Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа экономики и управления
Кафедра «Информационные технологии в экономике»

ПРОЕКТ ПРОВЕРЕН
Рецензент, ген. директор
ООО «Уралмрамор»
_____(Ю.В. Абдурахимов)
«____» _____ 2020 г.

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

_____(Б.М. Суховилов)
«____» _____ 2020 г.

Разработка математических моделей для оценки вероятности ключевого действия клиентами банка

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ
ЮУрГУ–38.04.05.2020.463.ПЗ ВКР

Руководитель работы, д.т.н., профессор
_____(В.В. Мокеев)
«____» _____2020г.

Автор работы,
студент группы ЭУ-226
_____(М.Р. Умурзаков)
«____» _____2020 г.

Нормоконтролер, к.т.н., доцент
_____(Е.В. Бунова)
«____» _____2020 г.

Челябинск 2020

АННОТАЦИЯ

Умурзаков М.Р. Разработка математических моделей для прогнозирования совершения ключевого действия клиентов банка – Челябинск: ЮУрГУ, ЭУ–226, 92 с.,38 ил., 16 табл., библиогр. список – 68 наим., прил. – 0.

Выпускная квалификационная работа выполнена с целью повышения эффективности процесса прогнозирования совершения ключевого действия клиентов банка.

В квалификационной работе рассматривается процесс оттока клиентов, расмматриваются методы машинного обучения, производится их сравнения, строятся модели по выбранным методам, разрабатывается план коммерциализации проекта.

Основные задачи работы:

1) описание процесса маркетинга отношений;
2) теоретическое описание машинного обучения, его виды и алгоритмы;
3) анализ и сравнение имеющихся прогнозных методов машинного обучения;
4) формулировка требований к прогнозной модели;
5) построение прогнозной модели;
6) сравнительный анализ результатов прогнозирования;
7) формирование плана коммерциализации.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Актуальность темы.

В настоящее время отток клиентов является одной из основных проблем для финансовых организаций. Непрерывная конкуренция на рынке и высокая стоимость привлечения новых клиентов заставили организации сосредоточиться на более эффективных стратегиях удержания клиентов. Хотя банковский и финансовый секторы показывают низкие показатели оттока клиентов по сравнению с другими, потеря клиентов гораздо сильнее влияет на прибыль нежели в иных секторах экономики. Исходя из этого, управление оттоком клиентов играет жизненно важную роль для повышения долгосрочной прибыльности организации.

По теме оттока было проведено много исследований для понимания поведения клиентов при уходе из организации и определения причин данного ухода. Анализируя работы, связанные с клиентооттоком, нельзя не отметить, что произошло значительное развитие методов для эффективного моделирования оттока клиентов в банковском и финансовом секторах.

После анализа текущей конъюнктуры в сфере анализа оттока клиентов, становится понятно, что в данной сфере существуют некоторые проблемы.

В частности, одной из основных остается недостаток точности прогнозирования; в результате чего, банку или кредитной организации порой бывает сложно предоставить конкретный вид услуг клиенту до того, как он уйдет к конкурентам.

Также недостатком является тот факт, что данная проблема слабо разработана в трудах российских ученых.

С проблемой оттока клиентов банка столкнулась компания Santander Group.

Банк Сантандер, Santander Group — крупнейшая финансово-кредитная группа в Испании. Помимо Испании Santander занимает одно из ведущих мест в Великобритании и в ряде стран Латинской Америки, также представлена в США.

Исходя из вышесказанного появляется необходимость поиска путей, направленных на решения проблемы связанной с прогнозированием совершения ключевого действия клиента.

Теоретической и методологической основой магистерской диссертации являются труды зарубежных и отечественных ученых в области машинного обучения. Так, например, в российской литературе известны такие авторы как:

Боднар А.Ю[1], Владыка М.В.[2], Пальмов С.В.[6], Романов В.В.[9], Сергеенкова А.А.[11], Хаустова М.Н.[13], Прохоренкова Л.А[46].

В зарубежной литературе известны следующие авторы:

Лабрам А.[15], Чан Ц.[17], Бакинкс В[18], Эстэр М.[22], Фаркуад М.[24], Феррера Дж.[25], Сегал Дж.[26], Гарланд Р.[27], Го-эн З.[29], Гулин К[30], Халкиди М.[31], Хуанг Б.[32], Хванг Х.[34], Жанг Дж.[36], Лей Дж.[37], Кевени С.[39], Керамати А.[40], Джафари-Маранди Р.[41], Аббаси Ю.[41], Брейман Л.[42], Лин С.[45], Мэйсон Л.[48,49], Бакстер Дж.[48,49], Литл М.[50], Сохраб Махмуд М.[51], Хуанг М.[52], Мозер М.[53], Ние Дж.[54], Сарадхи В.[57], Шарма А[58], Ши Н.[59], Мутанен Т.[61], Чен Т.[62], Ху Т.[63], Дуан Т[64], Юи З.[67], Жанг Т.[68].

Рассмотрев различные точки зрения в отечественной и зарубежной литературе, можно прийти к выводу, что существуют различные методы прогнозирования, каждый из которых имеет свои преимущества и недостатки.

Объект исследования процесс обеспечения и улучшение качества обслуживания клиентов.

Предмет исследования методы прогнозирования совершения ключевого действия клиентов банка.

Цель исследования – повышение эффективности процесса прогнозирования потенциальной ценности услуги для клиентов.

На основе результатов прогнозирования, банк в дальнейшем может сделать вывод о том какая услуга будет важна для клиента и своевременно её предоставить, тем самым предотвратив его уход из банка.

Для достижения цели были поставлены следующие задачи:

1) описание понятий оттока клиентов, персонализации.
2) провести исследование и анализ современных подходов к прогнозированию оттока клиентов.
3) анализ имеющихся прогнозных методов машинного обучения.
4) формулировка требований к прогнозной модели.
5) построение прогнозной модели.
6) сравнительный анализ результатов прогнозирования методами машинного обучения.
7) разработка плана коммерциализации.

Новизна работы заключается в том, что на основе комплексного анализа:

1) рассмотрены и проанализированы методы прогнозирования.
2) проведен сравнительный анализ методов.
3) разработаны модели прогнозирования потенциальной стоимости услуги для клиентов, результаты которых можно использовать для оценки вероятности совершения ключевого действия клиентами банка.

Практическая значимость работы обусловлена применением результатов исследования на практике, для анализа оттока клиентов в банке.

ГЛАВА 1 МАРКЕТИНГ ОТНОШЕНИЙ: ПРОЦЕСС ОБЕСПЕЧЕНИЯ И УЛУЧШЕНИЕ КАЧЕСТВА ОБСЛУЖИВАНИЯ КЛИЕНТОВ БАНКА

1.1 Сущность процесса маркетинг отношений

Каждое человеческое взаимодействие и транзакция строятся вокруг отношений. Сети взаимоотношений – это основа человеческого общества. Неудивительно, что этот фундаментальный факт был признан и исследован всеми предприятиями, в которых строились бизнес-стратегии вокруг клиента, и целью ставилось установление отношений с каждым клиентом.

Маркетинг отношений поэтому развился не только как маркетинговая стратегия, но и был основой, на которой компании строят свои основные ценности и этику. «Маркетинг взаимоотношений» определяет рамки, в которых компания может охватывать и ориентироваться на внешних рынках, как для конечного потребителя, так и для деловых партнеров, поставщиков и продавцов. Маркетинг взаимоотношений не ограничивается только клиентами и поставщиками, а расширен для охвата внутренних сотрудников, а также является эффективным способом привлечения лучших специалистов. Если отсканировать какую-либо рекламу ведущей корпорации в газете, можно увидеть, что основная часть рекламы для подбора персонала связана с историей Компании, культурой и усилиями по охвату потенциальных сотрудников. Рекламные объявления предназначены для того, чтобы повлиять на читателей, побуждая его подать заявку на работу.

В эпоху высоких технологий, когда концепции и инструменты маркетинга претерпели серьезные изменения с появлением электронной коммерции, онлайн-продаж, сетевого маркетинга, прямого маркетинга, бизнес-моделей B2B и B2C, маркетинг отношений стал основой, на которой строятся как бизнес-стратегии, так и маркетинговые стратегии.

В текущих реалиях организации начали распознавать и учитывать как человеческий фактор, так и эмоциональный фактор деловых отношений. Маркетинг взаимоотношений развивался как дисциплина, которая помогает

предприятиям выходить за рамки транзакций для долгосрочных деловых взаимодействий. Успешная стратегия маркетинга взаимоотношений помогает предприятию увеличивать свою прибыль на долгосрочной основе.

Фокус маркетинга взаимоотношений – создать долгосрочного постоянного клиента. Однако в центре внимания традиционной маркетинговой стратегии лежит привлечение большего количества клиентов. Хотя эти два подхода могут показаться слишком разными друг для друга, но в действительности применение обоих методов актуально в наше время. Таким образом, организация использует сочетание традиционной маркетинговой стратегии и маркетинга взаимоотношений.

Первым этапом любого цикла маркетингового планирования является оценка стратегии. При оценке важно выяснить, заинтересован ли клиент в каких-либо отношениях с организацией. Отношения были бы полезны для обеих сторон, если бы была возможность успеха. Если клиенты не находят никакой добавленной стоимости в отношениях, они не будут готовы вступить в них.

Еще одним моментом оценки является качество обслуживания. Если качество обслуживания будет на низком уровне, то менее вероятно, что потребители захотят иметь отношения с компанией. С другой стороны, если обслуживание клиента будет носить персонифицированный характер, вероятность долгих взаимоотношений с клиентами увеличивается. В случае бизнес-транзакций развитие отношений будет иметь важное значение.

При оценке стратегии организация хотела бы знать, на какого клиента она рассчитывает; активен ли клиент, а также стоимость обслуживания данного клиента.

В нынешних глобальных условиях динамика рынка постоянно меняется, поэтому организациям необходимо оценить время, затраченное на выполнение стратегии, и время, через которое данная стратегия принесёт результаты.

Маркетинг отношений может иметь огромное значение для банковского сектора. Клиенты хотят больше, чем просто одалживать деньги или использовать

свои сбережения. Они хотят персонализированных долгосрочных отношений с банком, построенных на доверии. Это означает, что они хотят больше, чем просто имя в базе данных. Точно так же пакеты услуг и продуктов представляют собой просто количественный ответ на запросы клиентов и не являются источником доверия или предпосылками для лояльности.

В своих лучших проявлениях они могут предложить удовлетворение сокращения затрат. Банки должны иметь расширенную базу данных с особыми финансовыми потребностями, в которой есть место для персонализации, уделяя больше внимания завоеванию доверия клиентов во время беспроигрышных долгосрочных отношений. Желаемое различие происходит от доверия и удовлетворения. Фактически, маркетинг отношений стал необходимостью, а не практикой огромного потенциала. Сегодня мы говорим о стратегиях построения отношений, ориентированных на клиентов и их реальные потребности, проходящих через экономические, поведенческие, эмоциональные и моральные фильтры. Удовлетворенность клиентов в течение всего жизненного цикла является основным основанием для их сохранения и будущей лояльности.

Растущее использование технологий позволяет активно развивать информационные и компьютерные информационные системы, используемые для содействия процессам принятия решений и управления взаимоотношениями с клиентами.

CBS (Core Banking System) увеличила эффективность. Это сделало работу также простой и прозрачной. Она объединяет все данные, профили клиентов и отзывы клиентов.

Компьютерная информационная система, такая как CRM, может улучшить информационный поток и процессы принятия решений между различными отделами банка.

Технологии самообслуживания, такие как банкоматы, мобильный банкинг и интернет-банкинг, являются важными инструментами маркетинга отношений. Эти банковские услуги просты в использовании, снижают затраты и экономят время.

Концепция управления взаимоотношениями с клиентами и управление жалобами помогают повысить лояльность клиентов.

Если банк поздравляет клиентов с днем рождения и юбилеем, это говорит нам о том, что банк использует персонифицированный подход к каждому клиенту, что будет благосклонно сказываться на их дальнейших взаимоотношениях.

Используя многоканальный подход, становится легко подойти к клиенту. Кроме того, использование мгновенных сообщений, электронной почты и системы коротких сообщений делает общение прозрачным и доступным.

Марк Дюркин и Барри Хоукрофт исследовали восприятие банкиров в Великобритании, Швеции и США в отношении интернета как инструмента маркетинга отношений. Было достигнуто согласие о том, что интернет играет ключевую роль в управлении взаимоотношениями, но было меньше согласия относительно скорости принятия клиентов и степени, в которой на это влияют стратегии банка[20].

Ифтихар Хуссейн, Мажар Хуссейн, Шахид Хуссейн и М.А. Саджид провели исследования в отдельных банках Пакистана. В исследовании изучалась и анализировалась стратегическая реализация CRM в отдельных банках Пакистана, выявлялись преимущества, а также факторы успеха и неудачи внедрения, и развивалось лучшее понимание влияния CRM на банковскую конкурентоспособность, а также обеспечивалось лучшее понимание того, что представляет собой хорошую практику CRM. Выяснилось, что CRM по-прежнему нуждается в активной повестке дня в Пакистане[33].

Джитес Пармар и Виджай Кумар Садананд изучили связь между методами CRM и лояльностью клиентов. Исследование состоит из двух частей. Первая часть называется опросом о лучших практиках CRM. Вторая часть, а именно исследование конкретного случая включает использование встроенного опроса лояльности клиентов. Результаты буквальной и теоретической репликации, выполненной с использованием метода сопоставления с образцом, указывают на отсутствие тесной связи между развертыванием лучших практик CRM в

запланированных коммерческих банках и уровнями лояльности розничных клиентов с высокой и средней стоимостью. Результаты также подразумевают, что переход к развертыванию CRM не может быть выгодной стратегией для розничных банков, особенно в индийском контексте[55].

Р.К. Уппал сравнил обслуживание клиентов относительно времени государственного сектора, частного сектора и иностранных банков в Амритсаре (Пенджаб). Выяснилось, что между тремя группами существует значительная разница. Электронные банки оказались более эффективными с точки зрения фактора времени. Было установлено, что это является очень важным фактором в перемещении клиентов в электронных банках[64].

Пиверс, Дуглас, Маршалл и Джэк оценили, что подтверждение транзакции важно для клиентов – посредством SMS-сообщения или внутри самого телефонного звонка. Клиенты оценили роль SMS для CRM как весьма желательную после денежных транзакций; они предпочитают версию банковской услуги телефонного звона. СМС подтверждение они оценили выше по качеству. Как следствие, разработанные инструменты и средства полезны для реализации банками стратегии CRM[56].

Доктор Ширмила Стэнли определила новые перспективы в банковском секторе с введением CRM. Предложенные ею меры защиты заключаются в том, что банки должны применять подходы построения отношений с клиентами, чтобы повысить ценность времени жизни клиента (CLV) и удовлетворенность клиента, что приведет к долгосрочным отношениям. Эта статья является попыткой увидеть факторы, необходимые для эффективной CRM в банковском секторе[60].

Кристин Энью, Мартин Бинкс, Брайан Чиплин исследовали степень удовлетворенности и удержания клиентов в британских банках и малом бизнесе, где ключевой стратегией является создание и поддержание лояльной базы данных. Они разработали предварительную модель. Был использовал дискриминантный анализ[21].

В текущем сценарии важность маркетинга отношений растет день ото дня. Банки должны поддерживать отношения с постоянными клиентами. Более того, сегодня клиенты стали более осведомленными, изощренными и напористыми, с растущим спросом на индивидуальные и инновационные продукты и услуги. Клиенты больше не заинтересованы в выкупе готовых решений, но нуждаются в средствах, которые соответствуют их бизнес-моделям и планам. Поэтому для банков крайне важно иметь прочные отношения со своими развивающимися клиентами, чтобы гарантировать, что они находятся в нужном месте в нужное время. Чтобы получить и создать преимущества, необходимо правильно управлять взаимоотношениями с клиентами, чтобы поддерживать, улучшать и развивать долгосрочные отношения между бизнесом и клиентами. Считается, что доверие, приверженность и связи играют важную роль в создании прочных отношений, которые создают как социальные, так и экономические выгоды, упомянутые выше, как для банков, так и для клиентов.

1.2 Сущность персонализации, и её место в маркетинге отношений

Затрагивая вопрос маркетинга отношений, нельзя не затронуть вопрос персонализации услуг, предоставляемых банком.

Хотя персонализация набирает обороты, определение еще не получило широкого распространения. К ошибочным описаниям относятся сегментирование предложений продуктов, настройка сообщений на главной странице и оцифровка пути клиента. Все это имеет решающее значение для обеспечения персонализации, но они не являются эквивалентами. Кроме того, предоставление уникального опыта клиентам по нецифровым каналам (таким как филиалы и колл-центры) может изменить ситуацию.

Истинная персонализация основана на глубоком понимании уникальных потребностей каждого клиента и организации набора адаптированного опыта по

цифровым и человеческим каналам. Персонализация потенциально создает беспроигрышный сценарий для банков и клиентов, которых они обслуживают.

По оценкам Boston Consulting Group, на каждые 100 миллиардов долларов активов, которые есть у банка, он может добиться роста доходов до 300 миллионов долларов, персонализируя взаимодействие с клиентами. Ожидаемо, что персонализированные банковские услуги обеспечат существенное конкурентное преимущество для тех банков, которые будут использовать персонализацию в течение следующих пяти лет.

С практической точки зрения, персонализированный банкинг означает предоставление нужного индивидуального опыта по нужному каналу в нужное время. Победившие банки могут восстановить доверие и обеспечить более глубокие и выгодные финансовые отношения с помощью простого уравнения[5]:

Частота взаимодействия × «вау-фактор» = доверие

(«Вау-фактор» отражает влияние каждого взаимодействия.) Хотя банковские услуги являются относительно редкими, многие люди часто взаимодействуют со своими банками, ежедневно используя мобильное приложение. Эта динамика создает возможность для банков иметь персонализированные взаимодействия, которые соответствуют контексту, уместны и ценны для клиентов.

По данным Aite Group, около ¾ опрошенных из выборки от 22 до 49 лет говорят, что хотели бы иметь своего собственного виртуального тренера по финансовому оздоровлению. Другие клиенты хотят, чтобы банки были больше похожи на других поставщиков услуг. (См. рисунок 1.) Другими словами, банки, которые предлагают соответствующие эквиваленты через управляемые услуги по оказанию финансовой помощи, одновременно анализируя данные, находятся в выигрышном положении, чтобы взять на себя новые и полезные роли[49].

Рисунок 1 – Опрос клиентов банков

Персонализация также стимулирует обмен информацией о продажах, хотя это не самое частое взаимодействие с клиентом. Каждый банковский клиент уникален, поэтому индивидуальные услуги будут лучше подходить, чем стандартные варианты. Один банк использовал персонализацию, чтобы поднять производительность продаж филиала более чем на 30%. В другой организации доходы выросли на 20% за три года[45].

Банки должны произвести революционные внутренние изменения, для оценки персонализации, направляя ее на этапы поиска, привлечения и удержания клиентов. Прежде всего, персонализация требует превращения клиента в центр внимания. Данная переработка требует реализации принципиально новых способов работы, которые требуют обновления ориентированных на клиента целей, создания стимулов для клиентов, переподготовки сотрудников, найма новых и создания нового набора аналитики и технологий. Масштабирование персонализации является чрезвычайно сложной задачей, и многие компании не выходят за рамки доказательств пилотных концепций и технологий.

Компании, которые предпринимают всеохватывающий поэтапный подход к персонализации в масштабе, как правило, не добиваются успеха.

Банки должны сосредоточиться на организации набора инструментов и ресурсов для персонализации. (См. рисунок 2.)

Рисунок 2 – Ресурсы для персонализации

Система должна иметь три основных элемента, которые формируют основу для персонализации в масштабе в банковской сфере[35]:

- **данные о клиенте.** Банковские системы должны обеспечивать единое общеорганизационное представление о каждом клиенте – представление, которое динамически отражает каждого клиента в любой момент.
- **персональный учебный план.** Предложения определяют желаемое поведение клиента и стратегию стимулирования такого поведения.
- **аналитический движок и рекурсивное обучение.** Используя машинное обучение и систематическое экспериментирование, банки обеспечивают гибкость и создают индивидуальные предложения и коммуникации, которые со временем развиваются. Способность к рекурсивному обучению, необходимая для персонализации – не меньше, чем способность непрерывно узнавать о каждом отдельном клиенте и дополнять эти знания с течением времени.

Для того, чтобы начать оценку персонализации, необходимо:

- определить несколько основных вариантов обслуживания клиентов, которые помогут банку добиться дифференциации.
- оценить текущие возможности персонализации. Банки, которые вкладывают средства в технологии, обычно имеют активы и ресурсы, которые они могут использовать.
- назначить старшего руководителя владение персонализацией в масштабе.
- оценить итеративность в шестимесячных волнах. Каждая волна должна предоставлять более полную версию целевого опыта при построении ключевых элементов системы.

1.3 Анализ работ, посвященных прогнозированию оттока клиентов банка

Отток клиентов тесно связан с уровнем удержания клиентов и лояльностью. Хванг и соавторы определяют уход клиента как самую важную проблему в высококонкурентной отрасли беспроводной связи. Их модель LTV (Модель пожизненной ценности клиента) предполагает, что скорость оттока клиентов оказывает сильное влияние на LTV, поскольку это влияет на продолжительность обслуживания и на будущие доходы. Хванг также определяет лояльность клиентов как индекс, который клиенты хотели бы остаться в компании. Отток отражает количество или процент постоянных клиентов, которые отказываются от отношений с поставщиком услуг[34].

Лояльность клиентов = 1 - показатель оттока

Моделирование оттока клиентов в чисто параметрической перспективе не подходит для контекста LTV, потому что функция удержания имеет тенденцию быть нестабильной, с пиками в даты окончания контракта.

Это позволяет отделу маркетинга, учитывая ограниченные ресурсы, быстро реагировать на возможность оттока клиентов.

