

Данные из всех систем связываются по ключевым идентификаторам и сджойниваются в одну матрицу признаков посредством системы работы с данным Hadoop.

4.3 Описание моделей

В данном исследовании сравнивается результат работы двух аналитических прогнозных модели машинного обучения: линейной регрессии и градиентного бустинга. Также для каждой из моделей проверяется гипотеза эффективности применения предварительной кластеризации объектов строительства. Схематичное изображение дататракта исследованных моделей представлено на (рис. 9).

Так как предлагается использовать моделей обучения с учителем, размеченную выборку разделяем по эпохам. В контрольную выборку попадают 3 последних размеченных месяца, в обучающую — все остальные.

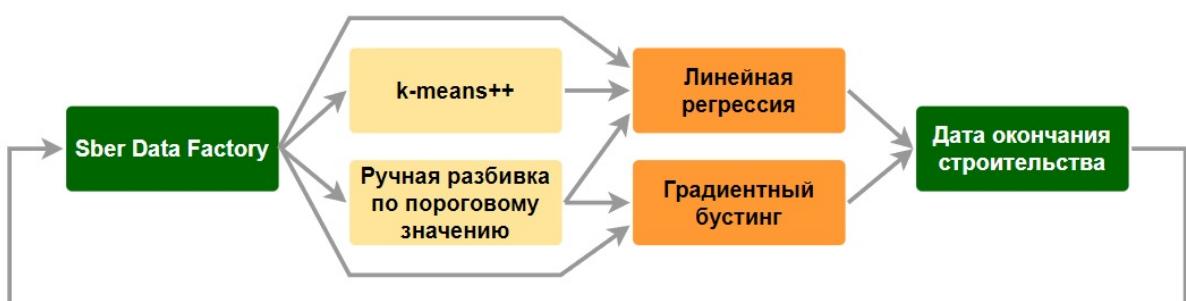


Рисунок 9 – Схематичное изображение исследованных моделей

4.3.1 Этап предобработки данных

На этапе предобработки данных исследовалась гипотеза, о возможности сгруппировать каким-либо образом объекты строительства для коллективного анализа и применения корректирующих мер на схожие объекты.

В качестве группирующих алгоритмов использовались алгоритм кластеризации k –means⁺⁺ и классификация по порогу фильтрующего значения.

k – means⁺⁺. Реализация алгоритма взята из библиотеки sklearn с модификацией выбора первого приближения. Алгоритм кластеризации неплохо себя показал, однако потребовал дополнительного времени для обучения. Кластеризация производилась с входным параметром 3, 4, и 5, задающим предполагаемое количество кластеров. Такие значения подобраны из соображений дальнейшей работы с моделями. Обслуживать более 5 регрессионных моделей сложно, а при количестве меньшем 3 кластеризация теряет смысл исходя из природы данных. Результат работы алгоритма представлен на (рис. 10), взята наиболее репрезентативная проекция, координаты не интерпретируемые. Однако видна тенденция к удачному разделению множества объектов на 4 группы это подтверждают и количественные метрики оценки ошибки в Таблице 2. Ввиду большого набора признаков в исходном датасете и небольшого количества объектов определить степень влияния конкретного признака на попадание объекта в тот или иной кластер становится проблематично, что снижает общую интерпретируемость модели.

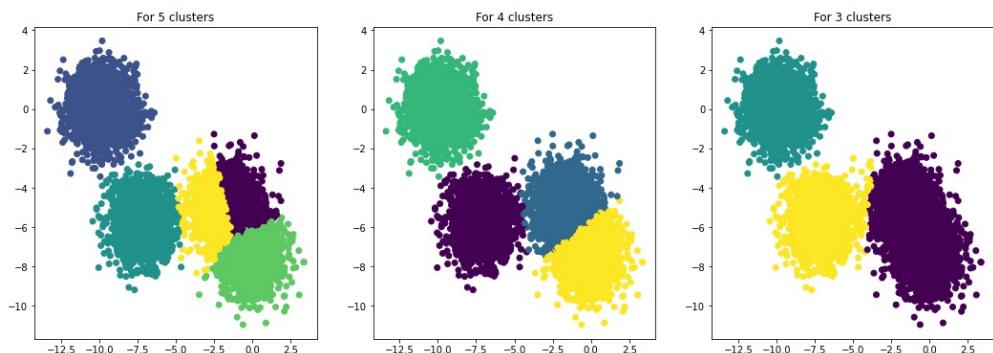


Рисунок 10 – Результат работы k-means для разного количества кластеров

Таблица 2 – Сравнение метрик качества $k-means^{++}$ для разного количества кластеров

