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

ПРОЕКТ ПРОВЕРЕН
Рецензент, начальник управления реализации проектов Министерства информационных технологий и связи Челябинской области
_____ (И.А. Филатов)
«____» _____ 2019 г.

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий кафедрой, д.т.н., с.н.с.
_____ (Б.М. Суховилов)
«____» _____ 2019 г.

РАЗРАБОТКА МАТЕМАТИЧЕСКОЙ МОДЕЛИ И СЕРВИСА КРЕДИТНОГО СКОРИНГА ДЛЯ АНАЛИЗА ПЛАТЕЖЕСПОСОБНОСТИ КЛИЕНТОВ БАНКА

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ
ЮУрГУ–38.04.05.2019.131.ПЗ ВКР

Руководитель проекта, д.т.н.
_____ (В.В. Мокеев)
«____» _____ 2019 г.

Автор проекта,
студент группы ЭУ– 244
_____ (А.А. Тютёва)
«____» _____ 2019 г.

Нормоконтролер, доцент
_____ (Е.В. Бунова)
«____» _____ 2019 г.

Челябинск 2019

АННОТАЦИЯ

Тютёва А.А., Разработка математической модели и сервиса кредитного скоринга для анализа платежеспособности клиентов банка. – Челябинск: ЮУрГУ, ЭУ-244, 2019. – 71 с., 18 ил., 8 табл., библиографический список – 35 наим.

Выпускная квалификационная работа посвящена разработке математических моделей и сервиса кредитного скоринга для анализа платежеспособности клиентов банка.

В работе представлены материалы исследования кредитного скоринга для выявления платежеспособности клиентов банка.

Целью работы является снижение расходов и увеличение прибыли банков.

Проведен анализ машинного обучения, а также обоснован выбор, почему именно его следует использовать для анализа платежеспособности клиентов банков. Представлены методы машинного обучения, а также сделан выбор в сторону метода, использованного при анализе платежеспособности клиентов. В работе присутствует описание выбранного метода, исходные данные предоставленные банком Хоум Кредит, предварительная обработка данных, а также результаты проведенной работы. Описана дорожная карта коммерциализации проекта, создан сайт по предоставлению услуги прогнозирования платежеспособности клиентов банка. Рассчитан медиаплан и ценовая политика коммерциализации проекта.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Кредитные отношения – один из наиболее важных аспектов современной экономической деятельности. Эффективность кредитной системы обуславливает успешное развитие производства и социально-экономического прогресса.

При помощи кредита сокращается время на удовлетворение хозяйственных потребностей. Предприятие – заемщик за счет дополнительных средств имеет возможность увеличить свои ресурсы, расширить хозяйство, ускорить достижение производственных целей. Таким образом, кредит выступает опорой современной экономики и неотъемлемой частью экономического развития. Его используют как крупные предприятия и объединения, так и малые производственные, сельскохозяйственные и торговые структуры.

Доходы от кредитных операций являются основным источником прибыли. Однако невозврат кредитов может привести к банкротству. Именно поэтому так важно подобрать для каждого клиента правильный кредитный продукт, а так же заранее распознать проблемного заемщика.

В настоящее время доход банка непосредственно зависит от качества оценки кредитного риска. В зависимости от принадлежности клиента к определенной группе риска, банк принимает решение о его кредитовании или не кредитовании. В современных банках используют два подхода для оценки риска кредитования:

– на основе мнения экспертов;

– с помощью системы кредитного скоринга.

Для оценки кредитоспособности физических лиц главным образом используют подход кредитного скоринга. Кредитный скоринг представляет собой систему, основанную на математических и статистических методах, которая, используя кредитную историю банка, прогнозирует вероятность того, что потенциальный заемщик вовремя вернет кредит. Скоринг оценивает не только вероятность возврата кредита, но и обязательность и надежность клиента.