Интеллектуальный анализ данных с помощью эволюционного обучения (DMEL) может показать причину или вероятность ухода; деревья решений, однако, мог показать только причину[17].

Го-ан и Вэй-дон сосредоточились на построении модели прогнозирования оттока клиентов с использованием метода опорных векторов в телекоммуникационной отрасли. Они сравнили этот метод с другими методами, такими как деревья решений, нейронные сети, наивный байесовский классификатор и логистическая регрессия. Результаты доказали, что метод опорных векторов является простым методом классификации с высокими возможностями, но при этом хорошей точностью[29].

Анил Кумар и Рави использовали интеллектуальный анализ данных для прогнозирования оттока клиентов по кредитным картам. Они использовали многослойный персептрон, логистическую регрессию, деревья решений, случайный лес, сеть радиально-базисных функций и метод опорных векторов[16].

Ние и соавторы построили модель прогнозирования оттока клиентов, используя логистическую регрессию и методы на основе деревьев решений в контексте банковской индустрии[54].

Шарма и Паниграхи применяли нейронные сети для прогнозирования оттока клиентов сфере услуг сотовой связи. Результаты показали, что нейронные сети могут прогнозировать отток клиентов с точностью выше 92%[58].

Сарадхи и Палшикар сравнили методы машинного обучения, использованные для построения модели прогнозирования оттока сотрудников[57].

В своем исследовании Лин и соавторы использовали грубую теорию множеств и методы принятия решений, основанных на правилах для выявления закономерностей, связанных с оттоком клиентов на счетах кредитных карт, с использованием графа транспортной сети[45].

Юи и соавторы применили методы нейронной сети, метод опорных векторов, деревья решений и метод расширенных опорных векторов для прогнозирования

оттока клиентов. Из изученных методов метод расширенных опорных векторов работал лучше всего[67].

Хуанг и соавторы представили логистическую регрессию на основе новых функций, линейный классификатор, нейронные сети, наивный байесовский классификатор, деревья решений, многослойный перцептрон и метод опорных векторов. В своих экспериментах каждая техника давала разные результаты[32].

Логистическая регрессия, наивный байесовский классификатор и многослойный перцептрон могут предоставить вероятности различного поведения клиентов. Фаркуад и соавторы использовали метод опорных векторов для прогнозирования оттока клиентов по банковским кредитным картам. Они представили гибридный подход для извлечения правил из данного метода для целей управления взаимоотношениями с клиентами[24].

Подход состоит из трех этапов, на которых:

1) Основанное на методе опорных векторов – рекурсивное устранение признаков применяется для сокращения набора функций;

2) полученный набор данных используется для построения модели;

3) используя наивный байесовский классификатор, генерируются древовидные правила.

Керамати и соавторы не только представили различные подходы к анализу данных и методам классификации, таким как деревья решений, нейронные сети, метод опорных векторов и метод k-ближайших соседей, но также сравнили характеристики этих подходов. В качестве примера они проанализировали данные иранской мобильной компании[40].

Лестер объясняет сегментационный подход в анализе оттока клиентов [43]. Она также указывает на важность правильных характеристик, изучаемых в анализе оттока клиентов. Например, в контексте банковского обслуживания эти изученные сигналы могут включать уменьшение баланса счета или уменьшение количества покупок по кредитной карте.

Подобный тип описательного анализа был проведен Кевени и соавторами [39]. Они изучали поведение клиентов в онлайн-сервисах на основе анкет, разосланных клиентам. Гарланд провел исследование о прибыльности клиентов в сфере персонального розничного банкинга [27].

Хотя их основное внимание уделяется ценности клиентов для исследований банка, они также исследуют продолжительность и возраст отношений с клиентами на основе прибыльности. Его исследование основано на опросе клиентов по почте, который помог ему определить долю клиента в кошельке, удовлетворенность и лояльность из качественных факторов.

Браян Грегори, в своей работе рассмотрел работу экстремального градиентного бустинга (вариация стандартного градиентного бустинга) при работе с данными банковских клиентов. Свою модель он использовал в соревновании WSDM Cup 2018 (Международное соревнование между учеными, занимающимися анализом больших данных) и занял в нём 1 место среди 575 команд[28].

Маркос Роберто Мачадо, также в своей работе рассмотрел эффективность работы градиентного бустинга через оценки лояльности клиентов банка. Однако, в своём исследовании он использовал другую вариацию градиентного бустинга – LightGBM. Данная методология, разработанная компанией Microsoft, является доработанной версией стандартного градиентного бустинга, алгоритмы которой позволяют обучать модель гораздо быстрее, чем делают стандартный GBM и другие его вариации. В своей работе он продемонстрировал работу LightGBM, сравнил данный метод XGBoost, и выявил, что LightGBM способен выдавать гораздо более точные результаты, чем его аналоги[47].

В своем исследовании Тиму Мутанен использовал и проанализировал базу данных клиентов из финского банка. Данные состояли только из личных клиентов. Данные были собраны с периода с декабря 2002 года по сентябрь 2005 года. Интервал выборки составлял три месяца, поэтому для этого исследования у него были соответствующие данные 12 точек времени [t (0) -t (11)]. В логистическом регрессионном анализе он использовали выборку из 151 000 клиентов[61].

Всего из базы данных клиентов было собрано 75 переменных. Эти переменные связаны с темами следующим образом: (1) транзакции счета IN, (2) транзакции счета OUT, (3) показатели обслуживания, (4) информация личного профиля и (5) объединенная информация уровня клиента.

В интересах компании, проводить мероприятия по удержанию клиентов. Соответственно перед отделом маркетинга стоит вопрос: что можно сделать, чтобы сохранить их. Достаточно ли трехмесячного периода прогнозирования, чтобы оказать положительное влияние, чтобы клиент остался? Или прогноз должен быть сделан, например, на шесть месяцев вперед?

Анализ оттока клиентов в этом исследовании может быть неинтересным, если клиенты оцениваются на основе стоимости жизни клиента. Определение оттока в его исследовании было основано на текущей информации. Но если определение оттока было основано, например, на учетной записи программы лояльности или на активном использовании интернет-сервиса, тогда клиенты, находящиеся в центре внимания, могли бы иметь большую ценность в течение всего «срока жизни», и, следовательно, было бы более важно сохранить этих клиентов.

В таблице 1 представлены примеры исследований по прогнозированию оттока, найденные в литературе: анализ клиентов, перестающих пользоваться услугами компаний, проводился в различных областях.

Таблица 1 – Анализ исследований, посвященных оттоку клиентов в различных секторах рынка

Автор	Секторы рынка и использованные методы	
Ау и соавторы[17]	Беспроводные технологии	DMEL-метод
	Телекоммуникации	Эволюционные алгоритмы
Бакинкс и соавторы[18]	Продажи	Логистическая регрессия
	Бизнес	Деревья решений
Феррера и соавторы[25]	Беспроводные технологии	Нейронная сеть, деревья решений
	Телекоммуникации	Метод эволюции правил
Гарланд и соавторы[27]	Продажи	Множественная регрессия
	Банковские услуги	Множественная регрессия
Хванг и соавторы[34]	Беспроводные технологии	Логистическая регрессия, нейронная сеть
	Телекоммуникации	Деревья решений
Мозер и соавторы[53]	Беспроводные технологии	Логистическая регрессия, нейронная сеть
	Телекоммуникации	Деревья решений
Кевени и соавторы[39]	Онлайн	Описательная статистика, основанная на опросах, проводимых среди клиентов
	Услуги	
Го-ан и Вэй-дон[29]	Телекоммуникации	Метод опорных векторов, деревья решений, логистическая регрессия, наивный Байесовский классификатор
Анил Кумар и Рави[16]	Банковские услуги	Многослойный перцептрон Румельхарта, логистическая регрессия, деревья решений, случайный лес, сеть радиально-базисных функций, метод опорных векторов
Ние и соавторы[54]	Финансы	Логистическая регрессия, деревья решений

Окончание таблицы 1 – Анализ исследований, посвященных оттоку клиентов в различных секторах рынка

Лин и соавторы[45]	Банковские услуги	Грубый набор, техника принятия решений, основанная на правилах, транспортная сеть
Хуанг и соавторы[32]	Телекоммуникации	Логистическая регрессия, линейный классификатор, наивный Байесовский классификатор, деревья решений, многослойный перцептрон Румельхарта, метод опорных векторов
Шарма и Панигрохи[58]	Телекоммуникации	Нейронные сети
Юи и соавторы[67]	Электронная коммерция	Нейронные сети, метод опорных векторов, деревья решений, метод расширенных опорных векторов
Фаркуад и соавторы[24]	Банковские услуги	Метод опорных векторов, наивный Байесовский классификатор, правила деревьев
Керамати и соавторы[40,41]	Телекоммуникации	Деревья решений, нейронные сети, метод опорных векторов, метод k-ближайших соседей
Грегори[28]	Банковские услуги	XGBoost
Мачадо и соавторы[47]	Банковские услуги	LightGBM
Мутанен[61]	Банковские услуги	Логистическая регрессия

Тематику банковских услуг в своих работах затронули следующие авторы:

Гарланд[27], Анил Кумар[16], Лин[45], Фаркуад[24], Грегори[28], Мачадо[47] и Мутанен[61]. Отдельно они представлены в таблице 2.

Таблица 2 – Анализ исследований, посвященных оттоку клиентов из банка

Автор	Методология
Гарланд и соавторы[27]	Множественная регрессия
Анил Кумар и Рави[16]	Многослойный перцептрон Румельхарта, логистическая регрессия, деревья решений, случайный лес, сеть радиально-базисных функций, метод опорных векторов
Лин и соавторы[45]	Грубый набор, техника принятия решений, основанная на правилах, транспортная сеть
Фаркуад и соавторы[24]	Метод опорных векторов, наивный Байесовский классификатор, правила деревьев
Грегори[28]	XGBoost
Мачадо и соавторы[47]	LightGBM
Мутанен[61]	Логистическая регрессия

Наиболее эффективные решения предложили Браян Грегори (Экстримальный градиентный бустинг – XGBoost), Маркос Роберто Мачадо (LightGBM), и Тиму Мутанен, который использовал метод логистической регрессии в своей работе.

Сектор потребительских розничных банковских услуг формируется клиентами, которые остаются в компании очень долгое время. Клиенты обычно отдают свой финансовый бизнес одной компании, и они не будут часто менять поставщика финансовой помощи. С точки зрения компании, это создает стабильную среду для управления взаимоотношениями с клиентами.

При постоянных отношениях с клиентами потенциальная потеря дохода из-за оттока клиентов в этом случае может быть огромной. Массовый маркетинг не может преуспеть в многообразии потребительского бизнеса сегодня. В связи с этим, анализ потребительской ценности наряду с прогнозами оттока клиентов поможет маркетинговым программам ориентироваться на более конкретные группы клиентов.

1.4 Постановка задачи

В современных реалиях ведение успешной банковской деятельности напрямую зависит от прогнозирования оттока клиентов. Исходя из этого целью данной работы будет являться повышение эффективности прогнозирования совершения ключевого действия клиентов банка.

Объектом исследовательской работы является процесс обеспечения и улучшение качества обслуживания клиентов.

Предметом исследования являются методы прогнозирования совершения ключевого действия клиентом банка с использованием методов машинного обучения.

Целью исследования является разработка математических моделей прогнозирования потенциальной ценности услуги для клиентов.

На основе результатов прогнозирования, банк в дальнейшем может сделать вывод о том какая услуга будет важна для клиента и своевременно её предоставить, тем самым предотвратив его уход из банка.

Для достижения цели были поставлены следующие задачи:

1) описание процесса маркетинга отношений;

2) теоретическое описание машинного обучения, его виды и алгоритмы;

3) анализ и сравнение имеющихся прогнозных методов машинного обучения;

4) формулировка требований к прогнозной модели;

5) построение прогнозной модели;

6) сравнительный анализ результатов прогнозирования;

7) формирование плана коммерциализации.

Исходя из анализа работ, посвященных прогнозированию совершения ключевого действия клиентов банка, было принято следующее решение: обратить внимание на вариации градиентного бустинга, т.к. они хорошо зарекомендовали себя для решения данной задачи.

Для достижения поставленных целей и задач построим модели следующими методами: воспользуемся вариациями градиентного бустинга: XGBoost (экстремальный градиентный бустинг) LightGBM, а также CATBoost (Категориальный бустинг). После формирования моделей, и сравнения их результатов, сформируем усредненные модели, которые в теории должны дать еще большую точность.

Для средних наборов данных (от 10 000 до 100 000 строк) метод экстремального бустинга подходит лучше остальных, т.к. показывает высокую точность. Этот алгоритм основан по схеме поэтапного подхода. В данной работе именно на метод XGBoost будет сделан основной упор. Мы не стали делать упор LightGBM, потому что данный метод в первую очередь предназначен для анализа крайне больших данных (свыше 100 000 строк). Категориальный бустинг также будет находиться в нашем исследовании на задних позициях, в связи с тем, что в первую очередь данный метод предназначен для анализа категориальных данных, у нас же все данные числовые.

XGBoost, по его названию, является улучшенным алгоритмом дерева решений. Это расширение подхода под названием Gradient Boosting, которое само по себе является расширением алгоритма AdaBoost.

К типичным особенностям решаемой задачи также можно отнести возможное наличие пропущенных данных.

Выводы по главе

Проведя анализ процесса «маркетинг отношений» и персонализации услуг, а также роли прогнозирования в данном процессе можно сделать выводы по данной главе. На сегодняшний день существует острая проблема в банковском секторе, данная проблема заключается в медленном реагировании банков на спрос клиентов, а также в недостаточной персонализации своих услуг. В связи с этим, не получая конкретную услугу, клиент отказывается от дальнейшего взаимодействия с таким банком.

Прогнозирование потенциальной стоимости услуги для клиентов банка поможет справится с проблемой клиентооттока, за счет своевременного реагирования банков на спрос их клиентов.

ГЛАВА 2 МЕТОДЫ ГРАДИЕНТНОГО БУСТИНГА, КАК ИНСТРУМЕНТ ПРОГНОЗИРОВАНИЯ СОВЕРШЕНИЯ КЛЮЧЕВОГО ДЕЙСТВИЯ КЛИЕНТОМ БАНКА

2.1 Понятие градиентного бустинга

Градиентный бустинг – это метод машинного обучения для задач регрессии и классификации, который создает модель прогнозирования в виде множества моделей слабого прогнозирования, обычно деревьев решений. Он строит модель поэтапно, как и другие методы повышения, и обобщает их, позволяя оптимизировать произвольную дифференцируемую функцию потерь[62].

Идея градиентного бустинга возникла из наблюдения Лео Бреймана, что бустинг можно интерпретировать как алгоритм оптимизации подходящей функции стоимости[42]. Алгоритмы явного повышения градиента регрессии были впоследствии разработаны Джеромом Фридманом одновременно с более общей перспективой повышения функционального градиента Лью Мейсоном, Джонатаном Бакстером, Питером Бартлеттом и Маркусом Фрианом [48,49]. В последних двух статьях были представлены алгоритмы повышения в качестве итерационных алгоритмов функционального градиентного спуска.

То есть алгоритмы, которые оптимизируют функцию стоимости по функциональному пространству путем итеративного выбора функции, которая указывает в направлении отрицательного градиента. Такое функциональное градиентное представление о повышении привело к разработке алгоритмов повышения во многих областях машинного обучения и статистики, помимо регрессии и классификации.

Градиентный бустинг чувствительный к настройкам, но при правильно подобранных параметрах может дать существенный прирост качества модели.

В градиентном бустинге важен такой параметр как «learning_rate», с помощью которого появляется возможность контролировать скорость обучения.

Под скоростью обучения понимается то, на сколько сильно дерево будет исправлять ошибки предыдущего дерева.

Основным недостатком данного алгоритма является чувствительность к параметрам модели, а также то, что для обучения может понадобиться время. Так же, как и другие алгоритмы, которые базируются на дереве решений, алгоритм отлично работает на данных сочетающие в себе непрерывные и бинарные признаки.

Так же стоит выделить основные параметры градиентного бустинга, это learning_rate и n_estimators. Эти два параметра тесно связанны между собой, так как при низком значение learning_rate требуется большое количество деревьев, в отличие от вышеописанных методов большое количество деревьев в градиентном бустинге делает модель более сложной, что может привести к переобучению.

Существует общепринятая рекомендация, которая заключается в том, чтобы настраивать n_estimators в зависимости от возможности вычислительной машины, а затем подгонять learning_rate.

2.1.1 XGBoost

Впервые XGBoost был выпущен в 2014 году студеном PhD Тианки Чен. Чен, предыдущий победитель конкурса данных KDDCup, обнаружил, что его предпочтительный подход («расширенные деревья решений») плохо поддерживается существующим программным обеспечением.

Его альтернатива, выпущенная как отдельная программа командной строки, приобрела известность позже в том же году, когда она поднялась на вершину списка лидеров соревнований Kaggle. Конкурс машинного обучения Бозона-Хиггса попросил участников исследовать свойства этой частицы после ее открытия в 2012 году, в частности, сосредоточившись на идентификации событий распада Хиггса в смоделированных данных. XGBoost быстро выпустил модель с самой высокой предикативной оценкой[62].

После этого первоначального успеха алгоритм быстро приобрел огромную популярность в кругах машинного обучения: к 2015 году он использовался в более чем половине победных конкурсов Kaggle. С тех пор он продемонстрировал современную производительность в задачах, начиная от прогнозирования продаж в магазине до обнаружения движения и классификации вредоносных программ.

XGBoost является частью семейства алгоритмов машинного обучения, основанных на концепции «дерева решений». Дерево решений – это простая система, основанная на правилах, построенная вокруг иерархии разветвленных утверждений истина / ложь.

XGBoost, по его названию, является улучшенным алгоритмом дерева решений. Это расширение подхода под названием Gradient Boosting, которое само по себе является расширением алгоритма AdaBoost.

Стандартный алгоритм градиентного бустинга включал ряд специальных параметров, касающихся роста деревьев решений. Затем XGBoost стандартизировал эти параметры таким образом, чтобы лучше обобщать результаты [56].

XGBoost включает в себя ряд других изменений, предназначенных для ускорения вычислений или улучшения соответствия. Например, общая проблема с алгоритмами градиентного бустинга (в отличие от алгоритмов бэггинга) заключается в том, что их нельзя распараллелить: вы не можете разделить рабочую нагрузку пополам и передать половинки разным процессорам для одновременной работы. Это связано с тем, что каждое новое дерево основывается на результатах предыдущих деревьев, пытаясь устранить их остаточную ошибку.

XGBoost обходит эту проблему. Вместо того, чтобы пытаться распараллелить обучение целых деревьев, он распараллеливает обучение различных узлов в каждом дереве[15].

Другие ключевые оптимизации включают в себя:

- ускоренный поиск идеальной точки разбиения каждого узла дерева (с использованием подхода «взвешенного квантиля»);

- лучшее определение того, когда прекратить выращивать дерево (с помощью метода «обрезки»);
- эффективная обработка разреженных данных, таких как ключевые переменные.

Алгоритм XGBoost имеет много сильных преимуществ:

- он очень быстро генерирует прогнозы и сравнительно быстро обучается по сравнению с большинством алгоритмов обучения;
- в частности, его можно - в определенной степени - распараллелить, что позволяет обучать большие модели на нескольких процессорах;
- он может обрабатывать разреженные данные, с которыми борются многие алгоритмы;
- реализация доступна с открытым исходным кодом и хорошо поддерживается как R, так и Python.

Тем не менее, у него есть несколько проблем:

- это черный ящик: определить логику любого данного прогноза относительно сложно;
- он выполняет очень поверхностный анализ, не пытаясь понять природу базовой системы;
- в частности, предсказания являются прерывистыми - они переходят от одного значения к другому в точках разделения, даже когда объясняющая и целевая переменные непрерывны.

Гиперпараметры[66]

Структурные параметры XGBoost – те параметры, которые задают контекст, который обуславливает тренировку деревьев - следующие:

- количество раундов (Number of rounds). Количество деревьев решений, которые накладываются друг на друга, каждый из которых повышает производительность последнего По-умолчанию: нет;
- ранние остановки (Early stopping rounds). Количество раундов, после которого прекращается повышение, если улучшения не видно. Полезно с

большим количеством раундов. По умолчанию: ноль (функция не включена);

- скорость обучения (Learning rate (eta)). Масштабирующий множитель применяется к каждому дереву, чтобы уменьшить переоснащение и дать будущим деревьям больше возможностей для роста По умолчанию: 0,3.

Параметры, специфичные для подгонки отдельных деревьев решений, следующие:

- максимальная глубина или максимальное количество листовых узлов (Maximum depth or maximum number of leaf nodes). Управляет размером каждого дерева. По умолчанию: 6 слоев или 64 (= 26) узлов;
- minimum child weight: это минимальный вес ребенка. Наименьшая популяция, которую может иметь листовой узел. По умолчанию: 1 (функция не включена);
- minimum loss reduction (gamma): минимальное снижение потерь (гамма). Наименьшее влияние, которое раскол может оказать на добротность дерева. По умолчанию: 0 (функция не включена);
- subsample: это подвыборки. Произвольно выбранная пропорция набора данных (строк / наблюдений), используемая при построении каждого дерева, для уменьшения избыточного соответствия. По умолчанию: 1 (функция не включена);
- column sample by tree: это выборка столбца по дереву. Произвольно выбранная пропорция набора данных (столбцы / объекты), используемая при построении каждого дерева, для уменьшения избыточного соответствия. По умолчанию: 1 (функция не включена).

Два параметра регуляризации (представляющие регуляризацию «L2» и «L1» соответственно) следующие:

- lambda: лямбда. Наказывает деревья с большими значениями листьев, чтобы уменьшить вероятность переобучения модели. По умолчанию: 0 (функция не включена).

- alpha: альфа. Наказывает деревья с большим количеством листьев, чтобы уменьшить вероятность переобучения модели. По умолчанию: 0 (функция не включена).

2.1.2 LightGBM

LightGBM – это платформа для повышения градиента, использующая алгоритмы обучения на основе дерева.