Модель	Внутреннее расстояние	Внешнее расстояние	Индекс Данна
$k-means^{++}$ 5	18,7	28,9	1,15
$k-means^{++}$ 4	15,9	28,1	1,31
$k-means^{++}$ 3	19,4	23,9	1,05
Пороговая классификация	20,2	27,4	1,23

Пороговая классификация. Здесь, проверялась гипотеза о возможности разделить объекты недвижимости по уровню, а именно, за основу был взят признак f :

$$a_i \leqslant \left(f = \frac{p_j \cdot |R|}{\sum_{k \in R} p_k} \right) \leqslant a_{i+1}, \quad (4.1)$$

где p_j — актуальная стоимость строительства исследуемого объекта, R — все объекты жилищного строительства в регионе, $|R|$ — количество этих объектов, а a_i — выведенные эмпирически пороговые значения. Таким образом, по сути объект строительства оценивается по отношению стоимости его строительства к стоимости среднего объекта стройки в регионе, и на основе пороговых значений этого признака происходит классификация. На (рис. 11) представлена визуализация значения признака оценки. Не трудно заметить, что точки можно удачно разделить на 4 группы. На основе данных были выбраны 3 пороговых значения по центрам отрезков крайних точек групп: $a_1 \approx 0,8$; $a_2 \approx 1,2$; $a_3 \approx 1,5$.

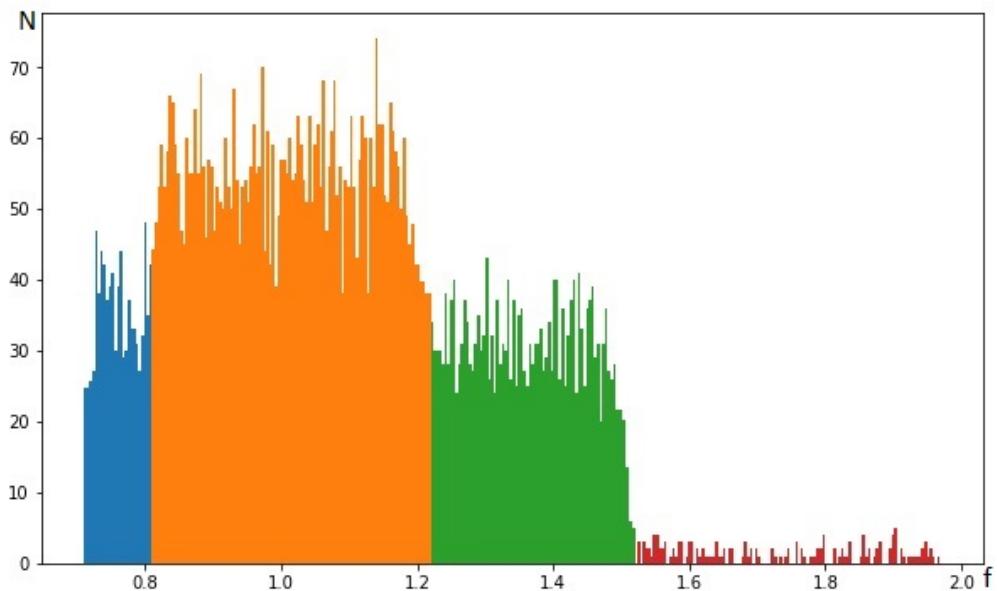


Рисунок 11 – Распределение значений признака f

Заметим, что величины ошибок для этого подхода не сильно хуже результатов работы алгоритмов кластеризации (Таблица 2). Однако, интерпретируемость пороговой классификации значительно выше, так выделяются 4 слоя объектов строительства, условно их можно назвать: "эконом" , "средний класс" , "бизнес" и "элитное" жилье. Интерпретируемость высока еще и потому, что такой терминологии придерживаются маркетологи крупных девелоперских компаний.