Актуальность темы обусловлена необходимостью прогнозирования данной-вероятности и надежности клиента с точки зрения его платежеспособности. Исследования и прогнозирование платежеспособности будут произведены на примере данных банка Home Credit. Благодаря предоставленным данным, может быть построена более разумная система кредитного скоринга. Все это позволит снизить риски банков и их деятельности, а также увеличить прибыль.

Основной целью работы является – снижение расходов и увеличение прибыли банков и коммерческих организаций, которые занимаются выдачей кредитов населению.

Чтобы достичь поставленную цель, необходимо решить следующие задачи:

– проанализировать процесс кредитования клиентов;

– проанализировать методы классификации прогнозирования платежеспособности потенциальных заемщиков;

– объяснить выбор использованных метрик качества;

– проанализировать предоставленный набор данных;

– провести предварительную обработку данных;

– исследовать эффективность прогнозирования платежеспособности;

– разработать коммерциализацию проекта.

Научной новизной является использование метода градиентного бустинга для прогнозирования сбоев технологических линий.

Практическая значимость – использование данного подхода позволяет снизить расходы и увеличить прибыль банков и коммерческих организаций, занимающихся выдачей кредитов населению.

Апробации работы:

1. Лайко С.А. WEB-ресурс как способ продвижения предприятия / С.А. Лайко, А.А. Тютёва // Научные исследования: теория, методика и практика: материалы III Междунар. науч.-практ. конф. (Чебоксары, 19 нояб. 2017 г.). В 2 т. Т. 2 / редкол.: О.Н. Широков [и др.] – Чебоксары: ЦНС «Интерактив плюс», 2017. – С. 258-260. – ISBN 978-5-6040208-7-6.

2. Тютёва А.А. Оболочка для создания компьютерных тестов как способ оценки уровня знаний / А.А. Тютёва, С.А. Лайко // Образование и наука в современных реалиях: материалы IV Междунар. науч.–практ. конф. (Чебоксары, 26 февр. 2018 г.) / редкол.: О.Н. Широков [и др.] – Чебоксары: ЦНС «Интерактив плюс», 2018. – С. 209-210. – ISBN 978-5-6040732-7-8.

3. Тютёва А.А. Электронные сервисы в школе: социально-техническая эффективность / А.А. Тютёва, А.А. Лесняк // Роль технических наук в развитии общества: сборник материалов Международной научно-практической конференции (Кемерово, 26-27 ноября 2015г.) / редкол.: А.Г. Пимонов [и др.] – С. 43-51. – ISBN 978-5-906805-29-4.

1 ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ КРЕДИТОВАНИЯ КОРПОРАТИВНЫХ КЛИЕНТОВ

1.1 Сущность и функции кредита

Кредит выступает в качестве экономической категории и представляет собой экономические отношения, связанные с формой движения денежного капитала на условиях возвратности и с уплатой процентов.

Порядок и условия кредитования в России юридически закреплены в ч.2 ГК РФ и регулируются гл. 42 "Заем и кредит".

Согласно ст. 807 ГК РФ заем – это отношения в виде договора, по которому одна сторона (заимодавец) передает или обязуется передать в собственность другой стороне (заемщику) наличные деньги или безналичные денежные средства либо определенные родовыми признаками вещи, документарные или бездокументарные ценные бумаги, а заемщик обязуется возвратить заимодавцу такую же сумму наличных денег или безналичных денежных средств (сумму займа) или равное количество полученных им вещей того же рода и качества либо ценных бумаг того же рода. Договор займа между гражданами должен быть заключен в письменной форме, если его сумма превышает 10 000 руб., а в случае, когда заимодавцем является юридическое лицо, – независимо от суммы.

Согласно ст. 819 ГК РФ кредит – это отношения в виде договора, по которому банк или иная кредитная организация (кредитор) обязуются предоставить денежные средства (кредит) заемщику в размере и на условиях, предусмотренных договором, а заемщик обязуется возвратить полученную денежную сумму и уплатить проценты на нее. Кредитный договор должен быть заключен только в письменной форме, при этом несоблюдение письменной формы влечет недействительность кредитного договора.