Данный метод обладает следующими преимуществами[47]:

- Более быстрая скорость обучения и высокая эффективность;
- Более низкое использование памяти;
- Лучшая точность;
- Поддержка параллельного и обучения с использованием графического процессора;
- Способен обрабатывать крупные данные.

Недостатком является высокая чувствительность к переобучению модели.

Гиперпараметры[44]:

- max_depth: описывает максимальную глубину дерева. Этот параметр используется для обработки перенастройки модели. Каждый раз, когда вы чувствуете, что ваша модель перегружена, мой первый совет - понизить max_depth;
- min_data_in_leaf: это минимальное количество записей, которое может иметь лист. Значение по умолчанию - 20, оптимальное значение. Данный гиперпараметр используется, когда появляются проблемы переобучением модели;
- feature_fraction: используется, когда бустинг – это случайный лес. Фракция 0,8 означает, что LightGBM будет выбирать 80% параметров случайным образом в каждой итерации для построения деревьев;

- bagging_fraction: указывает часть данных, которые будут использоваться для каждой итерации, и обычно используется для ускорения обучения и предотвращения переобучения;
- early_stopping_round: этот параметр может помочь ускорить анализ. Модель прекратит обучение, если одна метрика из одной проверочной информации не улучшится в последних раундах early_stopping_round. Это уменьшит чрезмерные итерации;
- Lambda: лямбда указывает на регуляризацию. Типичное значение варьируется от 0 до 1;
- min_gain_to_split: этот параметр будет описывать минимальное усиление для сплитов. Может использоваться для контроля количества полезных сплитов в дереве;
- max_cat_group: когда число категории велико, нахождение точки разделения на ней легко переобучается. Поэтому LightGBM объединяет их в группы «max_cat_group» и находит точки разделения на границах группы, по умолчанию: 64;
- Boosting: определяет тип алгоритма, который запустить, по умолчанию = gdbt:
- gbdt: традиционное дерево решений с градиентным усилением;
- RF: случайный лес;
- goss: выборочная градиентная односторонняя выборка;
- num_boost_round: количество ускоряющих итераций, обычно более 100;
- learning_rate: определяет влияние каждого дерева на конечный результат. GBM работает, начиная с начальной оценки, которая обновляется с использованием выходных данных каждого дерева. Параметр обучения контролирует величину этого изменения в оценках;
- num_leaves: количество листьев в полном дереве, по умолчанию: 31;
- Device: по умолчанию: процессор, также может использоваться графический процессор.

2.1.3 CATBoost

CatBoost — открытая программная библиотека, разработанная компанией Яндекс и реализующая уникальный патентованный алгоритм построения моделей машинного обучения, использующий одну из оригинальных схем градиентного бустинга. Основное API для работы с библиотекой реализовано для языка Python, также существует реализация для языка программирования R[46].

Практически любой современный метод на основе градиентного бустинга работает с числовыми признаками. Если у нас в наборе данных присутствуют не только числовые, но и категориальные признаки (англ. *categorical features*), то необходимо переводить категориальные признаки в числовые. Это приводит к искажению их сути и потенциальному снижению точности работы модели. Именно поэтому было важно разработать алгоритм, который умеет работать не только с числовыми признаками, но и с категориальными напрямую, закономерности между которыми этот алгоритм будет выявлять самостоятельно, без ручной «помощи». CatBoost — библиотека для градиентного бустинга, главным преимуществом которой является то, что она одинаково хорошо работает «из коробки» как с числовыми признаками, так и с категориальными[41].

Идея бустинг-подхода заключается в комбинации слабых (с невысокой обобщающей способностью) функций, которые строятся в ходе итеративного процесса, где на каждом шаге новая модель обучается с использованием данных об ошибках предыдущих. Результирующая функция представляет собой линейную комбинацию базовых, слабых моделей. Более подробно можно посмотреть в статье про градиентный бустинг.

Далее будет рассматриваться бустинг деревьев решений. Будем строить несколько деревьев, чтобы добавление новых деревьев уменьшало ошибку. Итого при достаточно большом количестве деревьев мы сможем сильно уменьшить

ошибку, однако не стоит забывать, что чем больше деревьев, тем дольше обучается модель и в какой-то момент прирост качества становится незначительным.

Главным плюсом CATBoost является тот факт что данный алгоритм может работать с категориальными данными.

К минусам можно отнести относительно медленную работу алгоритма, и невысокую точность при работе с некатегориальными данными.

Гиперпараметры[19]:

- depth: глубина дерева. Диапазон поддерживаемых значений зависит от типа единицы обработки и типа выбранной функции потерь:
- CPU - любое целое число до 16;
- GPU - любое целое число до 8 парных режимов (YetiRank, PairLogitPairwise и QueryCrossEntropy) и до 16 для всех других функций потерь.
- iterations: количество итераций. Максимальное количество деревьев, которое можно построить при решении задач машинного обучения;
- learning_rate: частота обучения. Используется для уменьшения шага обучения;
- l2_leaf_reg: коэффициент регуляризации L2. Коэффициент регуляризации l2 функции стоимости. Допускается любое положительное значение;
- bagging_temperature: температура бэггинга. Определяет настройки байесовского бутстрэпа. Он используется по умолчанию в режимах классификации и регрессии. Байесовский бутстрэп используется, чтобы назначать случайные веса объектам. Веса выбираются из экспоненциального распределения, если значение этого параметра установлено «1». Все веса равны 1, если значение этого параметра установлено в «0». Возможные значения находятся в диапазоне. Чем выше значение, тем агрессивнее бэггинг. Этот параметр можно использовать, если выбранный тип начальной загрузки – байесовский.

Выводы по главе

Мы рассмотрели сам градиентный бустинг и три его вариации, а конкретно:

- XGboost;
- LightGBM;
- CATBoost.

Для решения нашей задачи повышение эффективности прогнозирования совершения ключевого действия клиентами банка будем строить модели на следующих алгоритмах: XGBoost (экстремальный градиентный бустинг), LightGBM и CATBoost (категориальный бустинг).

В целом можно сказать, что для решения задачи классификации подходят множество алгоритмов, каждый алгоритм имеет свои достоинства и недостатки, связанные с их спецификой.

Таблица 3 – Сравнение алгоритмов

Алгоритм	Точность	Время обучения	Тип данных	Решаемая задача
XGBoost	Высокая	Зависит от параметров модели	Смесь бинарных и непрерывных признаков; Не требует масштабируемости данных.	Классификация; Регрессия.
LightGBM	Высокая	Зависит от параметров модели (но в основном быстрее чем другие вариации GBM)	Смесь бинарных и непрерывных признаков; Не требует масштабируемости данных.	Классификация; Регрессия.
CATBoost	Средняя (высокая при использовании категориальных данных)	Зависит от параметров модели (Самое медленное среди GBM)	Смесь бинарных, категориальный и непрерывных признаков; Не требует масштабируемости данных.	Классификация; Регрессия.

ГЛАВА 3 ПРОГНОЗИРОВАНИЕ СОВЕРШЕНИЯ КЛЮЧЕВОГО ДЕЙСТВИЯ КЛИЕНТОМ БАНКА

3.1 Метрика качества модели

В качестве основной метрики проекта будем использовать ошибку RMSLE. RMSLE расшифровывается как «Mean Squared Logarithmic Error» и переводится как «Средняя квадратическая логарифмическая ошибка». Суть метода заключается, по сути, в том, чтобы минимизировать сумму квадратов отклонений фактических значений от расчётных (SSE — «Sum of Squared Errors»). Формула целевой функции в этом случае выглядит следующим образом:

$$\text{RMSLE} = \sqrt{\sum_{t=1}^{T}(\log(x_i + 1) - (\log(y_i + 1))^2} \tag{3.1}$$

где y_i – истинные целевые значения;

x_i – предсказанные целевые значения.

Если бы в формуле (3.1) не было квадрата, то положительные и отрицательные отклонения друг друга погашали, из-за чего минимизировалось бы не расстояние между фактическими и расчётными значениями, а просто бы получалась разница. То есть наличие квадратов позволяет получить некоторую оценку расстояния от фактических значений до линии (расчётных значений).

3.2 Разведочный анализ данных.

Для работы с данными составим следующий план:

1. разведочный анализ данных (exploratory data analysis);
2. преобразование признаков (feature engineering);
2. построение базовых моделей;
3. улучшение модели;

4. интерпретация модели.

В качестве задачи для исследования был выбран конкурс с сайта kaggle.com

В частности, предлагается, с использованием языка программирования «Python», произвести анализ данных, предоставленных компанией «Santander Group» и построить предикативную модель персонализации обслуживания клиентов. Данные находятся в открытом доступе на сайте www.kaggle.com

Банк Сантандер, Santander Group – крупнейшая финансово-кредитная группа в Испании. Помимо Испании Santander занимает одно из ведущих мест в Великобритании и в ряде стран Латинской Америки, также представлена в США.

Цель соревнования – создать методику оценки ценности услуги для каждого потенциального клиента.

Нам предоставлены анонимизированный набор данных, содержащий числовые переменные, числовые целевые столбцы и строчные значения столбцов (их ID).

Задача состоит в том, чтобы предсказать значение целевого столбца в тестовой выборке.

В качестве набора данных компания предоставляет 2 файла в формате csv:

Train.csv – Обучающий набор;

Test.csv – Тестовый набор.

Основываясь на результатах предыдущей главы, мы сделали вывод, что для решения данной задачи, будет более целесообразно построить модели на следующих алгоритмах: XGBoost (экстремальный градиентный бустинг), LightGBM и CATBoost (категориальный бустинг).

Для обучения модели разобьем обучающую выборку на обучающую и тестовую с соотношением 80 на 20 соответственно.

Обучающей выборкой называют такую выборку, по которой происходит настройка модели. Проверочную выборку будем использовать для контроля

переобучения модели. Тестовая выборка (контрольная) служит для оценки итоговой модели, данная выборка дает честную оценку предсказания.

Таблица 4 – Разбиение исходных данных на выборки

Наименование выборки	Размер выборки, наблюдений	Доля, %
Обучающая:	4459	100
обучающая	3567	80
проверочная	892	20
Тестовая	49342	100

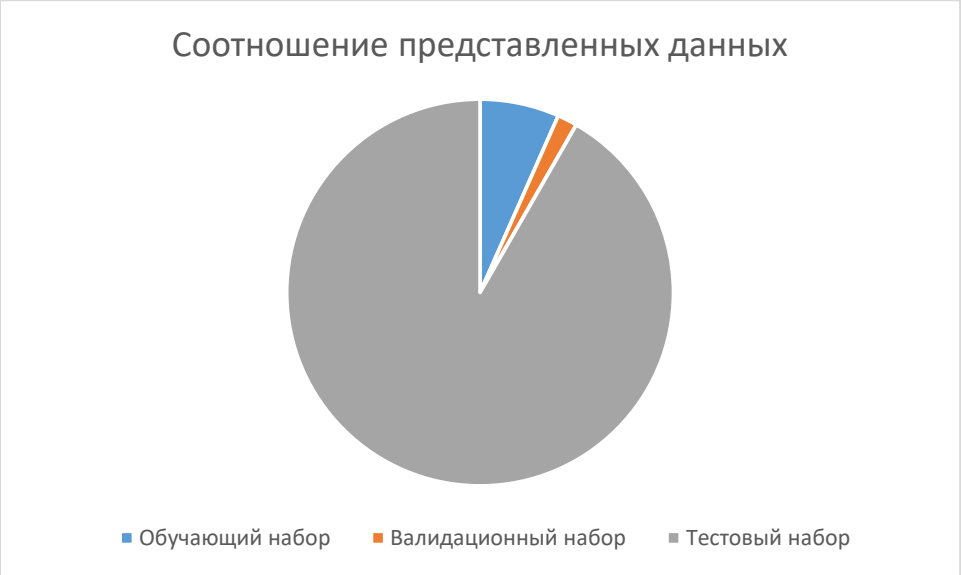

Рисунок 3 – Соотношение представленных данных

После первичного осмотра данных можно сделать следующие выводы:

1. У нас есть 41 признак и 49342 строки в тестовом наборе. У нас также есть 42 столбца, включая столбцы target и id, и 4459 строк в обучающей выборке. Расхождение, очевидно, вызвано тем, что целевого признака target в тестовой выборке нет, его нам и нужно предсказывать.

2. С первого взгляда можно сразу заметить, что тестовый набор почти в 11 раз больше, чем обучающий и валидационный вместе взятые.

Рассмотрим данные более внимательно. В таблице 5, представлены первые 7 столбцов обучающейся выборки.

Таблица 5 – Первые 7 столбцов обучающей выборки

ID	target	f190486d6	58e2e02e6	eeb9cd3aa	9fd594eec	6eef030c1
000d6aaf2	38000000.0	1866666.66	12066666.66	700000.0	600000.0	900000.0
000fbd867	600000.0	0.00	2850000.00	2225000.0	1800000.0	800000.0
0027d6b71	10000000.0	0.00	0.00	0.0	0.0	0.0
0028cbf45	2000000.0	2000000.00	0.00	0.0	0.0	0.0
002a68644	14400000.0	0.00	0.00	0.0	0.0	37662000.0

В таблице 6 показан формат данных.

Таблица 6 – Типизация данных

Column Type	Count
float64	41
object	1

Все столбцы имеют тип данных с плавающей точкой. Существует только один строковый столбец, который является ничем иным, как столбцом «ID».

Для оптимизации данных, проверим обучающуюся выборку на наличие пропущенных значений. После проверки данных, выявлено, что пропущенных значений нет.

Далее проверим выборку на наличие констант.

Выявлено что в данной выборке нет постоянных значений.

Исходя из вышеприведённых проверок, можно сделать вывод что данные оптимизированы и уже готовы к дальнейшей обработке.

После оптимизации данных перейдем к корреляционному анализу переменных.

Для начала проверим зависимость между целевой переменной и другими 40 признаками.

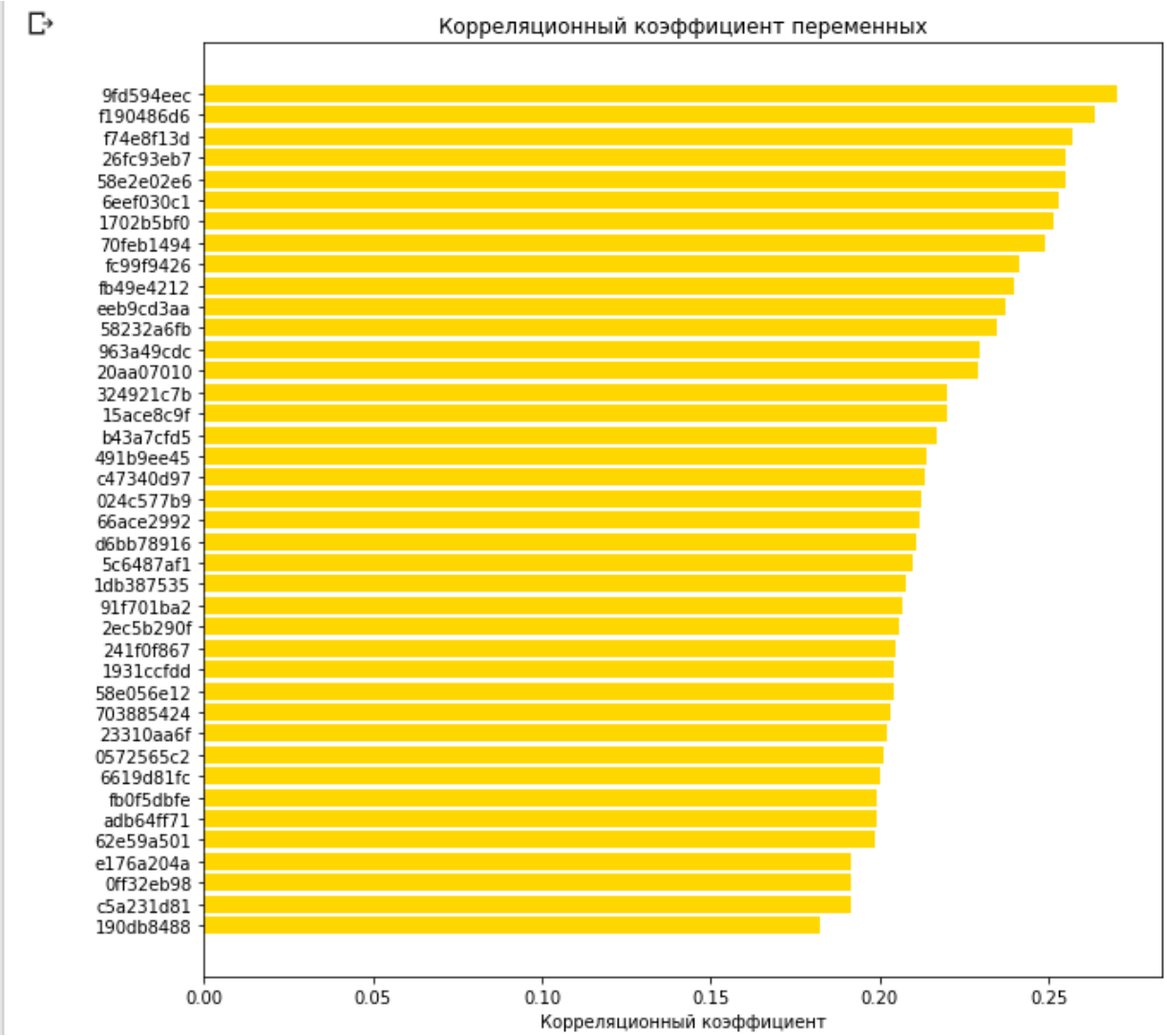

Рисунок 4 – Корреляционный коэффициент переменных

На основе данных, представленных на рисунке 3 можно сделать следующие выводы:

1) Все переменные в той или иной степени имеют низкую корреляцию с целевым признаком;

2) Только 7 переменных имею корреляцию около 0.25

Рассмотрим корреляционную зависимость между 7 переменными, имеющими корреляцию с целевым признаком около 0.25, для этого воспользуемся корреляцией Спирмена:

Корреляция Спирмена между двумя переменными равна корреляции Пирсона между значениями рангов этих двух переменных; в то время как корреляция Пирсона оценивает линейные отношения, корреляция Спирмена оценивает монотонные отношения. Если нет повторяющихся значений данных, идеальная корреляция Спирмена +1 или -1 происходит, когда каждая из переменных является идеальной монотонной функцией другой.

Интуитивно понятно, что корреляция Спирмена между двумя переменными будет высокой, если наблюдения имеют одинаковый ранг. Низкую корреляцию можно наблюдать, когда наблюдения имеют разный ранг между двумя переменными.

Для выборки размером n, n-оценки колонок X_i, Y_i, конвертируются в ранги $rg\ X_i, rg\ Y_i$, где корреляция Спирмена равна:

$$r_s = p_{rgx,rgy} = \frac{cov(rg_x, rg_y)}{\sigma_{rgx}\sigma_{rgy}}, \qquad (3.1)$$

где

p = Коэффициент Пирсона, примененный к ранговым переменным;

$cov(rg_x, rg_y)$= ковариация ранговых переменных;

$\sigma_{rgx}\sigma_{rgy}$ = стандартная девиация ранговых переменных.

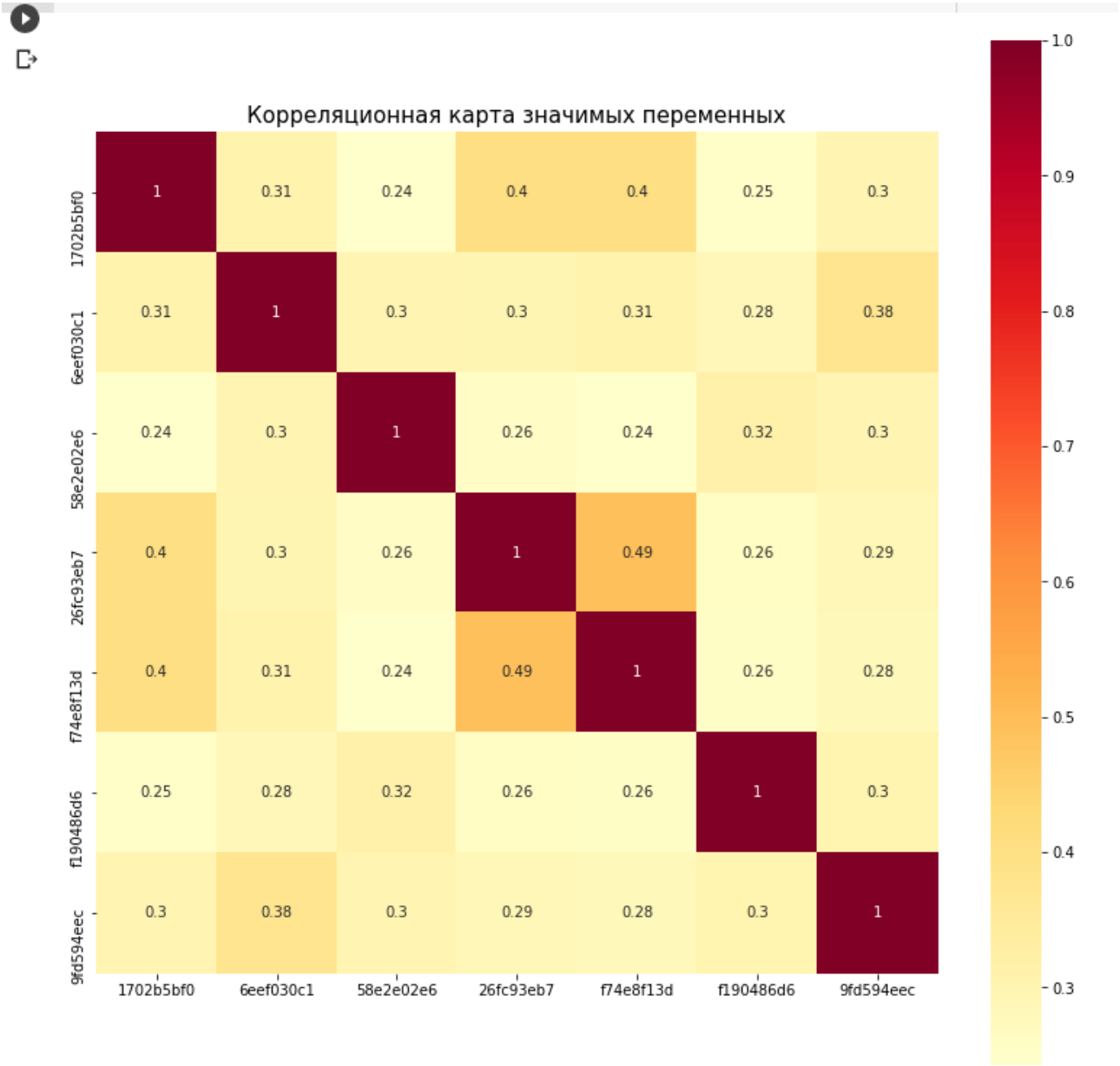

Рисунок 5 – Корреляционная карта значимых переменных

Похоже, что ни одна из выбранных переменных не имеет корреляции Спирмена более 0,49 друг с другом.