Резюмируя, разница в величине ошибки $\approx 10\%$ на данном этапе не является критической, так как напрямую не влияет на точность модели, а только на то, в какую регрессионную группу попадет объект. Поэтому на данном этапе в силу лучшей интерпретируемости предполагаем использование модели классификации с большим приоритетом.

4.3.2 Ядро модели

В качестве ядра модели рассматриваются два регрессионных алгоритма: классическая линейная регрессия и градиентный бустинг.

Линейная регрессия. В применении данного подхода рассмотрены модель без предобработки и две модели с предобработкой данных: кластеризацией $k - means$ для четырех кластеров и пороговой классификацией. Реализация алгоритма взята из библиотеки sklearn.

Градиентный бустинг. При обучении модели, в основе, которой лежит градиентный бустинг были рассмотрены 2 подхода с применением пороговой классификации и без нее. Для обучения взята модель LightGBM[13] с открытым исходным кодом. Эта имплементация расширяет алгоритм градиентного бустинга, добавлен тип автоматического выбора объектов. Также заметим, что в основе этой модели лежат простые деревья решений, поэтому пороговая классификация, фактически представляющая из себя дерево из одного шага, может быть включена в бустинговую модель и поэтому в качестве отдельного шага большого смысла не имеет. Также стоит отметить, что LightGBM имеет стохастический характер алгоритма, поэтому для получения корректных оценок точности его стоит запустить несколько раз и отражать средний результат, именно такой подход и будет применен. Далее рассмотрим насколько теоретические выводы коррелируют с практикой.

4.4 Результаты работы алгоритмов

Сравнение результатов.

Показатели качества предложенных выше моделей (рис. 9) представлены в Таблице 3. Нетрудно заметить, что модель градиентного бустинга в среднем показала себя значительно лучше линейной регрессии. Если смотреть на линейную регрессию, то метрики качества модели существенно лучше при применении предварительной группировки, при этом классификация по f -метрике на практике показала себя лучше чем $k - means$. Последнее скорее всего связано с выбором начального приближения. Для градиентного бустинга же напротив пороговая классификация положительного результата не принесла, и оказывается лучше использовать алгоритм в классической вариации.

Таблица 3 – Показатели качества построенных моделей

Модель	MSE	R^2	MAE	MAPE
LinearRegression	61564	0,516	307	20%
$k - means^{++}$ & LinearRegression	45671	0,634	260	17,7%
f-clas & LinearRegression	43282	0,652	248	17,3%
f-clas & LightGBM	24467	0,808	224	13,6%
LightGBM	22009	0,827	196	12%

Индикация проблемы.

На основе исследуемой даты фактического окончания строительства может быть построен индикационный дашборд, который будет сигнализировать о проблемах с тем или иным финансируемым объектом жилищного строительства. Так были выбраны 2 пороговых значения: 3 года и 7 лет(рис. 12). Если разница между плановой датой окончания строительства, установленная в проектной декларации на начало года, и предсказанной фактической датой готовности прогнозируемой моделью превышает 3 года, то объект попадает в желтую зону и выводится соответствующая индикация. Если же разница — превысит 7 лет, то даже в условия текущей нестабильности на рынке недвижимости такое смещение более чем существенно, и объект попадает в зону повышенного внимания и индикатируется красным. Внутри каждого сектора возможны разные оттенки индикации, интервал смены оттенка - год.

Таким образом, такая система индикации позволяет верхнеуровнево оценить состояние портфеля финансируемых проектов, понять какая доля объектов требует внимания и оценить насколько остро стоит проблема. Так результат работы модели на первое июня показал, что доля объектов в красной зоне около 1,5%, а объектов в желтой зоне