Согласно ст. 689 ГК РФ ссуда – это отношения в виде договора, по которому одна сторона (ссудодатель) обязуется передать или передаст вещь в безвозмездное временное пользование другой стороне (ссудополучателю), а последняя обязуется вернуть ту же вещь в том состоянии, в каком она се получила, с учетом

нормального износа или в состоянии, обусловленном договором. Право передачи вещи в безвозмездное пользование принадлежит ее собственнику, но при этом коммерческая организация не вправе передавать имущество в безвозмездное пользование лицу, являющемуся ее учредителем, участником, руководителем, членом ее органов управления или контроля.

Таким образом, при кредите заимодавцем выступает банк или кредитная организация, а при займе – физические и юридические лица. Ссуда предполагает безвозмездное временное пользование, кредит и заем выдаются на основе возвратности и платности.

Кредитные отношения, возникающие между субъектами экономических отношений, позволяют использовать кредит как финансовый инструмент по перераспределению свободных денежных средств на взаимных условиях. Кредитование представляет собой форму финансового обеспечения хозяйствующего субъекта путем покрытия его расходов за счет привлечения банковских и других форм кредитов. Другими словами, можно говорить о таком экономическом понятии как кредитование, которое представляет собой процесс привлечения кредитных ресурсов для обеспечения бесперебойного кругооборота денежных средств, необходимого для стабильного и эффективного функционирования деятельности экономического субъекта (граждане, предприятия, организации, государство).

Объектом в кредитных отношениях выступает ссужаемая стоимость, которая может предоставляться в денежной, товарной и смешанной формах.

В товарной форме кредит предназначен для передачи во временное пользование стоимости в виде какой-либо вещи. В денежной форме кредит означает, что предоставление и погашение его происходит в денежно-наличной или безналичной форме. Смешанная форма предполагает предоставление кредита в товарной, а погашение в денежной форме или наоборот.

Кредит возникает при передаче свободных денежных средств во временное пользование от одного субъекта другому, который испытывает недостаток в

этих средствах. В процессе кредитных отношений участвуют два субъекта: кредитор и заемщик. Кредитор – это участник кредитных отношений, предоставляющий ссуду на условиях возвратности и с уплатой процентов за ее использование. Заемщик – участник кредитных отношений, получающий ссуженную стоимость и принимающий на себя обязательство возвратить ее в установленный срок и уплатить процент за временное пользование. Заемщик, у которого возникают обязательства по возврату ссуженной стоимости при взятии данного кредита, становится должником и дебитором по отношению к своему кредитору.

Сущность кредита проявляется в перераспределении и аккумуляции временно свободных денежных средств и отражает степень развития кредитных отношений в рыночной экономике, при котором происходит восполнение временного недостатка собственных оборотных денежных средств.

Направленность и содержание действия кредита отражены в следующих функциях.

1. Перераспределительная функция – заключается в передаче временно свободных финансовых ресурсов из одних сфер хозяйственной деятельности в другие, для обеспечения более высокой прибыли. При этом происходит перераспределение временно высвободившейся стоимости на условии возврата. Активную роль в перераспределительном процессе играют банки, которые выступают в качестве посредников, размещая привлеченные денежные средства (за счет открытия депозитных вкладов) от своего имени в качестве кредита на условиях возвратности и платности.

При осуществлении государственного кредитования важной задачей для государства являются определение экономических приоритетов и привлечение финансовых ресурсов в те отрасли или регионы, ускоренное развитие которых объективно необходимо с позиции национальных интересов.

2. Функция замещения – это замещение наличных денег кредитными знаками обращения в виде векселей, чеков, кредитных карточек и безналичных расчетов. Использование безналичных расчетов по денежным обязательствам и зачет вза-

имной задолженности значительно сокращают налично-денежный оборот и уменьшают издержки обращения связанные с изготовлением, пересчетом, транспортировкой и охраной наличных денежных средств.

3. Функция аккумуляции и концентрации денежного капитала – это быстрое накапливание капитала, использование которого необходимо для какой-либо цели.