Приведенные выше графики помогли определить важные индивидуальные переменные, которые связаны с целевым признаком.

Представим распределение признаков с высоким корреляционным коэффициентом в обучающейся и тестовой выборках:

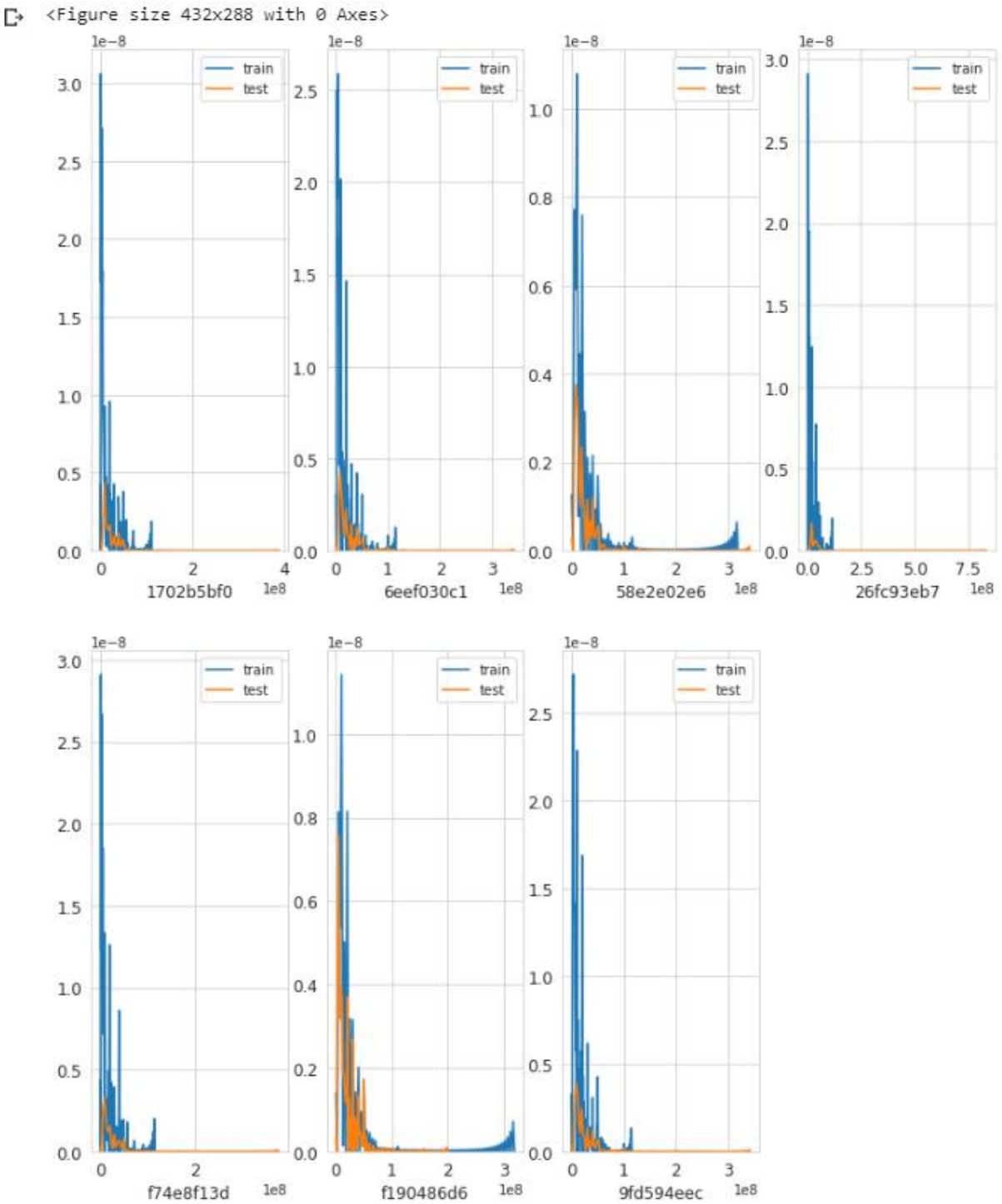

Рисунок 6 – Распределение признаков с высоким корреляционным коэффициентом

Как видно из рисунка 6, из 7 ключевых переменных, переменные 6eefO30c1, 58e2e02e6, f190486d6 имеют наибольший корреляционный коэффициент.

3.3 Преобразование признаков

Для дальнейшей оптимизации переменных для построения модели проведем следующие итерации.

Для начала выберем признаки по важности. Задействуем RandomForestRegressor, данный регрессор используется для определения важности функции, здесь мы выбираем признаки (NUM_OF_FEATURES).

Данный гиперпараметр установим в значение 40.

Далее попробуем проверить данные обучающейся и тестовой выборок с помощью теста Колмогорова-Смирнова.

Одновыборочный критерий проверки нормальности Колмогорова-Смирнова основан на максимуме разности между кумулятивным распределением выборки и предполагаемым кумулятивным распределением:

$$D_n = \sup |F_n(x) - F(x)| \tag{3.2}$$

$F_n(x)$ - кумулятивное распределение выборки

$F(x)$ - ожидаемое кумулятивное распределение (с известными параметрами)

Если D статистика Колмогорова-Смирнова значима, то гипотеза о том, что соответствующее распределение нормально, должна быть отвергнута.

Выводимые значения вероятности основаны на предположении, что среднее и стандартное отклонение нормального распределения известны априори и не оцениваются из данных.

Если в обучающем наборе функция имеет другое распределение, чем в наборе тестирования, мы должны удалить эту функцию, поскольку то, что мы узнали во время обучения, не может обобщаться. Пороговое значение (THRESHOLD_P_VALUE) и Пороговая статистика (THRESHOLD_STATISTIC) являются гиперпараметрами. Установим их в значение 0.01 и 0.3 соответственно:

После проверки, добавим дополнительные статистические функции к оригинальным функциям. Также методом «слабый случайный лес» добавим низкоразмерные представления как функции. NUM_OF_COM – гиперпараметр. Установим его в значение 10, для того чтобы не страдала точность модели:

С помощью данной итерации, мы добавили 21 дополнительный признак к имеющимся 40, который позволит нам в дальнейшем увеличить точность модели.

На предварительном этапе построения моделей у нас имеется 61 признак

3.4 Кросс-валидация

Процедура кросс-валидации, которую иногда называют перекрестной проверкой, это техника валидации модели для проверки того, насколько успешно применяемый в модели статистический анализ способен работать на независимом наборе данных. Обычно кросс-валидацию используют в ситуациях, где целью является предсказание, и хотелось бы оценить, насколько предсказывающая модель способна работать на практике. Один цикл кросс-валидации включает разбиение набора данных на части, затем построение модели на одной части (называемой тренировочным набором), и валидация модели на другой части (называемой тестовым набором). Чтобы уменьшить разброс результатов, разные циклы кросс-валидации проводятся на разных разбиениях, а результаты валидации усредняются по всем циклам.

3.4.1 Кросс-валидация по K блокам (K-fold cross-validation)

Набор данных разбивается на K одинаковых по размеру блока. Из K блоков один оставляется для тестирования модели, а остающиеся K-1 блока используются как тренировочный набор. Процесс повторяется K раз, и каждый из блоков используется один раз как тестовый набор. Получаются K результатов, по одному на каждый блок, они усредняются или комбинируются каким-либо другим способом, и дают одну оценку. Преимущество такого способа перед случайным

сэмплированием (random subsampling) в том, что все наблюдения используются и для тренировки, и для тестирования модели, и каждое наблюдение используется для тестирования в точности один раз.

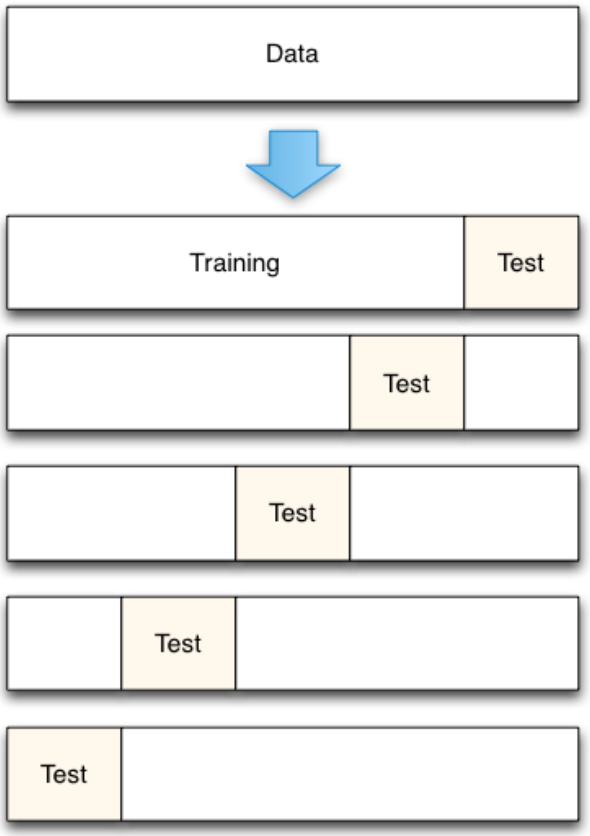

Рисунок 7 – Кросс-валидация по К блокам

В данной работе данные были разделены на 5 фолдов как было показано на рисунке 7.

3.5 Построение базовых моделей.

3.5.1 Модель XGBoost

В качестве первой базовой модели сформируем несколько моделей градиентного бустинга в вариации XGBoost.

Для обучения по данному методу, были построены 5 моделей с параметром ошибки RMSLE, с количеством бустинговых деревьев (n_estimators) от 500 до 2000, с максимальной глубиной дерева (max_depth) 50, с параметром гамма 1.1(gamma).

В ходе работе получились следующие результаты. Наименьшая ошибка при построении 750 деревьев, она же и будет являться лучшей. (рисунок 8)

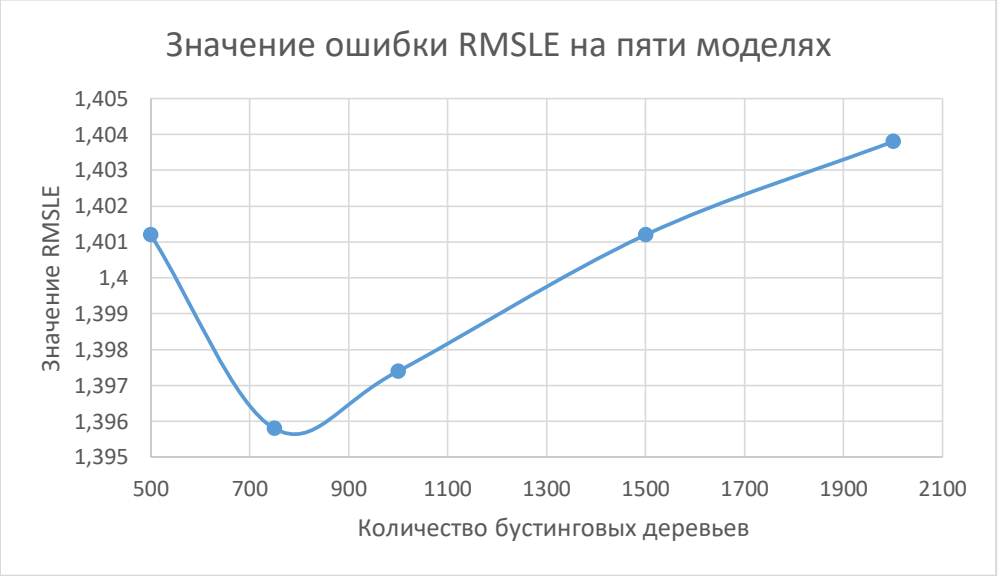

Рисунок 8 – Сравнение влияния количества бустинговых деревьев на валидационную ошибку RMSLE на валидацинной выборке

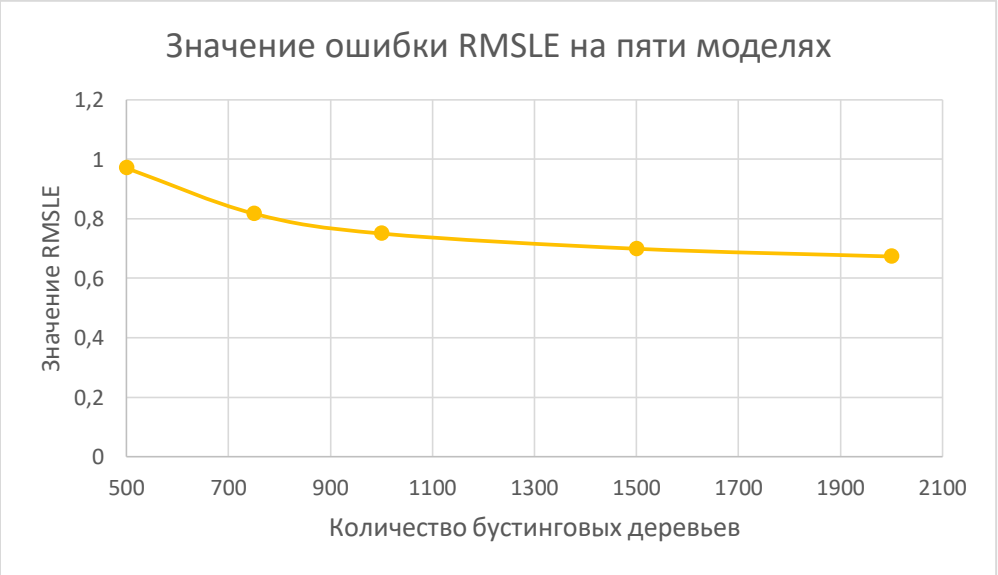

Рисунок 9 –Влияние количества бустинговых деревьев на валидационную ошибку RMSLE на обучающей выборке

Далее также на примере пяти моделей проверим значение RMSLE, при этом изменяя глубину бустинговых деревьев, с количеством бустинговых деревьев (n_estimators) 750, с максимальной глубиной дерева (max_depth) от 5 до 50, с параметром гамма 1.1(gamma). Наименьшая ошибка при глубине 45, она же и будет являться лучшей. (рисунок 10)

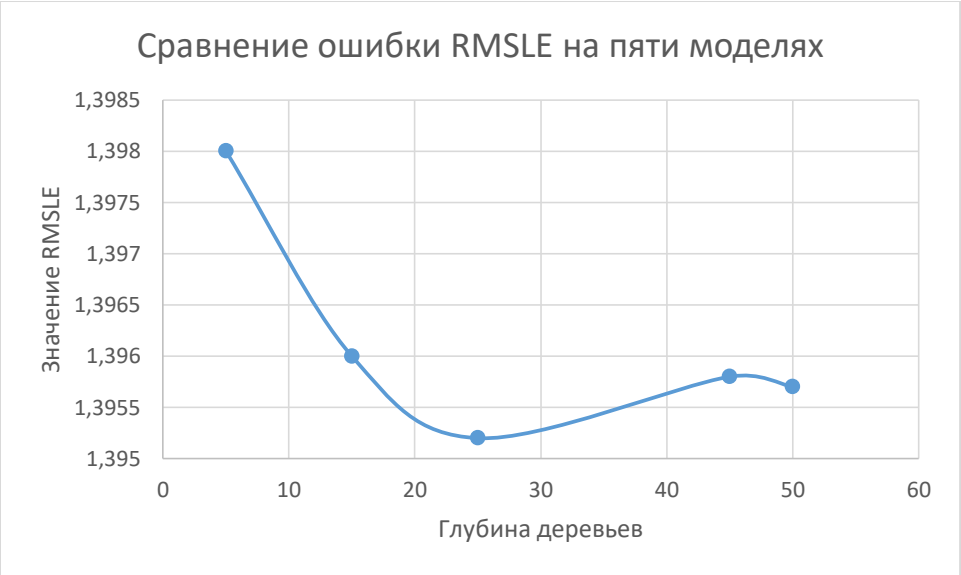

Рисунок 10 – Влияние глубины бустинговых деревьев на валидационную ошибку RMSLE на валидацинной выборке.

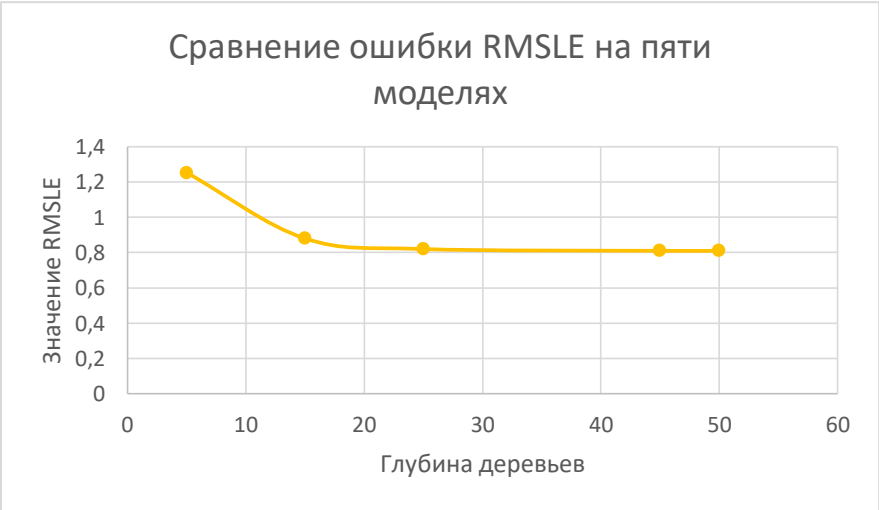

Рисунок 11 – Влияние глубины бустинговых деревьев на валидационную ошибку RMSLE на обучающей выборке.

Далее также на примере пяти моделей проверим значение RMSLE, при этом изменяя гиперпараметр гамма, с количеством бустинговых деревьев (n_estimators) 750, с максимальной глубиной дерева (max_depth) 45, с параметром гамма от 1 1.5(gamma). Наименьшая ошибка при значении 1,5, однако на обучающей выборке, увеличение данного гиперпараметра ухудшает точность, а значит сигнализирует о переобучении модели. Исходя из этого значение 1.1 будет являться лучшим. (рисунок 12, 13)

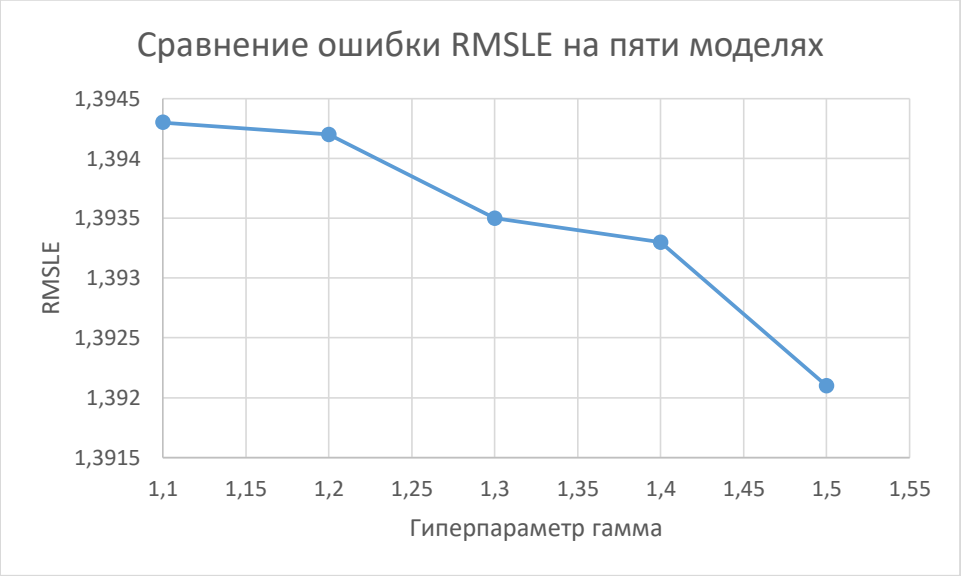

Рисунок 12 – Сравнение моделей по влиянию гиперпараметра Gamma на валидационной выборке

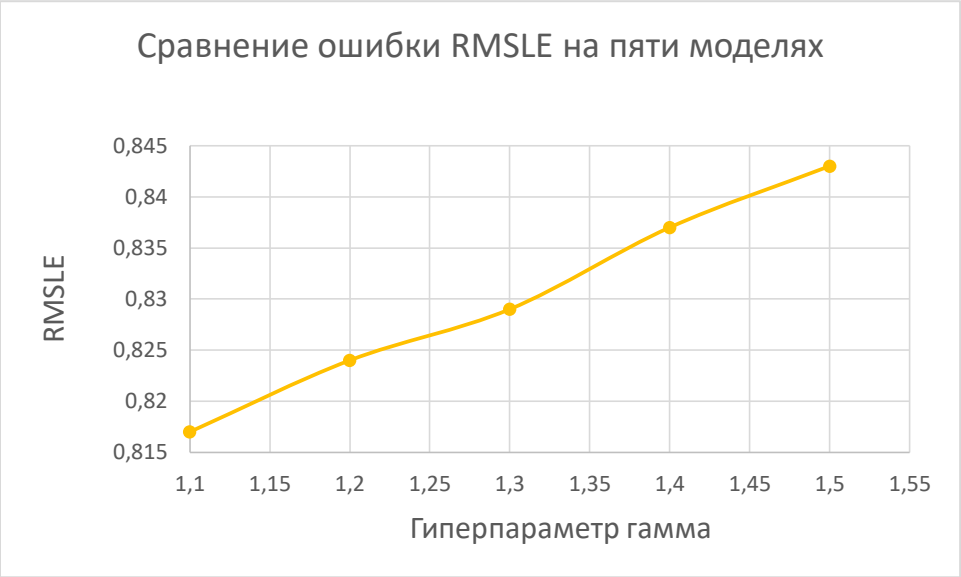

Рисунок 13 – Сравнение моделей по влиянию гиперпараметра Gamma на обучающей выборке

На основе предыдущих вычислений, составим финальный стек моделей, изменяя гиперпараметр learning_rate(шаг обучения) от 0,01 до 0,05, используя при этом предыдущие гиперпараметры. Наименьшая ошибка при значении 0,01, она же и будет являться лучшей. (рисунок 14)

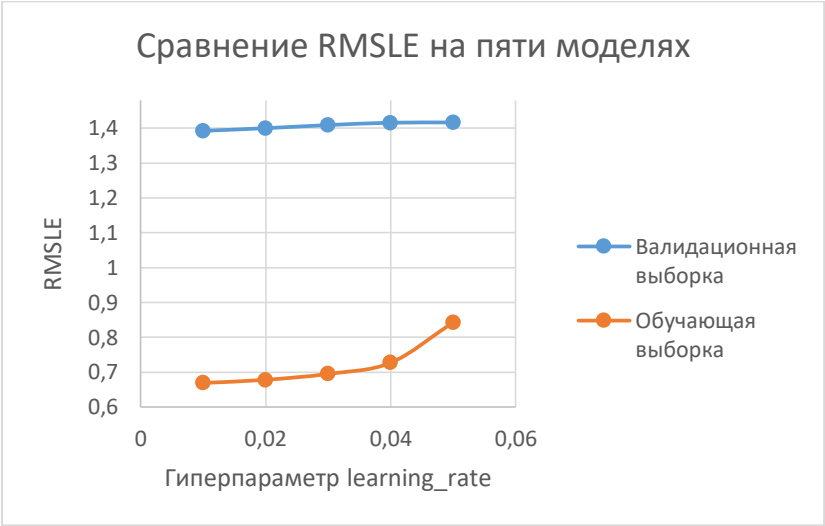

Рисунок 14 – Сравнение моделей по влиянию гиперпараметра learning_rate

Ввиду того что метрика RMSLE является достаточно редкой и специфичной, имеет смысл проверить нашу финальную базовую модель на метрике R^2(Коэффициент детерминации).