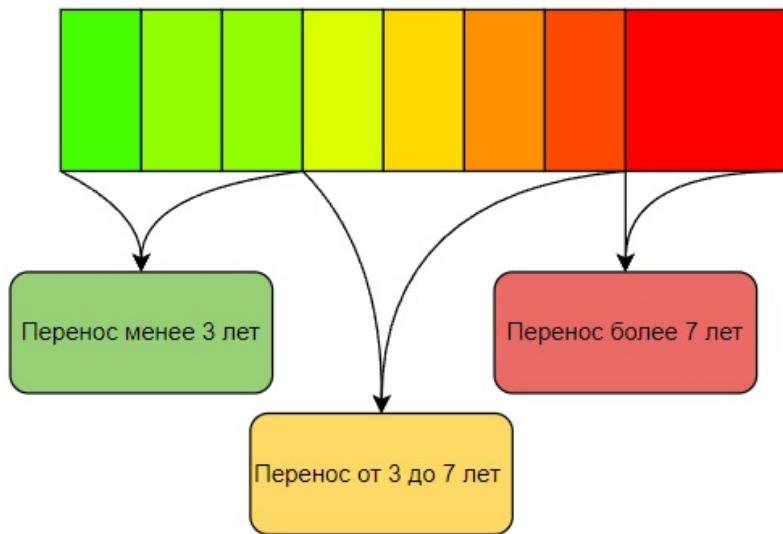


Рисунок 12 – Градиентная визуализация

порядка 20%. Таким образом объем требующих внимания риск-аналитиков объектов снижена до 1/5, при этом особое внимание необходимо уделить всего 1/200. То есть объем проводимых аналитических работ можно сократить практически в 5 раз.

Примеры работы модели.

Рассмотрим в качестве первого примера "Клубный дом на Пришвина" данный строительный объект изначально должен был быть сдан в конце третьего квартала 2019 года, однако плановый срок готовности постепенно переносился, планируемая стоимость строительства увеличивалась, а на момент подготовки модели у застройщика была просрочка срока передачи актуальной проектной декларации в ЕИСЖС. Но несмотря на все это моделью был спрогнозирован срок окончания готовности в 180 дней и дом обозначен индикацией не был. На текущий момент времени опубликована проектная декларация, в которой застройщик обещает сдать объект в начале 4 квартала 2022 года, то есть спустя 150 дней. Аналитически просто выявить причину оптимистичного прогноза модели, фактическое строительство дома уже завершено и сейчас производится получение разрешения на ввод в эксплуатацию объекта недвижимости. Однако на получение этого результата не потребовалось тратить ресурсы аналитиков.

Примером успешной подсветки проблемного объекта может послужить результат работы алгоритма для жилого дома "Виктория" в Костромской области. Плановый срок готовности объекта на начало года — конец 3 квартала 2022 года. Но фактически конструктив дома готов даже визуально не более чем на 60% (рис. 13). Более того застройщик уже переносил в прошлом плановую дату готовности дома более чем на 3 года. Размер чистой прибыли, согласно отчетности застройщика составляет 0 руб. за последний отчетный год. При этом фактическая стоимость строительства превыси-

ла плановую на 30%, а согласно проектной декларации к концу 1 квартала 2022 года, готовность объекта должна быть более 80%. Аналитические выводы о проблемности объекта подтверждают и результат работы алгоритма, предсказанный плановый срок готовности по нему составляет еще 3,5 года, что относит дом в категорию желтой индикации, требует дополнительного внимания риск-аналитиков и показывает состоятельность подхода.



Рисунок 13 – Сравнение рендера и фото жилого дома “Виктория”

5 Экономический эффект от внедрения модели

Говоря об эффекте от внедрения модели в первую очередь стоит сказать о том, что в текущих реалиях политика банка, поставившего задачу, заключается в повышении эффективности и производительности команд в условиях сохранения численности штата. А это значит, что на плечи сохранившегося числа сотрудников ложится больший объем аналитической работы с потенциально рисковыми объектами. Поэтому любая возможность ресурсоемко упростить и ускорить работу риск-менеджеров уже ценна.

В описанных выше условиях экономический эффект решаемой задачи можно рассмотреть в двух плоскостях. Оценить, существует ли вообще необходимость мониторинга степени готовности объектов строительства, а также сравнить эффективность машинной оценки по сравнению с оценкой, полученной аналитически, сотрудником банка.