Источниками капитала могут быть собственные средства (в виде прибыли) и заемные средства – в виде кредита. Получение прибыли и амортизационных отчислений требует длительного времени, в то время как получение кредита в банке возможно в течение нескольких дней или часов. Поэтому денежные средства, взятые в кредит, значительно сокращают время в процессе накопления капитала, необходимого для какой-либо цели.

Рассматривая функции кредита, можно также выделить дополнительные производные функции, которые выступают в качестве роли при осуществлении различных экономических отношений. Роль кредита в экономической системе заключается в следующем:

1) кредит позволяет получать во временное пользование денежные средства. Зачастую в процессе кредитования происходит восполнение временного недостатка собственных оборотных средств;

2) кредит обслуживает товарооборот, т.е. активно воздействует на товарное и денежное обращение. Вводя в сферу денежного обращения такие инструменты, как векселя, чеки, кредитные карточки и т.д., он обеспечивает замену наличных расчетов безналичными операциями, что упрощает и ускоряет механизм экономических отношений на внутреннем и международном рынках;

3) кредит способствует ускорению научно-технического прогресса, так как привлечение финансовых ресурсов необходимо для бесперебойного функционирования научных центров;

4) кредит экономит издержки обращения, например снижаются затраты продавца на хранение и продажу товаров. Чем быстрее скорость обращения товаров, тем меньше издержки обращения;

5) кредит приносит определенный доход кредитору от передачи им во временное пользование заемщику своих денежных средств. Если заемщик намерен досрочно погасить всю сумму своего основного долга, то кредитор может потребовать уплаты причитающихся по этому кредиту части процентов, в качестве неустойки в виде штрафа;

6) кредит способствует рациональному использованию полученных ресурсов в силу их платности;

7) на базе кредитных отношений осуществляется контроль за эффективностью деятельности экономических субъектов, который строится на основе наблюдения за деятельностью заемщиков и кредиторов, при этом оцениваются кредитоспособность и платежеспособность субъектов. Кредитор, используя свои методы, старается контролировать финансовое состояние заемщика, стремясь обеспечить своевременный возврат основного долга и начисленных процентов.

1.2 Повышение эффективности процесса кредитования

Повышение доходности кредитного портфеля банка напрямую зависит от грамотного управления кредитными рисками. Именно скоринговые системы позволяют снизить риски без потери доходности, предложив ответ на ключевые вопросы: насколько проблематичной будет работа банка с конкретным заемщиком, какое значение кредитного лимита установить, и вернет клиент кредит или нет.

Скоринг является методом классификации совокупности заемщиков на различные группы. На практике, в зависимости от задач анализа заемщика, кредитный скоринг включает:

Application-скоринг – оценка кредитоспособности заемщиков для получения кредита. Переводит в количественную плоскость риски банка, которые связаны с правильной оценкой социальных, демографических, финансовых и других дан-

ных заемщика для принятия решения о выдаче кредита. При принятии решения о выдаче кредита быстрый анализ заемщика при помощи скоринговых моделей позволяет получить наиболее объективную оценку на основании не субъективных мнений, а аналитически проверенных закономерностей.

Behavioral Scoring – поведенческий скоринг, принятие банком решений в рамках "управления" отдельными кредитными счетами заемщиков и кредитным портфелем в целом. Основная задача поведенческого скоринга – это прогнозирование потенциальных рисков, связанных с заемщиками, которые составляют кредитный портфель. Риски, связанные с обслуживанием кредитов, бывают разные, поэтому скоринговые модели для поведенческого скоринга используют различные критерии оценки и ранжирования заемщиков. Основные из них это: оценка риска неплатежеспособности, риска дефолта (преждевременного закрытия счета), а также скоринг доходности клиентов.