Коэффициент детерминации характеризует долю вариации (дисперсии) результативного признака y, объясняемую регрессией, в общей вариации (дисперсии) y.

Коэффициент детерминации рассчитывается для оценки качества подбора уравнения регрессии. Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 0,5.

Оценим влияние гиперпараметра learning_rate на изменение коэффициента детерминации. На финальном стеке моделей наибольшая точность составила 0,82 пункта при параметре learning_rate 0,01(рисунок 15).

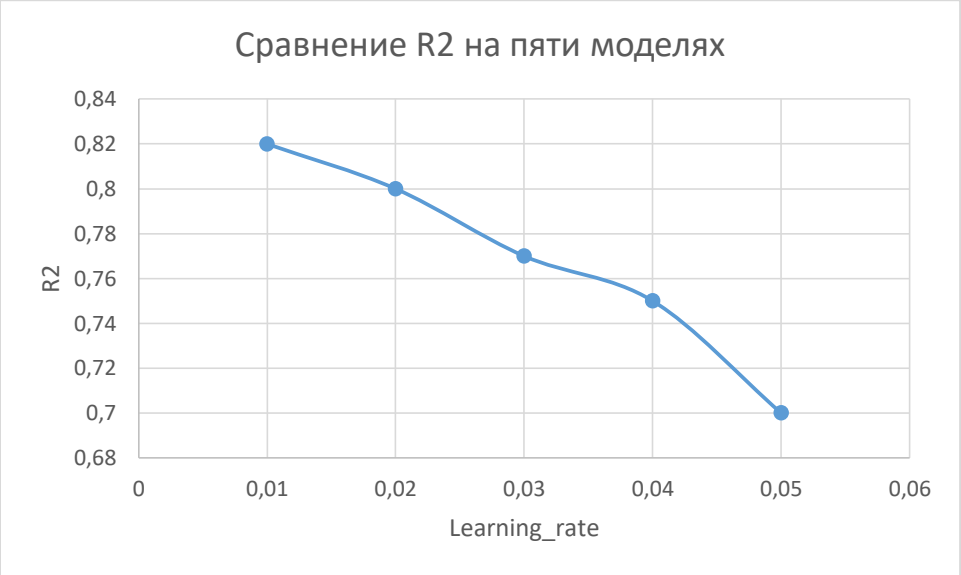

Рисунок 15 – Сравнение моделей по влиянию гиперпараметра learning_rate на коэффициент детерминации

После настройки всех гиперпараметров данная модель имеет валидационную оценку среднеквадратичной логарифмической ошибки в 1.3957 пункт и значение коэффициента детерминации в 0,82 пункта.

Посмотрим на ключевые переменные в данной модели (Рисунок 16)

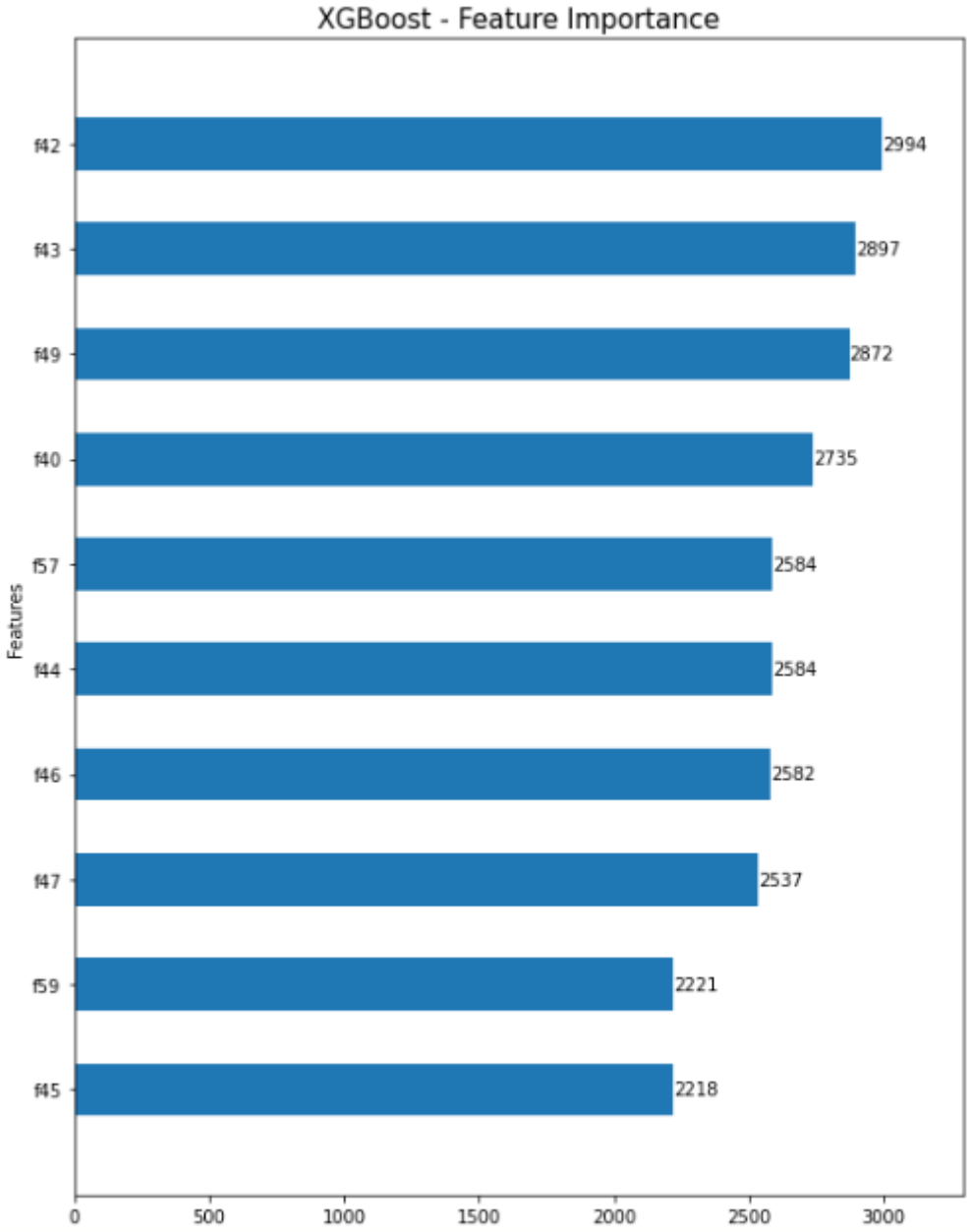

Рисунок 16 – Ключевые переменные в модели XGBoost

На рисунке 16, мы видим, что ключевым переменными для оценки модели являются переменные f42, f43, f49, f40.

Разберем поподробнее гиперпараметры данной модели:

- colsample_bylevel – данный параметр отражает отношение выборки столбцов для каждого уровня. Подвыборка выполняется один раз для каждого нового уровня глубины, достигнутого в дереве. Столбцы подвергаются выборке из набора столбцов, выбранных для текущего дерева. Нельзя не отметить, что данный параметр имеет большую зависимость от

параметра max_depth, который будет рассмотрен ниже. Наиболее оптимальном весом для представленных данных будет являться показатель в 0.5. Отклонение данного показателя хотя бы на 0.1 пункт негативно влияет на точность модели;

- num_leaves – количество листьев в дереве. Данный параметр почти никак не отражается на точности модели в данной ситуации;
- gamma – минимальное уменьшение потерь, необходимое для создания дальнейшего разделения данных на листовом узле дерева. Обычно, чем больше гамма, тем более консервативным будет алгоритм. Значения гаммы выше 1.1 вызывают переобучение модели;
- learning_rate – размер шага, используемый в обучении, предотвращает переобучение. Значение параметра установлено на пункте 0.01, потому что дальнейшее повышение негативно сказывается на точности модели;
- max_depth – максимальная глубина дерева. Увеличение этого значения сделает модель более сложной. В данном случае оптимальным показателем стало значение 50. В паре с гиперпараметром colsample_bylevel данный показатель в положительной степени повлиял на точность модели;
- objective – данный параметр отражает задачу обучения и её цель. Для градиентного бустинга с вариацией XGBoost используется значение 'reg:linear';
- booster – данный параметр отражает метод машинного обучения. В нашем случае это градиентный бустинг основанный на деревьях :'gbtree';
- min_child_weight – минимальное количество наблюдений в листе дерева. Значение параметра установлено по умолчанию на 1, т.к. в данной ситуации дальнейшее его увеличение сильно вредить точности предикативной модели;
- n_estimators – показатель отвечает за количество деревьев градиентного бустинга. Значение параметра установлено на 1000 деревьев, т.к. именно это значение позволяет добиться наивысшей точности нашей модели;

- reg_alpha – L1-регуляризация способствует разреженности функции, когда лишь немногие факторы не равны нулю. Оптимальным показателем для нашей модели будет значение 1, т.к. в данных присутствует много факторов, равных нулю;
- reg_lambda – L2-регуляризация способствует появлению малых весовых коэффициентов модели, но не способствует их точному равенству нулю. Для данного показателя также как и для показателя alpha будет адекватно значение 1, ввиду наличия в данных малых весовых коэффициентов;
- eval_metric – оценочная метрика модели, в нашем случае это среднеквадратичная логарифмическая ошибка, но т.к. в базовом библиотеке её нет, будет использовать просто средневадратичную ошибку;
- subsample – соотношение подвыборок обучающихся данных. При значении параметра в 1, модель начинается переобучаться и предикативная точность падает, в данных условиях оптимально будет использовать значение 0,9;
- silent – параметр отвечает за выключение отображения лишней информации, получаемой в процессе обучения, поставим его на 1;
- random_state – параметр, отвечающий за случайный номер зерна(сида). В процессе подбора весов, оптимальным показателем стало значение 7;
- nthread – количество потоков, использованных в процессе моделирования. Так как мы работает в Google Colab, будет логично использовать все мощности которые предоставляет нам платформа.

3.5.2 Модель LightGBM

В качестве второй базовой модели сформируем модель градиентного бустинга в вариации LightGBM.

Для обучения по данному методу, были построены 5 моделей с параметром ошибки RMSLE, с количеством бустинговых деревьев (n_estimators) от 500 до 1500, с количеством листьев 4, с параметром доля бэггинга(bagging_fraction) в 0,7.

В ходе работе получились следующие результаты. Наименьшая ошибка при построении 500 деревьев, она же и будет являться лучшей. (рисунок 17)

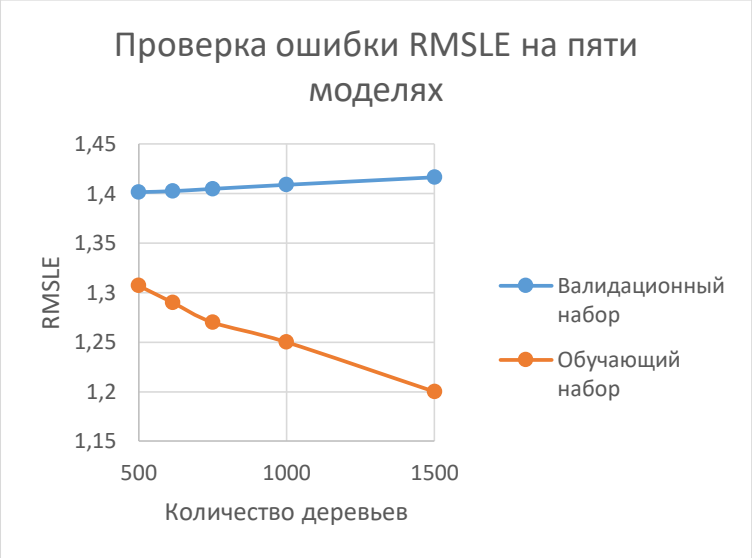

Рисунок 17 – Сравнение моделей по количеству бустинговых деревьев

Далее также на примере пяти моделей проверим значение RMSLE, с количеством бустинговых деревьев (n_estimators) 500, с количеством листьев от 100 до 200, с параметром доля бэггинга(bagging_fraction) в 0,7.

В ходе работе получились следующие результаты. Наименьшая ошибка при построении деревьев с 120 листьями, она же и будет являться лучшей. (рисунок 18)

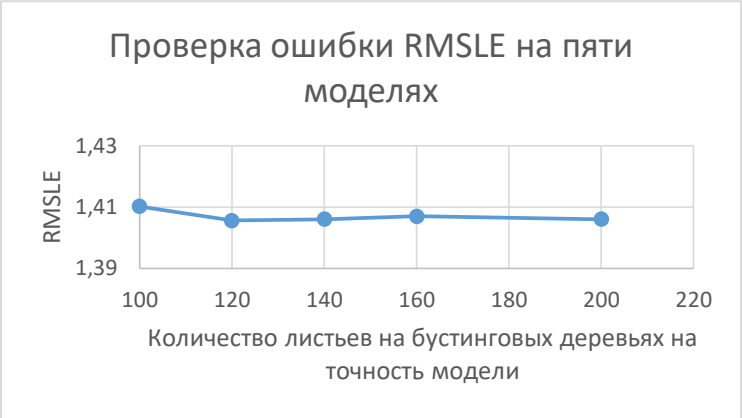

Рисунок 18 – Влияние количества листьев на бустинговых деревьях

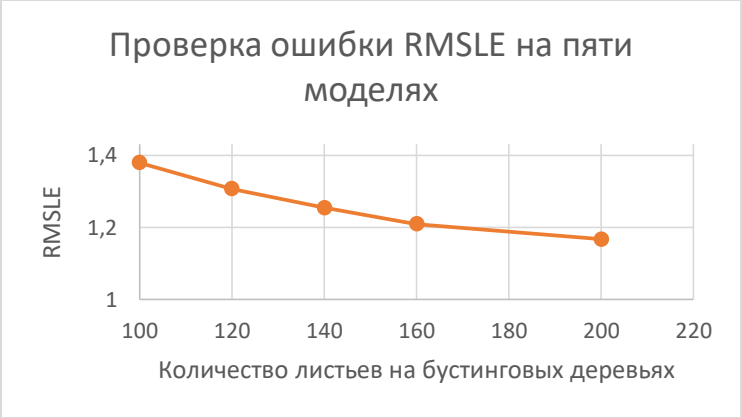

Рисунок 19 – Влияние количества листьев на бустинговых деревьях

Далее, руководствуясь предыдущими вычислениями, сравним параметр bagging_fraction в диапазоне от 0,2 до 1. Количество деревьев 500, количество листьев 4.

Наименьшая ошибка при значении 1, однако значение ошибки RMSLE на обучающей выборки начинает расти уже после 0,7, исходя из этого, именно данное значение будет являться лучшим (Рисунок 20,21)

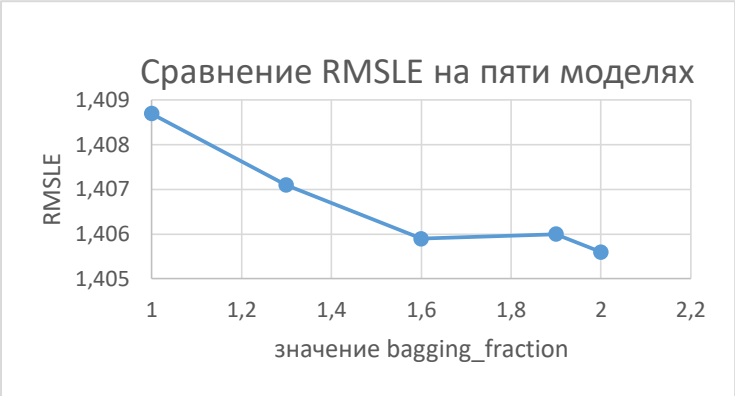

Рисунок 20 – Влияние гиперпараметра bagging_fraction на точность модели на валидационной выборке

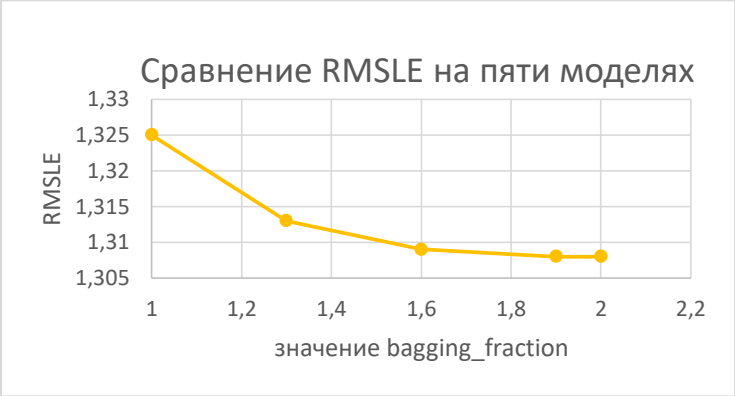

Рисунок 21 – Влияние гиперпараметра bagging_fraction на точность модели на обучающей выборке

На основе предыдущих вычислений, составим финальный стек моделей, изменяя гиперпараметр learning_rate(шаг обучения) от 0,01 до 0,05, используя при этом предыдущие гиперпараметры. Наименьшая ошибка при значении 0,01, она же и будет являться лучшей. (рисунок 22).