Если оценивать стоимость строительства одного объекта, то согласно данным портала [наш.дом.рф\[3\]](#), она может составлять более 25 млрд. руб. И в случае со схемой финансирования по эскроу счетам, банк оказываясь гарантом, рискует практически всей суммой займа. Тем временем стоимость труда отдела риск-аналитики из 10 специалистов будет стоить банку порядка 30-35 млн. руб. ежегодно, что составляет менее промилле от потенциальной суммы риска только по одному проекту. Однако если речь идет о портфеле порядка 4-5 тысяч объектов строительства, очевидно, что затраты оказываются ничтожны даже при количественном увеличении штата на порядок.

Нетрудно заметить что охватить объем в 4 тысячи объектов строительства условному штату из 10 человек объективно проблематично, так на одного сотрудника будет приходится порядка 400 объектов мониторинга, даже при условии, что будет возможность уделять внимание не всему портфелю сразу, специалисту придется просматривать порядка 20 объектов за один рабочий день. Провести качественный анализ за столь короткое время крайне проблематично. Однако, предложенная модель позволяет верхнеуровнево оценить состояние портфеля проектов и корректно приоритизировать порядок анализа проектов. При этом объем критически важных для анализа проектов сокращается до 60, и может быть первично отработан за один день. В это же время объекты в желтой зоне индикации составляют только 20%, а значит объем нагрузки на риск-аналитиков снижается практически в 5 раз, достигая уже 4 объектов мониторинга за 1 рабочий день. Резюмируя, со стороны использования человеческого ресурса, такой подход, очевидно, более эффективный.

Рассмотрим финансовую сторону реализации и внедрения элемента системы автоматического мониторинга, представленного в работе. Для тестирования, отладки и настройки системы необходимо развернуть приложение в нескольких контурах банка.

Под каждый контур необходимо заложить серверные мощности и ресурс поддержки.

Стоимость внедрения решения включает подготовку серверных мощностей, затраты на разработку и разворачивание модели. Данная сумма для описанной модели не превысит 5 млн. руб., при этом важно сказать, что в эту сумму уже заложена возможность масштабирования модели на другие признаки, то есть мощность заложена с запасом. Стоимость обслуживания модели будет включать затраты на поддержку работоспособности кластеров, ресурс разработки для возможности доработки модели и затраты на поддержку серверных мощностей. Описанные издержки не должны превысить 4 млн. руб. ежегодно.

Для того, чтобы достичь аналогичного эффекта понижения трудозатрат в отношении одного риск-аналитика, необходимо было бы увеличить штат сотрудников в 5 раз, при этом потенциальные дополнительные издержки составили бы 120 млн. руб. ежегодно или 10 млн.руб. ежемесячно. Предполагая, что в обоих случаях речь идет о наших регулярных издержках, пересматриваемых не чаще чем раз в год, можем не учитывать ставку дисконтирования и предположить срок потенциальной окупаемости проекта. На тестирование и развертку приложения на боевых средах потребуется порядка 3 месяцев, в течение которых серверной инфраструктуре будет необходима поддержка. Так стартовые инвестиции в проект в совокупности с первым кварталом поддержки не превысят 6 млн., а потенциальная экономия уже на первый месяц боевого использования приложения тем временем составит порядка 10 млн. Таким образом, несложно заметить, что срок потенциальной окупаемости проекта составляет всего 4 месяца, что является довольно хорошим показателем при последующей потенциальной экономии на регулярных издержках более чем в 30 раз.

6 Заключение

В работе предложен метод оценки фактической даты готовности объектов в сфере жилищного строительства. Для улучшения оценки были построены модели предобработки данных основанные на кластеризации и пороговой классификации. Результаты прогноза моделей были оценены по соответствующим метрикам качества и было показано явное превосходство модели градиентного бустинга. Средняя абсолютная ошибка модели не превысила 13%. Показана возможность использования построенной модели для индикации проблемных объектов и оценки качества портфеля финансируемых объектов. Обоснована экономическая эффективность внедрения модели уже на краткосрочном горизонте планирования. Возможными направлениями развития могут быть: обобщение модели на другие важные для проектного мониторинга признаки, переход от прогнозирования макропараметров объектов строительства к динамически изменяющимся показателям; добавление в признаконое описание объектов, изучаемых моделью показателей, описывающих уровень импортозависимости строительства. Последнее стало немаловажным фактором при оценке любого финансируемого проекта, как критерий повышенного риска.