Collection Scoring – определение приоритетных дел и направлений работы в отношении "плохих" заемщиков, состояние кредитного счета которых классифицировано как "неудовлетворительное". Эффективная система своевременного предупреждения просрочек является очень важной для снижения затрат банка в рамках работы по взысканию задолженности и работы с залоговым имуществом. Collection-скоринг способен улучшить эффективность работы банка на всех этапах процесса управления взаимоотношениями с должниками.

Fraud Scoring – это методология и процессы по выявлению и предотвращению мошеннических действий со стороны потенциальных и уже существующих клиентов-заемщиков. Скоринг по выявлению попыток мошенничества помогает принимать незамедлительные решения по определению тех заемщиков, чьи обращения по выдаче кредита должны быть отклонены либо отложены для более детального рассмотрения.

Известные сегодня разработки SAS, KXEN, Experian, SPSS, EGAR – это не специализированные программные средства для скоринга, а универсальные аналитические инструменты (Data Mining), так называемое «интеллектуальное яд-

ро», которое можно, в том числе использовать и для построения собственных скоринговых моделей. В полном понимании, скоринговая система изнутри представляет собой сложную систему автоматизации выдачи потребительских кредитов в банковских отделениях, торговых точках, через интернет, которая в качестве аналитического ядра использует решение одной из известных компаний-разработчиков.

Скоринг – это не только работа с определенными скоринговыми моделями, но и построение скоринговой инфраструктуры. Во многих Data mining (глубинных анализируемых данных) результат анализа статистических данных можно сохранить в виде программного кода, а его вставить в банковское программное обеспечение. Под скоринговой системой подразумевают специальное программное обеспечение, с помощью которого можно рассчитать необходимый показатель на основе исходных данных.

Скоринг представляет собой математическую или статистическую модель, с помощью которой на основе кредитной истории "прошлых" клиентов банк пытается определить, насколько велика вероятность, что конкретный потенциальный заемщик вернет кредит в срок.

Скоринг является методом классификации всей интересующей популяции на различные группы, когда нам неизвестна характеристика, которая разделяет эти группы (вернет клиент кредит или нет).

Применение кредитного скоринга позволит:

– увеличить кредитный портфель за счет уменьшения количества необоснованных отказов по кредитным заявкам;

– повысить точность оценки заемщика;

– уменьшить уровень невозвратов кредита;

– ускорить процедуру оценки заемщика;

– создать централизованное накопление данных о заемщиках;

– снизить формируемые резервы на возможные потери по кредитным обязательствам;

– быстро и качественно оценить динамику изменений кредитного счета индивидуального заемщика и кредитного портфеля в целом.

В скоринге кредитоспособность заемщика предсказывается созданной моделью. Она основана на характеристиках клиента, желающего получить кредит, и оценивает риск путем предсказания манеры погашения долга заемщиком. Скоринг позволяет выявить дополнительные факторы, которые влияют на кредитоспособность заемщика, установив взаимосвязь между событиями кредитной истории и различными его характеристиками.

1.3 Обзор работ

Данная тематика по прогнозированию платежеспособности клиентов банка была реализована исследователями, в следующих работах:

1) Скиба Сергей Александрович. 2012. Современный подход к оценке платежеспособности клиента при кредитовании;
2) Кочеткова В.В., Ефремова К.Д.. 2017. Обзор методов кредитного скоринга;
3) Лукашевич Никита Сергеевич. 2011. Управление кредитными заявками на основе автоматизированной системы кредитного скоринга.

В представленных работах были предложены различные методы, а в частности такие, как:

– k Ближайших Соседей;
– Случайный лес;
– Классификатор экстра-деревьев;
– Градиентный бустинг.

Данные методы будут рассмотрены далее.

1.4 Постановка задачи

Скоринговые системы позволяют снизить риски без потери доходности, предложив ответ на ключевые вопросы: насколько проблематичной будет работа банка с конкретным заемщиком, какое значение кредитного лимита установить, и вернет клиент кредит или нет. Повышение доходности кредитного портфеля

банка напрямую зависит от грамотного управления кредитными рисками. Именно скоринговые системы позволяют снизить риски банков без потери доходности и поэтому не маловажно использовать методы машинного обучения, с целью оценки платежеспособности клиентов банка.