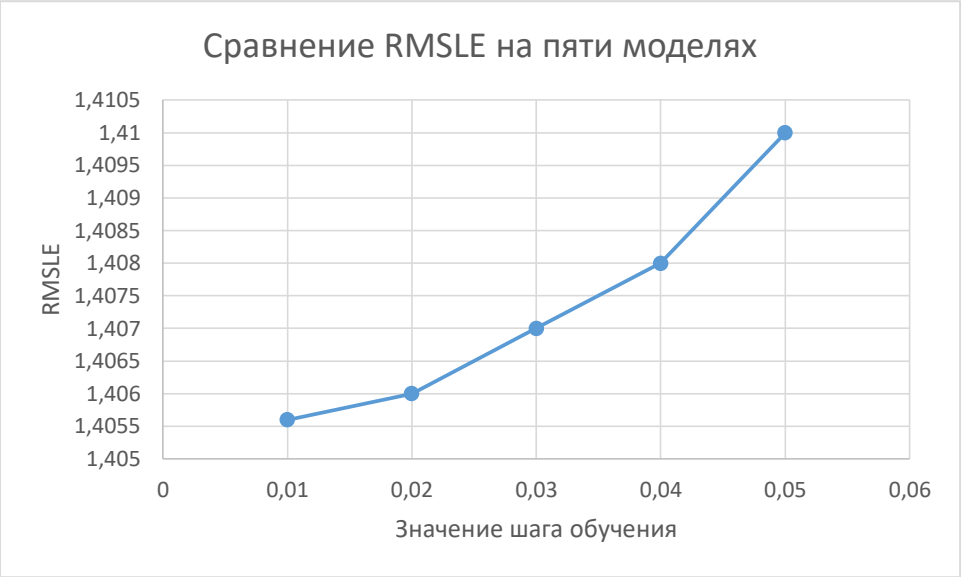

Рисунок 22 – Влияние гиперпараметра learning_rate на точность модели на валидационной выборке

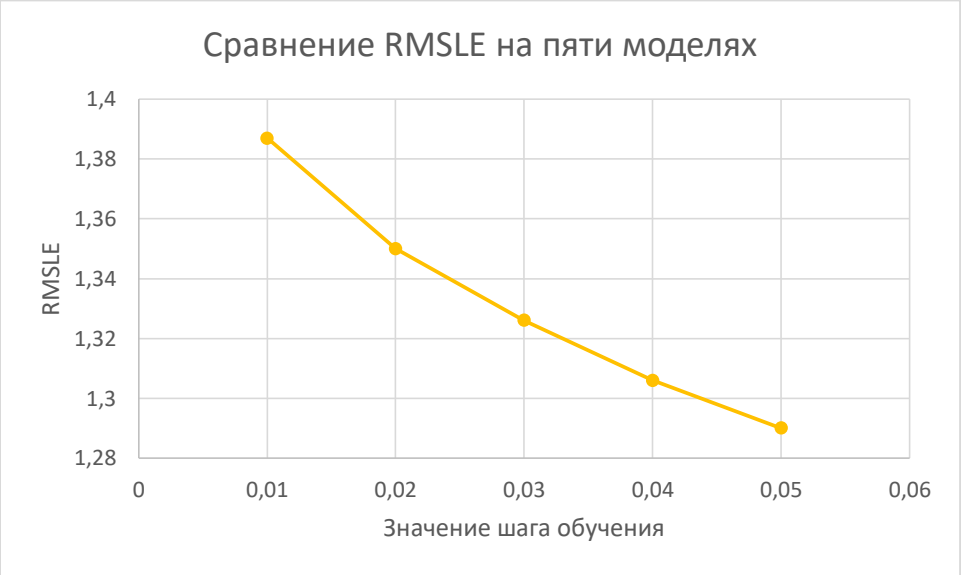

Рисунок 23 – Влияние гиперпараметра learning_rate на точность модели на обучающей выборке

По аналогии с предыдущей моделью, проверим коэффициент детерминации на финальном стеке моделей (Рисунок 24).

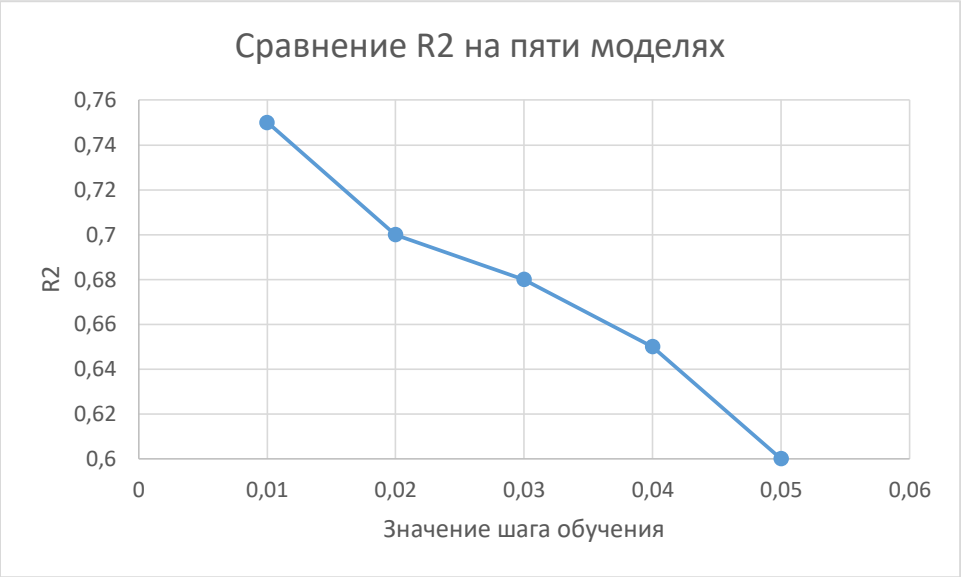

Рисунок 24 – Сравнение моделей по влиянию гиперпараметра learning_rate на коэффициент детерминации

После настройки всех гиперпараметров, данная модель имеет валидационную оценку RMSLE 1,4056 и значение коэффициента детерминации в 0,75 пунктов.

Посмотрим на ключевые переменные в данной модели (Рисунок 25).

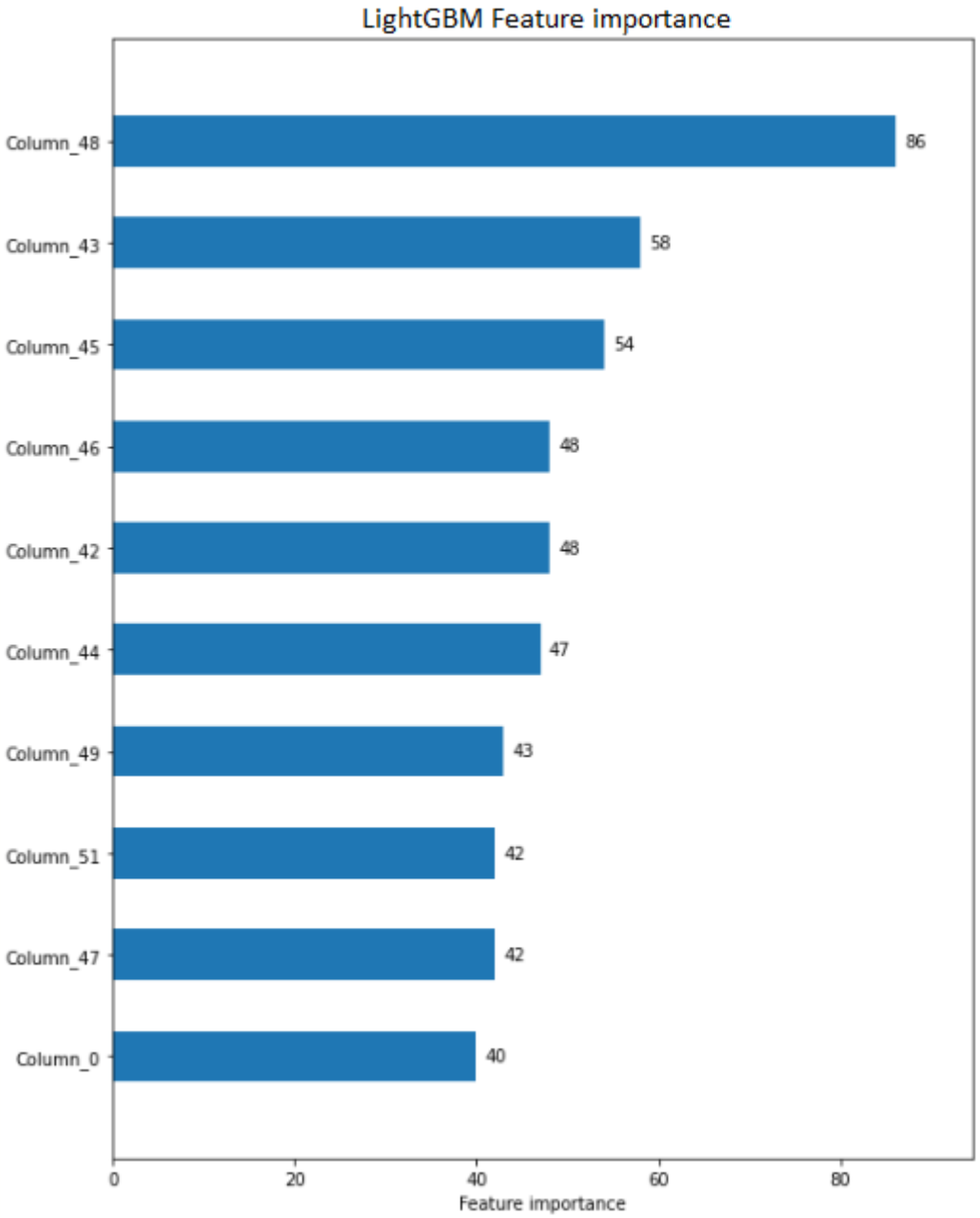

Рисунок 25 – Ключевые переменные модели на базе LightGBM

Самым важный признаком в данной модели стал f48, являющийся 48 колонкой. Разберем поподробнее гиперпараметры данной модели:

- num_leaves – количество листьев в дереве. В отличии от предыдущей модели в данном случае количество листьев имеют большое значение. После подбора весов, было эмпирически выявлено, что оптимальным значением данного параметра будет 120;

- learning_rate – размер шага, используемый в обучении, предотвращает переобучение. Значение параметра установлено на пункте 0.01;
- n_estimators – показатель отвечает за количество деревьев градиентного бустинга. Значение параметра установлено на 500 деревьев, т.к. именно это значение позволяет добиться наивысшей точности нашей модели;
- max_depth – максимальная глубина дерева. Увеличение этого значения сделает модель более сложной. В данном случае оптимальным показателем стало значение 10. Дальнейшей увеличение глубины дерева не дало никаких результатов, значения же ниже 10 ухудшают точность модели. В паре с гиперпараметром num_leaves данный показатель в положительной степени повлиял на точность модели;
- Metric – оценочная метрика модели, в нашем случае это среднеквадратичная логарифмическая ошибка, но т.к. в базовом библиотеке её нет, будет использовать просто средневадратичную ошибку;
- is_training_metric – параметр выводящий результаты работы модели после её обучения и тестирования;
- bagging_fraction – параметр, который случайным образом выбирает часть данных на каждой итерации. В данном случае параметр случайным отбирает 70% данных перед обучением каждого дерева;
- Verbose – отвечает за отображение данных в процессе обучения, поставим его на -1, чтобы не забивать лог ненужной информацией;
- bagging_freq – параметр, отражающий частоту бэгинга на к i итерации. В данном случае бэгинг проводится на каждой итерации (значение 1);
- feature_fraction – параметр, который случайным образом выбирает часть признаков на каждой итерации.

3.5.3 Модель CATBoost

В качестве третьей базовой модели сформируем модель градиентного бустинга в вариации CATBoost.

Для обучения по данному методу, были построены 5 моделей с параметром ошибки RMSLE, с количеством бустинговых деревьев (n_estimators) от 50 до 250, с глубиной дерева 8, с параметром количеством сплитов (max_bin) в 20.

В ходе работе получились следующие результаты. Наименьшая ошибка при построении 100 деревьев, она же и будет являться лучшей. (рисунок 26)

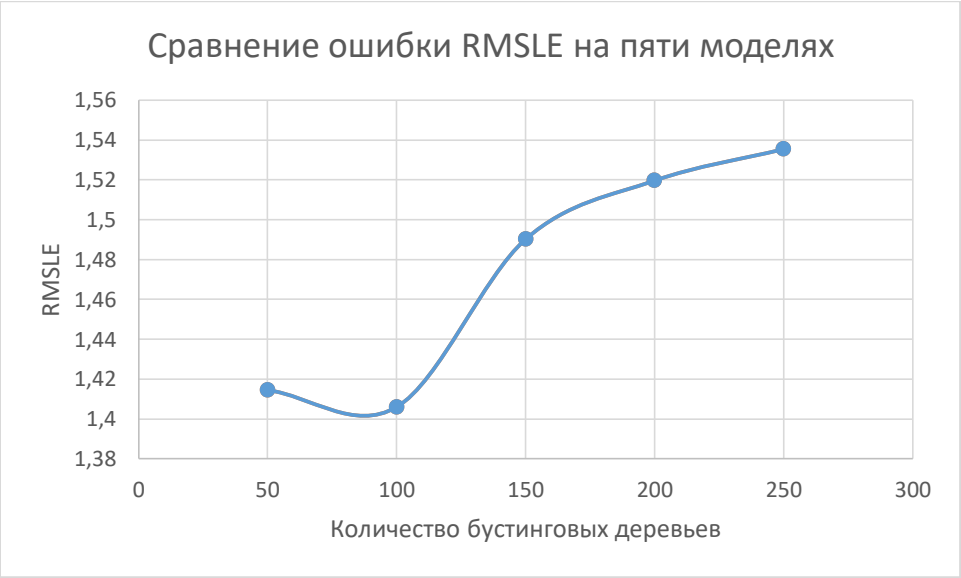

Рисунок 26 – Влияние количества бустинговых деревьев на точность модели на валидационной выборке

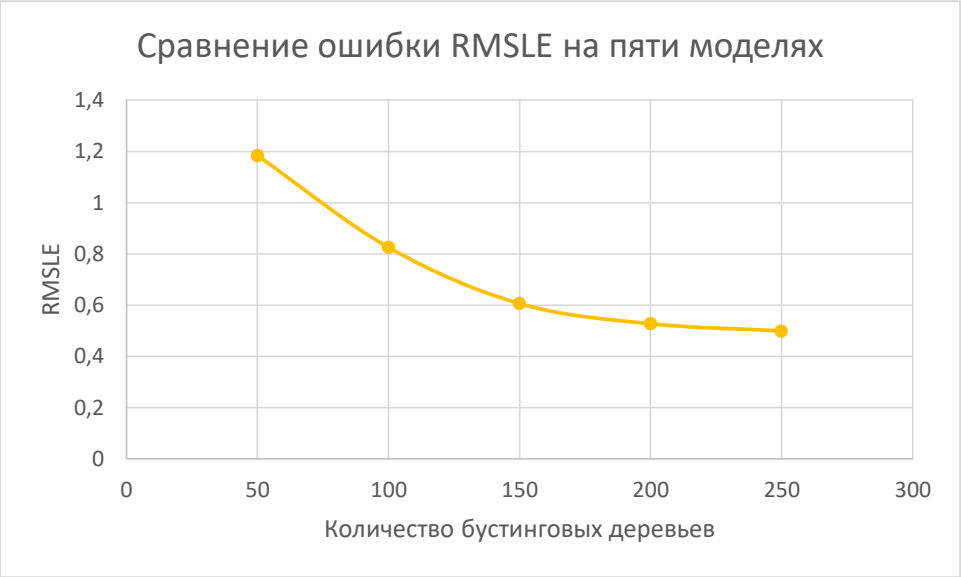

Рисунок 27 – Влияние количества бустинговых деревьев на точность модели на обучающей выборке

Далее, на основе предыдущей проверки, сравним глубину дерева от 5 до 10. Количество деревьев 50, сплитов 50. Наименьшая ошибка была при глубине дерева 7 (Рисунок 28).

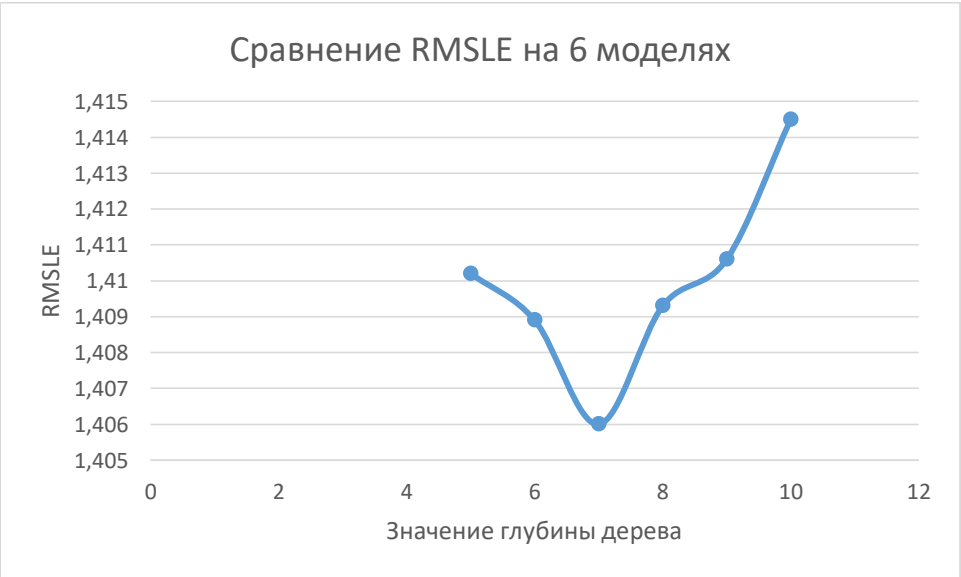

Рисунок 28 – Влияние глубины дерева на точность модели на валидационной выборке

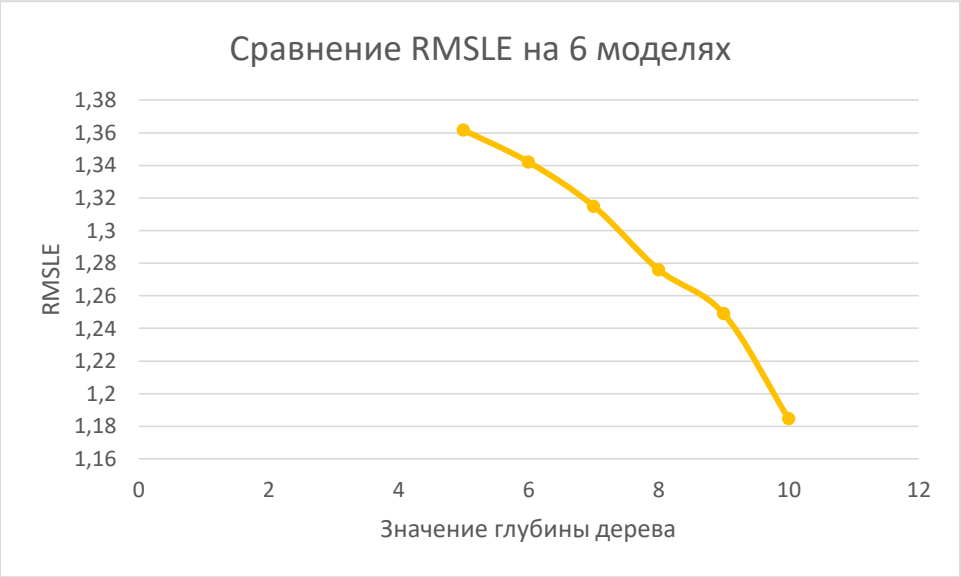

Рисунок 29 – Влияние глубины дерева на точность модели на валидационной выборке

Далее сравним количество сплитов в модели от 20 до 120. Количество деревьев 100, глубина дерева 7. Наиболее оптимальным вариантом стало количество сплитов 20 (Рисунок 30)

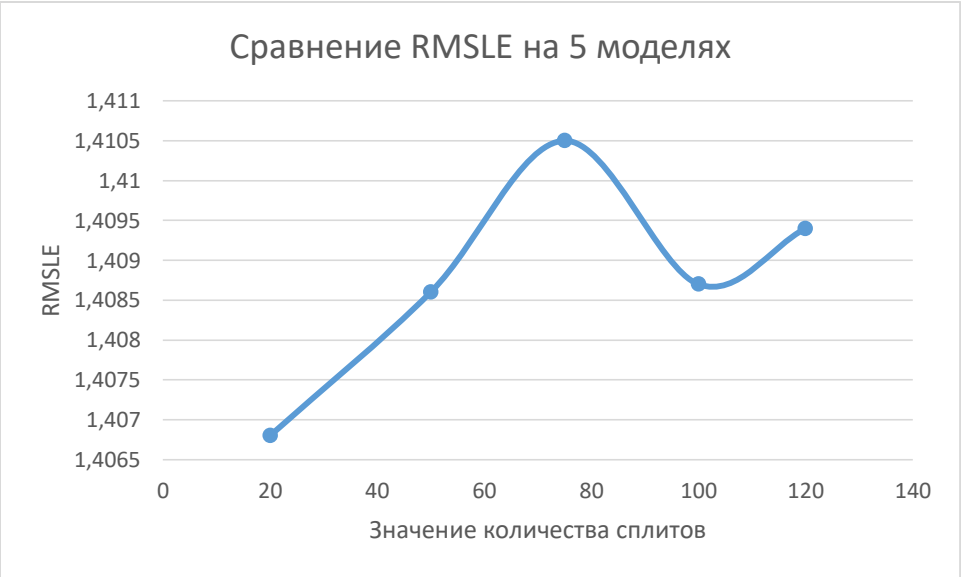

Рисунок 30 – Влияние количества сплитов на точность модели на валидационной выборке

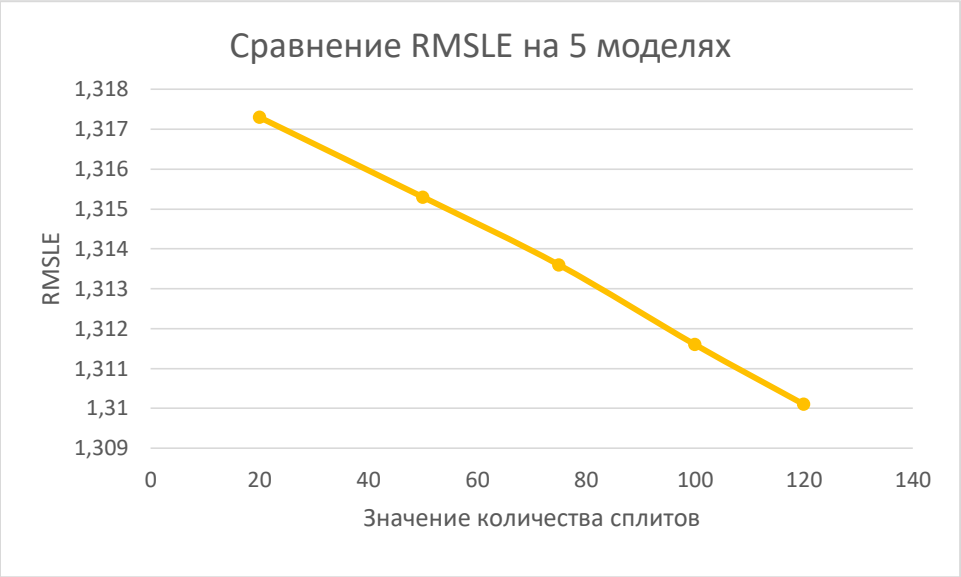

Рисунок 31 – Влияние количества сплитов на точность модели на обучающей выборке

На основе предыдущих вычислений, составим финальный стек моделей, изменяя гиперпараметр learning_rate(шаг обучения) от 0,1 до 0,5, используя при этом предыдущие гиперпараметры. Наименьшая ошибка при значении 0,1, она же и будет являться лучшей. (рисунок 32).