Список литературы

- [1] Антонов В.Г., Масленников В.В. и др. *journal = Палеотип т. у. . . л. . . р.*
- [2] Ассоциация банков России. Финансирование жилищного строительства в рамках достижения национальных целей развития до 2030 года // Аналитические материалы. — 2020.
- [3] Единая информационная система жилищного строительства [Электронный ресурс] // <https://наш.дом.рф/>.
- [4] Единый ресурс застройщиков [Электронный ресурс] // <https://erzrf.ru/>.
- [5] Кенчадзе Д.Д., Власенко Н.А. и др. Строительство в России. 2020 // Федеральная служба государственной статистики (Росстат). — 2020.
- [6] Мониторинг объема жилищного строительства | Минстрой России [Электронный ресурс] // <https://minstroyrf.gov.ru/trades/zhilishnaya-politika/8/>.
- [7] Никонова И.А. Проектный анализ и проектное финансирование // Альпина паблишер. — 2012.
- [8] Фатрелл Р., Шафер Д., Шафер Л. Управление программными проектами. Достижение оптимального качества при минимуме затрат // М.: Вильямс. — 2003. — С. 23–36.
- [9] Филлипс Дж. Управление проектами в области информационных технологий // ЛОРИ. — 2006.
- [10] Хейфец Е.Е. Анализ источников и механизмов финансирования девелоперских проектов жилищного строительства // Российский экономический интернет-журнал. — 2020. — № 4. — С. 56.
- [11] Центральный банк Российской Федерации | Банк России [Электронный ресурс] // <https://cbr.ru/>.
- [12] Чусавитина Г.Н., Макашова В.Н. Управление проектами по разработке и внедрению информационных систем // Общество с ограниченной ответственностью ФЛИНТА. — 2014.
- [13] Chen C., Zhang Q. et al. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion // Chemometrics and Intelligent Laboratory Systems. — 2019. — Vol. 191. — Pp. 54–64.

- [14] *Chen T., Guestrin C.* Xgboost: A scalable tree boosting system // *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. — 2016.
- [15] *Friedman J.* Greedy function approximation: a gradient boosting machine // *Annals of statistics*. — 2001. — Pp. 1189–1232.
- [16] *Friedman J.* Stochastic gradient boosting // *Computational statistics & data analysis*. — 2002. — Vol. 38, no. 4. — Pp. 367–378.
- [17] *Gulin A., Karpovich P.* Greedy function optimization in learning to rank // *Lection on the RuSSIR 2009 conference*. — 2009.
- [18] *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning: data mining, inference, and prediction // *MIT Press*. — 2009. — Vol. 2.
- [19] *Jaderberg M.* Decoupled neural interfaces using synthetic gradients // *International conference on machine learning*. — 2017. — Pp. 1627–1635.
- [20] *Jonathan T. Barron.* A general and adaptive robust loss function // *Google Research*. — 2019.
- [21] *Kerzner H.* Project management: a systems approach to planning, scheduling, and controlling // *John Wiley & Sons*. — 2017.
- [22] *Liu H.* Housing investment, stock market participation and household portfolio choice: Evidence from China's urban areas // *arXiv preprint arXiv:2001.01641*. — 2020.
- [23] *Mohri M., Rostamizadeh A., Talwalkar A.* Foundations of machine learning // *MIT Press*. — 2012.
- [24] *Nesterov Y.* Introductory lectures on convex optimization // *SpringerVerlag US* 2004. — 2004.
- [25] *Robbins H., Monro S.* A stochastic approximation method // *Annals of Mathematical Statistics*. — 1951.
- [26] *Schmidt M., Le Roux N., Bach F.* Minimizing finite sums with the stochastic average gradient // *Mathematical Programming*. — 2013.
- [27] *Yusupova A., Pavlidis N., Pavlidis E.* Adaptive dynamic model averaging with an application to house price forecasting // *arXiv preprint arXiv:1912.04661*. — 2019.