Выводы по главе 1

Рассмотрено понятие и функции кредита. Он выступает в качестве экономической категории и представляет собой экономические отношения, связанные с формой движения денежного капитала на условиях возвратности и с уплатой процентов.

Были рассмотрены преимущества кредитного скоринга в условиях современного мира, его влияние на повышение эффективности кредитования клиентов банков. признаки отказов производственных линий и их последствия.

Исходя из исследования предметной области, следует сделать вывод, что для решения задачи прогнозирования платежеспособности клиентов банков эффективно использовать методы машинного обучения.

2 МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ КРЕДИТНОГО СКОРИНГА

В машинном обучении задачи обычно делятся на широкие категории. Эти категории основаны на том, как обучение получено или как обратная связь по обучению предоставляется разработанной системе.

Двумя наиболее широко распространенными методами машинного обучения являются контролируемое обучение, которое обучает алгоритмы на основе примерных входных и выходных данных, помеченных людьми, и неконтролируемое обучение, которое обеспечивает алгоритм без помеченных данных, чтобы позволить ему находить структуру в своих входных данных. Давайте рассмотрим эти методы более подробно.

Кредитный скоринг может быть определен как технология, позволяющая кредитной организации решить вопрос о предоставлении кредита заявителю с учетом его характеристик, таких как возраст, доход, семейное положение, образование, должность. Естественно, подобные технологии возникли вместе с появлением торговли и потребностью в кредитовании. Идеи и методы скоринга, соответствующие их современному пониманию, были впервые сформулированы в работе Д. Дюрана [29].

После принятия соглашений Базель II(и особенно Базель III) стало возможным и необходимым применять процедуры внутреннего рейтинга для оценки общих параметров риска. Это сделало более значительной роль кредитного скоринга и заставило финансовые институты постоянно совершенствовать используемые ими количественные модели.

2.1 Основные методы машинного обучения в кредитном скоринге

2.1.1 Линейная регрессия

Линейная регрессия связывает характеристики заемщика, представленные вектором x с целевой переменной y.

у рассчитывается по формуле 1.

$$y = \beta_0 + \langle \beta, x \rangle + \varepsilon, \quad (1)$$

где ε – случайная ошибка с нулевым средним.

В последние годы линейная регрессия в чистом виде не используется, хотя по-прежнему служит важным инструментом в смешанных моделях.

2.1.2 Байесовские сети

Отправной точкой для применения байесовских сетей в кредитном скоринге послужила работа Н. Фридмана с соавторами [30]. В этом исследовании приведено обобщение так называемого простого (наивного) байесовского метода, в соответствии с которым выбирается решение с наибольшей апостериорной информацией. Применение наивного байесовского метода обосновано в случае, когда атрибуты независимы. В кредитном скоринге это предположение нереалистично: например, нельзя игнорировать взаимосвязь таких показателей, как возраст, образование, доход. В самом общем виде байесовская сеть представляет собой ацикличный ориентированный граф. При обучении формируются условные распределения вероятности.

Байесовская сеть определяет совместное распределение вершин. Например, наивный Байесовский метод получается, если взять категориальную переменную в качестве корневой вершины, а все атрибуты – в качестве ее «детей». Неформально обучение байесовской сети состоит в ее максимальной адаптации к обучающему набору данных.

Оптимизация проводится относительно скоринговой функции. Наиболее употребительными являются байесовская скоринговая функция и функция, основанная на принципе минимальной длины описания (MDL). Эти функции асимптотически приводят к одинаковому результату обучения.

2.1.3 Нейронные сети

Нейронная сеть преобразует набор входных переменных в набор выходных переменных и моделирует как линейные преобразования, так и нелинейные. Преобразования осуществляются с помощью нейронов, представляющих собой упрощенную модель нейронов головного мозга. Нейроны связаны в сеть односторонними каналами передачи информации. Каждый нейрон может быть активирован поступающими входными сигналами, и в активном состоянии выдает выходные сигналы.