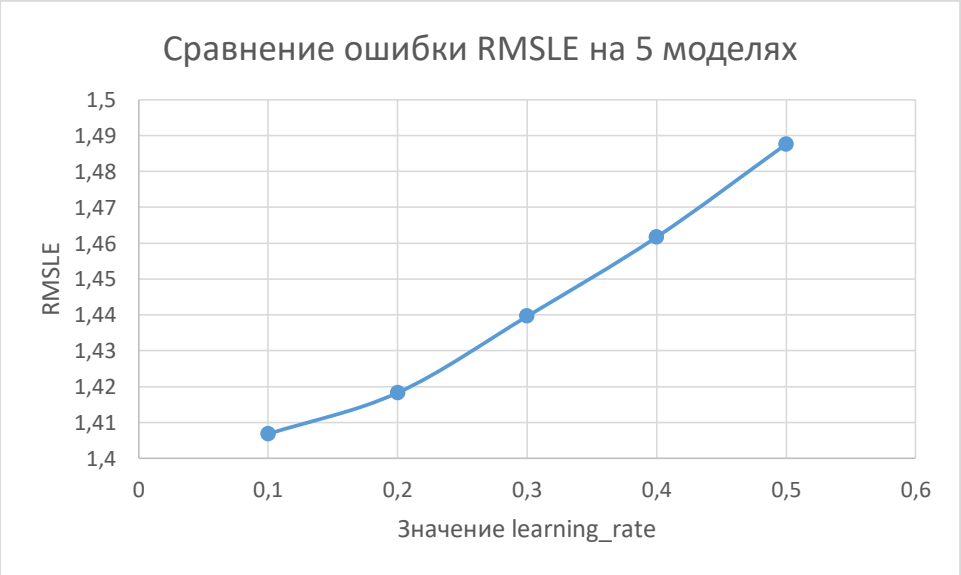

Рисунок 32 – Влияние гиперпараметра learning_rate на точность модели на валидационной выборке

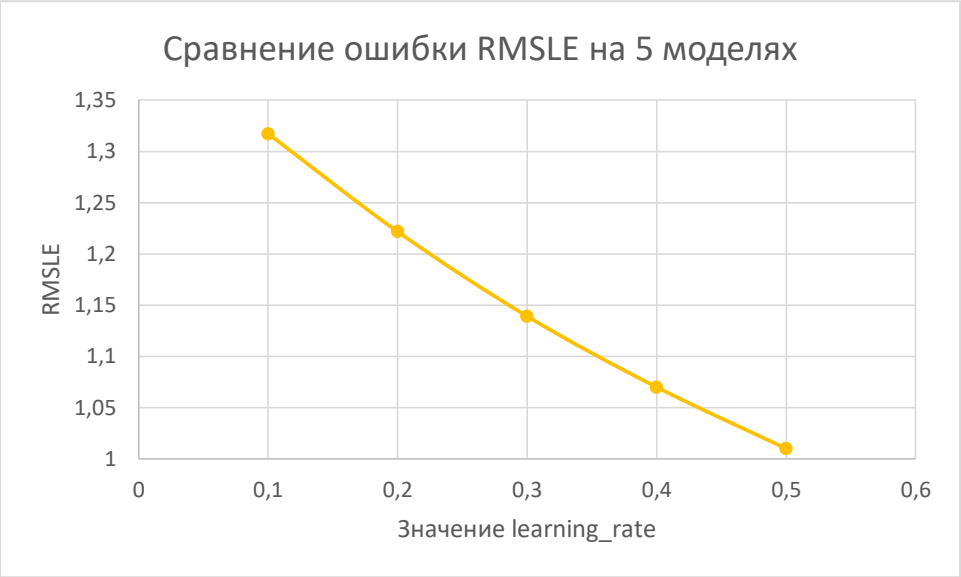

Рисунок 33 – Влияние гиперпараметра learning_rate на точность модели на обучающей выборке

По аналогии с предыдущими моделями, проверим коэффициент детерминации на финальном стеке моделей(Рисунок 34).

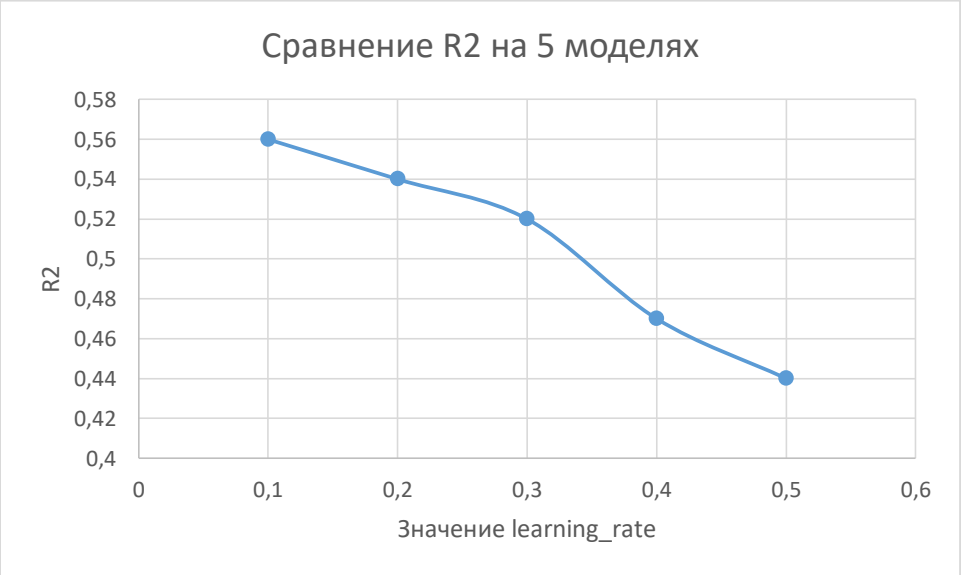

Рисунок 34 – Сравнение моделей по влиянию гиперпараметра learning_rate на коэффициент детерминации

После настройки всех гиперпараметров, данная модель имеет валидационную оценку RMSLE 1,406 и значение коэффициента детерминации в 0,56.

Низкая точность данной модели по сравнению с предыдущими может быть связана с тем что вариация градиентного бустинга CATBoost предназначена в первую очередь для работы с целочисленными данными(int), а не с данными с плавающей точкой(float). В нашем случае у нас все данные имеют тип float.

Разберем поподробнее гиперпараметры данной модели:

- iterations – данный параметр отвечает за количество деревьев в модели, в отличии от предыдущих моделей при использовании вариации CATBoost, модель начинает переучиваться уже на 120 итерации, ввиду этого было сделано решение остановиться на 100 деревьях;
- learning_rate – размер шага, используемый в обучении, предотвращает переобучение. Значение параметра установлено на пункте 0.1. Данное значение было выявлено в процессе подбора весов. Анализируя данный параметр также нельзя не заметить, что размер шага здесь значительно выше, нежели на предыдущих моделях
- depth – максимальная глубина дерева. Увеличение этого значения сделает модель более сложной. В данном случае оптимальным показателем стало значение 7. Дальнейшей увеличение глубины дерева дало только отрицательные результаты. Данный параметр также ниже по сравнению с предыдущими моделями;
- eval_metric – оценочная метрика модели, в нашем случае это среднеквадратичная логарифмическая ошибка, но т.к. в базовом библиотеке её нет, будет использовать просто средневадратичную ошибку;
- random_seed – параметр, отвечающий за случайный номер зерна(сида). В процессе подбора весов, оптимальным показателем стало значение 42;
- od_type – параметр отвечающий за детектор переобучения модели;
- thread_count – количество потоков, использованных в процессе моделирования. Так как мы работает в Google Colab, будет логично использовать все мощности которые предоставляет нам платформа;

- max_bin – данный параметр отвечает за количество сплитов для числовых функций. Оптимальным показателем для данной модели стало значение 20.

3.6 Улучшение модели. Построение усредненной модели

После построения трех базовых моделей следует создать усредненную модель. Ввиду того что последняя модель уступает своим предшественникам в точности, было решено создать две усредненные модели:

1) первая модель будет включать в себя две усредненных модели XGBoost и LightGBM;

2) вторая модель будет включать в себя все три выше представленные модели (XGBoost, LightGBM, CATBoost).

Сравним полученные базовые модели (Рисунок 35,36)

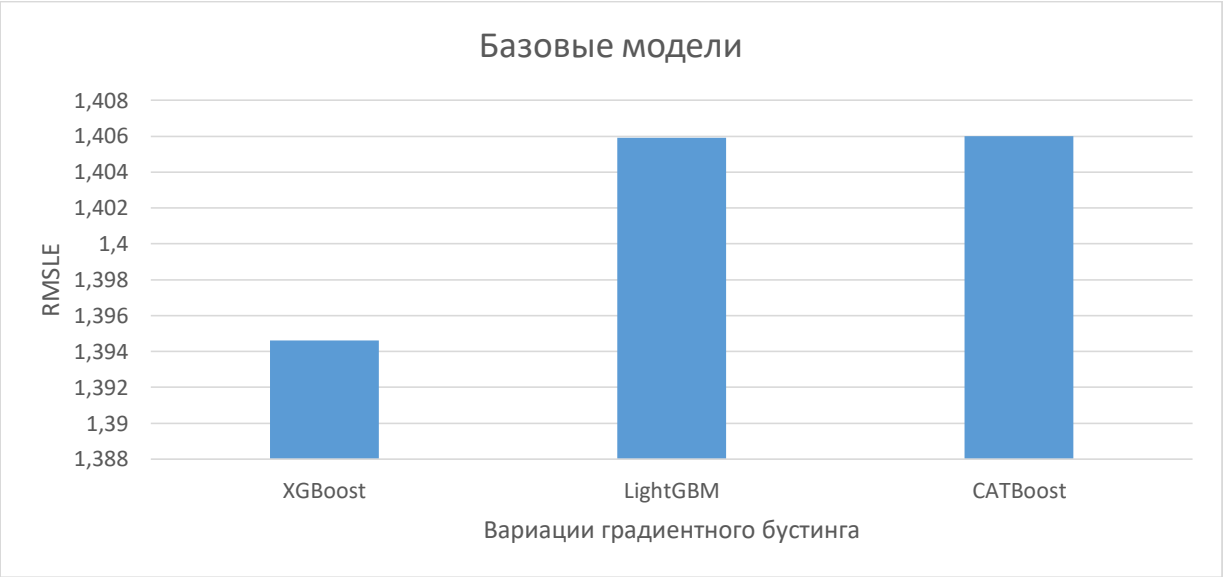

Рисунок 35 – Сравнение базовых моделей (оценка RMSLE)

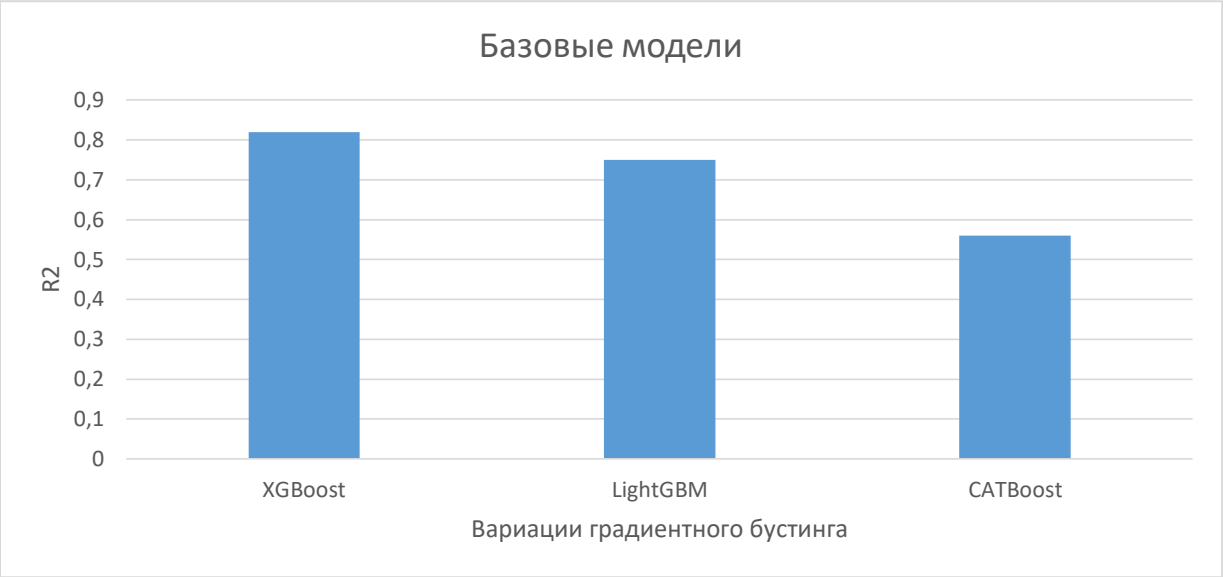

Рисунок 36 – Сравнение базовых моделей (оценка R^2)

Наибольшую точность показала модель на основе экстремального градиентного бустинга: RMSLE = 1,3946

Легкий градиентный бустинг и категориальный бустинг имеют примерно одинаковые значения RMSLE, однако в метрике R^2 CATBoost значительно отстает от своих конкурентов, причины отставания были описаны в предыдущем параграфе.

Сравним полученные усредненные модели(Рисунок 37).

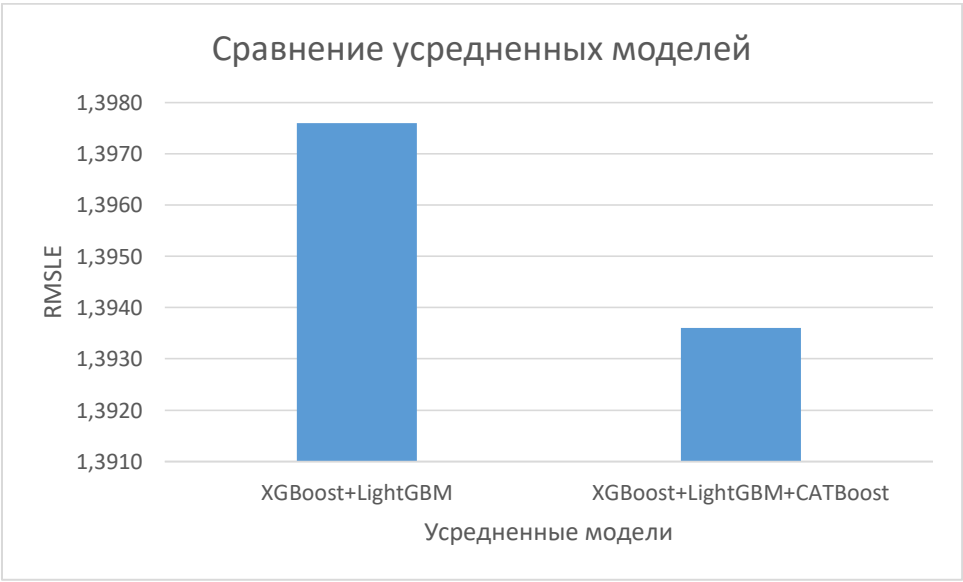

Рисунок 37 – Сравнение усредненных моделей

Модель XGBoost+LightGBM+CATBoost показала наиболее точный результат(1,3636), по сравнению с моделью XGBoost+LightGBM(1,3676), однако данное различие практически несущественно.

3.7 Сравнение моделей и обсуждение результатов

Для тестовой выборки ответы недоступны для пользователей напрямую, так как они хранятся на сайте kaggle. Поэтому оценить точность прогноза можно только путем загрузки полученного прогноза на сайт kaggle. На рисунке 38 представлен скриншот с загруженными данными и тестовой оценкой обеих усредненных моделей.

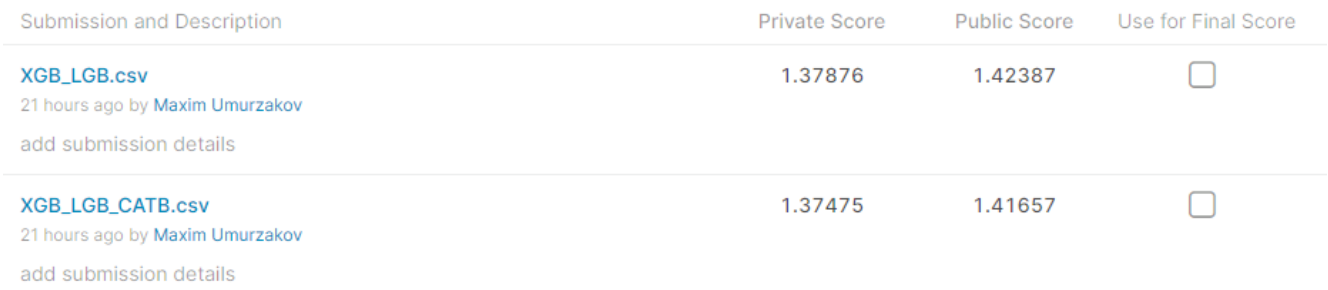

Рисунок 38 – Данные, загруженные на kaggle.com

Private Score дает наиболее полную и точную оценку модели, поэтому будем опираться на данный показатель.

Таблица 7 – Сравнение усредненных моделей

Модель	RMSLE на валидационной выборке	RMSLE на тестовой выборке
XGBoost+LightGBM	1,3976	1,37876
XGBoost+LightGBM+CATBoost	1,3936	1,37475

Как и на валидационной выборке усредненная модель со всеми вариациями показала лучший результат в 1,37475 пунктов.

Из сравнения моделей мы видим, что прогнозировать потенциальную стоимость услуги для клиентов банка лучше всего при использовании всех представленных вариаций градиентного бустинга: экстримального бустинга, «легкого» бустинга и категориального бустинга.

Выводы по главе

1) был проведен разведочный анализ данных, были построены сопутствующие графики;

2) произведено преобразование признаков и добавление новых переменных для увеличения точности моделей;

3) были построены три математические модели, основанные на методе градиентного бустинга в его различных вариациях (XGBoost, LightGBM, CATBoost);

4) были построены две математические модели на базе предыдущих моделей: усредненная XGBoost + LightGBM и усредненная XGBoost + LightGBM + CATBoost;

ГЛАВА 4 ПРОЕКТ КОММЕРЦИАЛИЗАЦИИ

4.1 Дорожная карта коммерциализации

Предлагаемое решение, в частности модель для оценки потенциальной ценности услуги для клиента, реализованное на языке программирования Python, с использованием библиотек машинного обучения, позволит коммерческим банкам минимизировать свои потери, повысить качество обслуживания, и как следствие повысить уровень доверия клиентов и свои конкурентные способности на рынке.

Решаемая задача – оценка потенциальной ценности услуги для клиента коммерческого банка.

В компании планируется работа 2 разработчиков, 1 из которых разрабатывает прогнозные модели, 1 разрабатывает интерфейсы, при условии выпуска 3 моделей каждые месяц, объем реально достижимого объема рынка – 36 программных продукта за год.

В таблице 8 представлена дорожная карта проекта на 2021 г.

Таблица 8 – Дорожная карта проекта на 2021 год

	2021			
	1 квартал	2 квартал	3 квартал	4 квартал
Исследования и разработки	Анализ существующих методов прогнозирования; Сравнение выбранных методов прогнозирования; Формирование оптимального решения на основе рассмотренных методов.	Расчет плановых затрат; Выявление источников финансирования; Формирование бизнес–модели.	Формирование бизнес–модели.	Формирование бизнес–модели.
Создание продукта	Подготовка к анализу, сбор данных; Разработка прототипа проекта.	Разработка прототипа проекта; Тестирование; Выявление слабых мест; Улучшения качества модели.	Работа над ошибками; Улучшения качества модели; Разработка интерфейса системы; Тестирование.	Вывод продукта на рынок
Общее организационное развитие и план по найму	Анализ объема работ.	Формирование плана развития кадрового потенциала; Подбор команды. Обучение команды;	Обучение команды; Анализ проделанной работы.	Анализ проделанной работы.
Защита интеллектуальной собственности		Подача заявки на регистрацию права собственности.		
Маркетинг, внедрение	Исследование рынка аналогичных систем.	Разработка web-представительства компании.	Продвижение на рынок.	
Привлечение инвестиций и продажа		Участие в программах государственной поддержки.	Участие в программах государственной поддержки.	Продажа программного продукта.

4.2 Бизнес-Модель

Таблица 9 – Бизнес-Модель

<table>
<tr><td>Ключевые партнеры</td><td>Ключевые виды деятельности</td><td>Предлагаемые преимущества</td><td>Отношение с клиентами</td><td>Сегменты клиентов</td></tr>
<tr><td rowspan="3">Интерсвязь – основной поставщик Интернета;

Reg.ru –поставщик ресурсов для размещения информации на сервере, постоянно находящемся в сети (Хостинг);

Новостные сайты и блоги.</td><td>Проектирование и разработка предиктивных систем;

Консультирование заказчиков;

Внедрение разработанных систем;

Поддержка разработанных систем.</td><td rowspan="3">Основу анализа составляют данные находящиеся во внутренне среде предприятия;

Продукт позволяет увеличить скорость реагирования банка спрос клиентов;

Невысокая цена по сравнению с конкурентами.</td><td>Индивидуальный подход к каждому клиенту, за счет специфики данных;</td><td rowspan="3">Финансовые учреждения.</td></tr>
<tr><td>Ключевые ресурсы</td><td>Каналы сбыта</td></tr>
<tr><td>Данные предприятия;
Трудовые ресурсы;
Финансовые ресурсы;
Информационные ресурсы;
Материальные ресурсы.</td><td>Web-представительство компании.</td></tr>
<tr><td colspan="3">Структура расходов</td><td colspan="2">Структура доходов</td></tr>
<tr><td colspan="3">Постоянные издержки: Налоги; Арендная плата; Оборудование; Оплата труда административного персонала.
Переменные издержки: Маркетинг; Затраты на консультирования; Оплата труда производственного персонала.</td><td colspan="2">Продажа прав на пользование программным продуктом;
Поддержка программного продукта.</td></tr>
</table>

4.3 Команда проекта

Таблица 10 – Команда проекта

Необходимые роли в проекте	Обоснование, краткое описание функций
Руководитель проекта	Менеджер проекта выполняет огромное количество работ, начиная от разработки плана проекта, оценки рисков, контроля функциональных и стоимостных рамок и заканчивая ежедневной работой с командой на проекте.
Разработчики (2 человека)	Это ключевые люди в любой ИТ команде, именно они занимаются непосредственным созданием программного продукта или сложным конфигурированием базового коробочного решения.
Бухгалтер	Специалист по бухгалтерскому учёту, работающий по системе учёта в соответствии с действующим законодательством.