В нейронной сети имеется слой входных нейронов – это те нейроны, на которые поступают значения входных переменных, слой выходных нейронов – из выходных сигналов этих нейронов формируются выходные переменные, и скрытые слои. Нейронные сети различаются своей структурой, числом скрытых слоев, функциями активации.

В работе Д. Веста [34] проанализированы пять моделей нейронных сетей, используемых в кредитном скоринге:

1) многослойный персептрон (MLP);
2) смесь экспертов (MOE);
3) сеть радиальных базисных функций (RBF);
4) квантование обучающего вектора (LVQ);
5) нечеткий адаптивный резонанс (FAR).

Эффективность применения нейронных сетей перечисленных типов в кредитном скоринге сравнивалась с эффективностью применения классических параметрических методов (линейный дискриминантный анализ и логистическая регрессия), непараметрических методов (k ближайших соседей или k-NN, ядерной оценке плотности) и классификационных деревьев. Полученные результаты подтвердили, что многослойные персептроны показывают далеко не самую высокую точность, сети типа смеси экспертов и сети радиальных базисных функций показывают в кредитном скоринге вполне удовлетворительный результат.

Сети, основанные на нечетком адаптивном резонансе, оказались наименее точными. Не уступающие другим сетям по распознаванию «плохих» заемщиков, они существенно хуже работают по распознаванию «хороших» заемщиков.

2.1.4 Комбинированные методы

К числу гибридных и комбинированных относятся методы, в которых применяются различные техники кредитного скоринга для повышения эффективности. Наиболее употребительны три метода комбинирования (ensemble methods): беггинг (bagging – bootstrap aggregating), бустинг (boosting) и стекинг (stacking).

Беггинг был введен в работе Л. Бреймана [28]. Основная идея метода – построение набора предикторов, которые в совокупности (после определенного агрегирования) дают более совершенный предиктор.

Применение беггинга оказывается особенно эффективным в тех случаях, когда основной алгоритм обучения неустойчив – сильно зависит от небольших изменений в обучающем множестве.

Основная идея бустинга – сформировать на основе слабого (в смысле точности) алгоритма сильный алгоритм классификации. В процессе формирования сильного алгоритма слабый алгоритм «доучивается» за счет того, что перераспределяются веса примеров из обучающей выборки: в случае верного распознавания вес снижается.

При стекинге происходит комбинирование нескольких алгоритмов с помощью некоторого комбинатора. Как правило, в роли комбинатора выступает логистическая регрессия [35].

1. Тестовые наборы данных

Широкое распространение среди разработчиков алгоритмов кредитного скоринга получили два набора данных с условными названиями австралийский (Australian scoring data) и немецкий (German Credit Data Set). Австралийский набор содержит в общей сложности данные о 690 заемщиках, из которых 307 состоятельны (выплачивают кредит), а 383 несостоятельны. Описание каждого

заемщика включает 14 атрибутов (6 непрерывных и 8 категориальных). Немецкий набор содержит 1 000 записей о заемщиках, из которых 700 состоятельны, 300 несостоятельны. Описание заемщика содержит 20 атрибутов.

2. Оценка качества алгоритмов кредитного скоринга

Одним из способов определения качества модели машинного обучения является разделение выборки на обучающую, которая используется для идентификации параметров алгоритма, и контрольную, для каждого объекта которой проводится сравнение класса, предсказанного алгоритмом, и истинного класса объекта.

При этом наиболее распространенные методы оценки алгоритмов кредитного скоринга основываются на матрице ошибок: все объекты контрольной выборки разбивают на четыре категории в зависимости от комбинации истинного ответа y:

Поскольку целью применения алгоритмов классификации в кредитном скоринге является сортировка объектов скоринга на хорошие и плохие, эффективность алгоритмов оценивается путем сопоставления для каждого объекта из контрольного набора данных класса, спрогнозированного алгоритмом, с реальным классом этого объекта.