4.4 Ценообразование

Для расчета себестоимости продукта необходимо знать объем затрат на разработку ПО. Группировку затрат будем производить по экономическим элементам, а именно:

1. материальные затраты;
2. затраты на оплату труда;
3. амортизация основных средств;
4. прочие затраты.

Материальные затраты рассчитываются по формуле 4.1.

$$З_м = \sum Q_i \cdot З_i, \quad (4.1)$$

где $З_м$ – затраты на материалы;

Q_i – количество;

$З_i$ – затраты на единицу.

Затраты на материалы представлены в таблице 11.

Таблица 11 – Затраты на материалы

Наименование	Единица измерения	Стоимость за единицу, руб.	Количество, шт.	Сумма, руб.
Бумага для принтера	Пачка	230	2	460
USB – флэш накопитель	Штук	800	3	2 400
Ручка	Штук	15	10	150
Картридж	Штук	880	1	880
Итого				3 890

Затраты на оплату труда будем рассчитывать следующим образом:

$$З_п = \sum(O_i + O_i \cdot C) \cdot G, \quad (4.2)$$

где $З_п$ – месячный фонд оплаты труда;

O_i – оклад;

C – страховые взносы, С=0,34;

G – занятость.

Таблица 12 – Затраты на оплату труда

Наименование	Оклад (без страховых взносов), руб.	Страховые сборы, руб.	Занятость, %	Сумма, руб.
Руководитель проектов	75 000	25 500	65	65 325
Python – разработчик	67 000	22 780	90	80 802
Разработчик интерфейсов	63 000	20 400	60	50 040
Итого				196 167

Расчет затрат на амортизацию будем производить по следующий формуле:

$$A_{мес} = \sum \frac{С_i}{С_c \cdot T} \cdot З_i \quad (4.3)$$

где $A_{мес}$ – амортизация за месяц;

$С_i$– первоначальная стоимость;

$С_c$ – срок службы (год);

Т – количество месяцев в году (12);

$З_i$ – загруженость.

Таблица 13 – Затраты на амортизацию

Наименование	Кол-во	Цена, руб.	Сумма, руб.	Срок службы, месяцев	Амортизация в месяц, руб.	Загруженность %	Сумма, руб.
Ноутбук LENOVO IdeaPad S145-15AST	1	21890	21890	36	608,06	85	516,85
Ультрабук MSI PS42 Modern 8MO-463XRU	2	56890	113780	48	2370,42	85	2014,86
Windows 10 корпорат-я	3	15000	60000	48	1250	100	1250
MS Office	3	5000	20000	36	556	100	556
Антивирус Kaspersky Security	3	1600	6400	12	533	100	533
Принтер лазерный XEROX Phaser 3020	1	6730	6730	36	190	30	57
ИТОГО							4 927, 71

Так же стоит отразить прочие затраты (см. таблицу 7). В состав арендных платежей входит стоимость аренды и обслуживания помещения.

Таблица 14 – Прочие затраты

Наименование	Затраты в месяц, руб.	Количество, шт.	Сумма, руб.
Аренда помещения	12000 за 26 м^2	1	12 000
Хостинг	333	1	333
Интернет	1100	1	1 100
Итого			13 433

Суммарные затраты на разработку рассчитываются по формуле:

$$З = \sum З_{мес} \cdot \text{tp} \quad (4)$$

где З – суммарные затраты;

$З_{мес}$ – затраты за месяц;

tp – время на разработку.

Таблица 15 – Суммарные затраты

Наименование	Затраты в месяц, руб.
Материальные затраты	3 890
Затраты на оплату труда	196 167
Амортизация основных средств	4927, 71
Прочие затраты	13 433
Итого	218 417, 71

Для расчета себестоимости 1 единицы продукта, разделим затраты на переменные и постоянные, а также воспользуемся формулой

$$\text{Полная себестоимость} = \frac{З_{пер} \cdot Q + З_{пост}}{Q} \quad (4.5)$$

где $З_{пер}$ – переменные затраты;

Q – целевой объем продаж;

$З_{пост}$ – постоянные затраты.

Таблица 16 – Переменные и постоянные издержки

Наименование	Затраты на разработку	Время на разработку, месяцев	Сумма, руб.
Переменные издержки			
Затраты на оплату труда	176 880	1	176 880
Итого			176 880
Постоянные издержки			
Материальные затраты	4 160	1	4 160
Амортизация основных средств	6 456	1	6 456
Прочие затраты	19 349	1	19 349
Итого			29 965

Себестоимость 1 единицы продукции составляет 218 417 руб. 71 коп. Установим целевой объем продаж в месяц равным 3. Зная себестоимость единицы продукции, мы можем рассчитать оптимальную цену и прибыль. Рассчитывать будем по формуле 28:

$$\text{Отпускная цена за 1 единицу} = C + (C * q) \quad (4.6)$$

где С – себестоимость единицы продукции;

q – процент прибыли от продаж (20 %).

$$\text{Пр} = \text{Отпускная цена за 1 ед} - \text{С} \quad (4.7)$$

где С – себестоимость единицы продукции;

ПР – прибыль.

$$S = \text{Отпускная цена за 1 ед} \cdot r \quad (4.8)$$

где S – цена для потребителя;

r – НДС (20 %).

Из этого следует что, при наценке компании в 20%, отпускная цена за единицу продукции составит 252 101 руб. 26 коп., а так как у нас еще есть НДС, то цена для потребителя составит 314 521 руб. 51 коп., соответственно прибыль составит 43 683 руб. 55 коп. с одного проекта, а так как планируется выполнять два проекта в месяц, то прибыль составит 131 051 рубль в месяц.

Выводы по главе

В данной главе был рассмотрен проект создания сервиса прогнозирования совершения ключевого действия клиентом банка, проанализированы и выбраны возможные методы коммерциализации, были рассчитаны приблизительные затраты, которые составили 218 472 рубля, составлена дорожная карта проекта, с описанием процедур и действий на этапе внедрения.

ЗАКЛЮЧЕНИЕ

В ходе написания магистерской диссертационной работы, был рассмотрен процесс клиентооттока, рассмотрены факторы, влияющие на него, выявлена проблема, а также рассмотрена роль прогнозирования совершения ключевого действия ключевого действия клиента. Для оценки совершения ключевого действия клиентов, были использованы математические модели прогнозирования потенциальной ценности услуги для клиентов. В качестве инструментов прогнозирования рассматривались методы машинного обучения, а конкретно три вариации градиентного бустинга. Сравнительный анализ методов позволил выбрать необходимые инструменты для анализа, а именно вариации XGBoost, LightGBM и CATBoost.

В ходе работы были построены три модели на базе выше приведенных методов. В процессе работы, были построены три базовые модели, реализованные с помощью метода градиентного бустинга были выбраны три библиотеки, это:

- XGBoost, которая показала результат равный 1.3946;
- LightGBM, которая показала результат равный 1.4059;
- CatBoost, которая показала результат равный 1.4060.

Далее были построены две усредненные модели: XGBoost+LightGBM+CATBoost и XGBoost+LightGBM Данные модели разрабатывались с помощью библиотек Sklearn, xgboost, lightgbm и catboost и дали среднеквадратическую логарифмическую ошибку равную 1.37475 и 1.37876 соответственно.

В заключение работы был рассмотрен проект коммерциализации. Была построена дорожная карта проекта, включающая в себя этапы необходимые для вывода проекта на рынок. Были произведены плановые расчеты затрат на реализацию проекта, а также рассчитана цена на единицу продукции.

Таким образом была достигнута цель исследования и были выполнены поставленные задачи в начале исследования.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1 Боднар А. Ю. Информационно-сервисные инструменты обслуживания клиентов как основа обеспечения конкурентоустойчивости банков: дис. ... канд. экон. наук: 08.00.05. - Ростов-на-Дону, 2012. – 166 с.

2 Владыка М.В., Гончаренко Т.В. Повышение финансовой грамотности населения региона как фактор лояльности клиентов банка./М.В. Владыка, Т.В. Гончаренко // Банковское дело. – 2012. №3(11). – С. 20–25.

3 Гаврилов, Л.П. Основы электронной коммерции и бизнеса : учебное пособие / Л.П. Гаврилов. — Москва : СОЛОН-Пресс, 2009. – 592 с. – Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. – URL: https://e.lanbook.com/book/13783 (дата обращения: 25.11.2019). – Режим доступа: для авториз. пользователей.

4 Коэльо, Л.П. Построение систем машинного обучения на языке Python / Л.П. Коэльо, В. Ричарт ; перевод с английского А.А. Слинкин. – 2-е изд. — Москва : ДМК Пресс, 2016. — 302 с. – Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. – URL: https://e.lanbook.com/book/82818 (дата обращения: 28.11.2019). – Режим доступа: для авториз. пользователей.

5 Мухаметзянова, Е. С. Повышение качества обслуживания физических лиц в коммерческих банках: на примере ОАО "Сбербанк России": дис. ... канд. экон. наук: 08.00.10. - М., 2014. – 201 с.

6 Пальмов С. В. Анализ и прогнозирование оттока клиентов в телекоммуникационных компаниях на основе технологии Data Mining : дис. ... канд. техн. наук: 05.13.13. – Самара, 2005. – 172 с.

7 Путилов, А.В. Коммерциализация технологий и промышленные инновации : учебное пособие / А.В. Путилов, Ю.В. Черняховская. –Санкт-Петербург : Лань, 2018. –324 с. – Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. — URL: https://e.lanbook.com/book/110937 (дата обращения: 28.11.2019). — Режим доступа: для авториз. пользователей.

8 Рашка, С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения : руководство / С. Рашка ; перевод с английского А.В. Логунова. – Москва : ДМК Пресс, 2017. – 418 с. – Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. — URL: https://e.lanbook.com/book/100905 (дата обращения: 26.11.2019). — Режим доступа: для авториз. пользователей.

9 Романов В. В. Система взаимоотношений российского коммерческого банка с клиентами: дис. ... канд. экон. наук: 08.00.10. – СПб., 2004. – 156 с.

10 Салмин А.А. Формирование оценки лояльности клиентов телекоммуникационной компании на основе байесовского подхода: дис. ... канд. техн. наук: 05.13.13. - Самара, 2008. – 181 с.

11 Сергеенкова А. А. Современные технологии обеспечения конкурентоспособности многофилиального коммерческого банка на рынке финансовых услуг: дис. ... канд. экон. наук: 08.00.10. - Ростов-на-Дону, 2007. – 151 с.

12 Сковиков, А.Г. Цифровая экономика. Электронный бизнес и электронная коммерция : учебное пособие / А.Г. Сковиков. – Санкт-Петербург : Лань, 2019. – 260 с. – Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. — URL: https://e.lanbook.com/book/119637 (дата обращения: 28.11.2019). — Режим доступа: для авториз. пользователей.

13 Хаустова М. Н. Индивидуальное банковское обслуживание состоятельных клиентов как сегмент международных финансов: дис. ... канд. экон. наук: 08.00.14. – СПб., 2012. – 139 с.

14 Шарден, Б. Крупномасштабное машинное обучение вместе с Python : учебное пособие / Б. Шарден, Л. Массарон, А. Боскетти ; перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2018. – 358 с. – Текст : электронный // Электронно-библиотечная система «Лань» : [сайт]. – URL:

https://e.lanbook.com/book/105836 (дата обращения: 25.11.2019). — Режим доступа: для авториз. пользователей.

15 Alex Labram: Fitting data with XGBoost // Institute and Facility of Actuaries, 2019. – P. 33–42.

16 Anil Kumar D, Ravi V: Predicting credit card customer churn in banks using data mining // International Journal of Data Analytics Technical Strategy, 2008. – P. 19–25.

17 Au W., Chan C.C., Yao X.: A Novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans. on evolutionary comp. 7, 2003. – P. 30–43.

18 Buckinx W., Van den Poel D.: Customer base analysis: partial detection of behaviorally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research 164, 2005. – P. 7–15.

19 CATBoost documentation – 2020. – 63 p. [Электронный ресурс] – Режим доступа: https://catboost.ai/docs/concepts/parameter-tuning.html (дата обращения 23.02.2020)

20 Durkin Mark G, Howcroft Barry: Relationship Marketing in banking Sector: Impact of New Technologies//Marketing Intelligence Planning Vol. 21 No.1, 2008. – P. 61–71.

21 Ennew, C. T., Binks, M. R., &Chiplin, B: Customer Satisfaction and Customer Retention: An Examination of Small Businesses and Their Banks in the UK. In Proceedings of the 1994 Academy of Marketing Science (AMS) //Annual Conference. Springer International Publishing, 2015. – P. 182–196.

22 Ester, M.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / M. Ester, H.-P. Kriegel, J. Sander, X. Xu // In Proc. ACM SIGMOD Int. Conf. on Management of Data, Portland, OR, 1996. –P. 226-231.

23 F. Pedregosa: Scikit-learn: machine learning in Python/G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Perrot, É. Duchesnay // Journal of Machine Learning Research. 12, 2011. – P 2825–2830.

24 Farquad MAH, Ravi V, Raju SB Churn prediction using comprehensible support vector machine:An analytical CRM application // Applied Soft Computing, 2014. – P. 27-32.

25 Ferreira J., Vellasco M., Pachecco M., Barbosa C.: Data mining techniques on the evaluation of wireless churn. ESANN2004 proceedings - European Symposium on Artificial Neural Networks Bruges, 2004. – P. 76-89.

26 G. Sehgal: Comparison of various clustering algorithms/G. Sehgal, K. Garg // International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014. – P. 3074–3076.

27 Garland R.: Investigating indicators of customer profitability in personal retail banking. Proc. of the Third Annual Hawaii Int. Conf. on Business, 2003. – P. 153–166.

28 Gregory, Bryan: Predicting Customer Churn: Extreme Gradient Boosting with Temporal Data // ResearchGate, Berlin, 2019. – P. 85–98.

29 Guo-en X, Wei-dong J: Model of customer churn prediction on support vector machine// Systems Engineering Theory Pract 28, 2008. – P. 55–70.

30 Guolin Ke , Qi Meng: LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Pekin University, People Republic of China, 2017. – P. 121–139.

31 Halkidi M On clustering validation techniques/ Halkidi M, Batistakis Y, Vazirgiannis M // J. Intell. Inf. Syst. 2001. – P. 107–145.

32 Huang B, Kechadi MT, Buckley B: Customer churn prediction in telecommunications // Expert Systems with Applications, 2012. – P. 143–167.

33 HussainIftikhar, HussainMazhar, HussainShahid and Sajid M.A.: Customer Relationship Management: Strategies And Practices In Selected Banks Of Pakistan, International Review of Business Research Papers Vol. 5 No. 6, 2009. – P.117–132.

34 Hwang H., Jung T., Suh E.: An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications 26, 2004. – P. 87–106.

35 J. McCaffrey Clustering Non-Numeric Data Using Python [Электронный документ] https://visualstudiomagazine.com/articles/2018/04/01/.. Проверено (дата обращения 25.11.2019)

36 Jin Zhang: LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity–Application to the Tox21 and Mutagenicity Data Sets // Karolinska Institutet, Sweden, 2019. – P. 254–276.

37 Jing Lei: Cross-Validation With Confidence // Journal of the American Statistical Association, Pittsburgh, 2019. – P. 165–183.

38 Karypis, G. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling / G. Karypis, E.-H. Han, V. Kumar // Journal Computer Volume 32 Issue 8. IEEE Computer Society Press Los Alamitos, CA, 1999. – P 68–75.

39 Keaveney S., Parthasarathy M.: Customer Switching Behaviour in Online Services: An Exploratory Study of the Role of Selected Attitudinal, Behavioral, and Demographic Factors. Journal of the Academy of Marketing Science 29, New-York, 2001. – P. 5–20.

40 Keramati A, Ardabili SMS: Churn analysis for an Iranian mobile operator // Telecommunication Policy, 2011. – P. 75–93.

41 Keramati A, Jafari-Marandi R, Aliannejadi M, Ahmadian I, Mozzafari M, Abbasi U: Improved churn prediction in telecommunication industry using data mining techniques // Applied Soft Computing, 2014. – P. 91–111.

42 Leo Breiman: Technical Report 486 , Statistics Department University of California, Berkeley, 1999. – P. 154–177.

43 Lester L.: Read the Signals//Target Marketing 28, Detroit, 2005. – P. 98–117.

44 LightGBM Documentation – 2020. – 103 p. [Электронный ресурс] – Режим доступа: https://lightgbm.readthedocs.io/en/latest/Parameters.html (дата обращения 23.02.2020)

45 Lin C-S, Tzeng G-H, Chin Y-C: Combined rough set theory and flow network graph to predict customer churn in credit card accounts // Expert Systems with Applications, 2011. – P. 23–37.

46 Liudmila Prokhorenkova, Gleb Gusev: CatBoost: unbiased boosting with categorical features // Cornell University, New York, 2017. – 30 p.

47 Machado, M: LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry // ResearchGate, Berlin, 2019. – P. 65-78.

48 Mason, L.; Baxter, J.; Bartlett, P. L.; Frean, Marcus : Boosting Algorithms as Gradient Descent // Advances in Neural Information Processing Systems 12. MIT Press, 1999. – P. 76–89.

49 Mason, L.; Baxter, J.; Bartlett, P. L.; Frean, Marcus: Boosting Algorithms as Gradient Descent in Function Space // Research School of Information Sciences and Engineering Australian National University Canberra, 1999. – P. 43–68.

50 Max A Little: Using and understanding cross-validation strategies. Perspectives on Saeb et al // GigaScience, Volume 6, Issue 5, May 2017. – P. 77–91.

51 Md. Sohrab Mahmud: Improvement of K-means clustering algorithm with better initial centroids based on weighted average /Md. Sohrab Mahmud, Md. Mostafizer Rahman, Md. Nasim Akhtar// 7th International Conference on Electrical and Computer Engineering 2012. – P 647–650.

52 Min Huang: Improved K-means clustering center selecting algorithm/ Min Huang, Lei Yu, Ying Chen// Information Engineering and Applications 2012 pp 373-379 Madhu Yedla Enhancing, K-means Clustering Algorithm with Improved Initial Center/ Madhu Yedla, Srinivasa Rao Pathakota, T.M Srinivasa //International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 1 (2) 2010. P. 121–125.

53 Mozer M. C., Wolniewicz R., Grimes D.B., Johnson E., Kaushansky H.: Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunication Industry. IEEE Transactions on Neural Networks, 2000. – P. 111–120.

54 Nie G, Rowe W, Zhang L, Tian Y, Shi Y: Credit card churn forecasting by logistic regression and decision tree // Expert Systems with Applications, 2011. – P. 99–121.

55 ParmarJitesh, Sadanand Vijay Kumar: Customer Relationship Management (CRM) Best Practices and Customer Loyalty: A Study of Indian Retail Banking Sector//European journal of social Sciences Vol. 11 No. 1, 2009. – P. 61–85.

56 Peevers, G., Douglas, G., Marshall, D., & Jack, M. A. :On the role of SMS for transaction confirmation with IVR telephone banking //International Journal of Bank Marketing, 29(3), 2011. – P. 206–223.

57 Saradhi VV, Palshikar GK: Employee churn prediction // Expert Systems with Applications, 2011. – P. 33–55.

58 Sharma A, Panigrahi PK: A neural network based approach for predicting customer churn in cellular network services // International Journal of Computer Applications, 2011. – P. 76–93. – Текст: электронный: [сайт].

59 Shi Na: Research on k-means clustering algorithm / Shi Na, Liu Xumin, Guan yong //2010 Third International Symposium on Intelligent Information Technology and Security Informatics 2010. – P. 55–67.

60 StanleyShirmila: New Perspectives in the banking sector –The CRM way// International Journal of Marketing, Financial Services & Management Research, Vol.1 Issue 11, November 2012. – P. 154–169.

61 Teemu Mutanen: Customer churn prediction - A case study in retail banking // Workshop on practical Data Mining // Berlin, Germany , 2006. – P. 5–19.

62 Tianqi Chen, Carlos Guestrin: XGBoost: A Scalable Tree Boosting System // University of Washington, 2016.

63 Ting Hu and Ting Song: Research on XGboost academic forecasting and analysis modelling // Journal of Physics: Conference Series, Volume 1324, United Kingdom, 2019. – P. 76–101.

64 Tony Duan, Anand Avati: NGBoost: Natural Gradient Boosting for Probabilistic Prediction // Cornell University, New York, 2019. – P. 167–191.

65 Uppal R.K.:Customer Service in Indian Commercial Banks: An Empirical Study//Asia Pacific Journal of Social Sciences, Vol. 1 No.1, 2010, – P. 127–141.

66 Xgboost Documentation – 2020. – 130 p. [Электронный ресурс] – Режим доступа: https://xgboost.readthedocs.io/en/latest/parameter.html (дата обращения 23.02.2020).

67 Yu X, Guo S, Guo J, Huang X: An extended support vector machine forecasting framework for customer churn in e-commerce // Expert Systems with Applications, 2011. – P. 57–79.

68 Zhang, T. BIRCH: An Efficient Data Clustering Method for Very Large Databases / T. Zhang, R. Ramakrishnan, M. Linvy // In Proc. ACM SIGMOD Int. Conf. on Management of Data. ACM Press, New York, 1996. – P.103–114.