Задача кредитного скоринга имеет две особенности. Во-первых, классификация плохого кредита как хорошего обходится дороже, чем классификация хорошего кредита как плохого, а во-вторых, в обучающей выборке хороших клиентов всегда больше чем плохих.

В связи с первой особенностью в задаче кредитного скоринга применяются следующие метрики качества алгоритмов.

Доля правильно классифицированных кредитов рассчитывается по формуле 2.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (2)$$

где TP – сокращение для True Positive;

TN– сокращение для True Negative;

FP– сокращение для False Positive;

FN– сокращение для False Negative.

Доля правильно классифицированных плохих кредитов среди всех наблюдений, отнесенных к плохим кредитам (точность) рассчитывается по формуле 3.

$$Precision = \frac{TP}{TP+FP}, \quad (3)$$

где TP – сокращение для True Positive;

FP– сокращение для False Positive.

Оценка распознавать плохие кредиты (полнота) рассчитывается по формуле 4.

$$Recall = \frac{TP}{TP+FN}, \quad (4)$$

где TP – сокращение для True Positive;

FN– сокращение для False Negative.

Доля правильно классифицированных хороших кредитов среди всех наблюдений, отнесенных к хорошим кредитам рассчитывается по формуле 5.

$$Negative\ Predictive\ Value = \frac{TN}{TN+FN}, \quad (5)$$

где TN– сокращение для True Negative;

FN– сокращение для False Negative.

Оценка способности алгоритма распознавать хорошие кредиты (специфичность) рассчитывается по формуле 6.

$$Specificity = \frac{TN}{TN+FP}, \quad (6)$$

где TN– сокращение для True Negative;

FN– сокращение для False Negative.

Доля плохих кредитов, неправильно, отнесенных к хорошим рассчитывается по формуле 7.

$$False\ Negative\ Rate = \frac{FN}{TP+FN}, \quad (7)$$

где TP – сокращение для True Positive;

FN– сокращение для False Negative.

Доля хороших кредитов, неправильно отнесенных к плохим рассчитывается по формуле 8.

$$False\ Positive\ Rate = \frac{FP}{TN+FP}, \quad (8)$$

где TN– сокращение для True Negative;

FP– сокращение для False Positive.

Программная реализация алгоритмов машинного обучения в кредитном скоринге

Программные продукты, используемые для автоматизации решения задач интеллектуального анализа данных и машинного обучения, можно разделить на три класса:

– коммерческие статистические пакеты;

– открытые среды;

– облачные решения.

Исторически в банках для решения задач анализа данных, в частности связанных со скорингом, использовались коммерческие программные продукты. Наиболее часто применялся набор продуктов SAS, реже пакеты IBM SPSS и Statistica. Эти три линейки продуктов предоставляют схожие функциональные возможности, включающие в себя средства аналитической подготовки данных, готовые и настраиваемые шаблоны алгоритмов машинного обучения, в том числе моделей линейной и логистической регрессии, деревьев и лесов решений, градиентного бустинга, опорных векторов, нейронных сетей. Кроме того, в этих пакетах возможна настройка параметров моделей и использование интерактивных техник оценки качества.

В последние годы банки, не отказываясь полностью от применения коммерческих пакетов типа SAS, стали использовать и открытые среды Python/R/Spark. Преимуществом этих сред является прежде всего возможность использования гораздо большего количества алгоритмов, чем в коммерческих пакетах.

Язык R создавался как специальное средство для статистических вычислений, он стал первой открытой средой, которая начала активно использоваться для анализа данных. Наиболее часто используемые библиотеки для машинного обучения в R – это rpart и CARET (алгоритмы классификации и регрессии), randomForest (алгоритм случайных лесов), nnet (нейронные сети), kernlab (метод опорных векторов), gbm (градиентный бустинг), ROCR (визуализация метрик качества алгоритмов классификации).

Язык Python стал самым популярным средством для анализа данных после выхода отлично документированной библиотеки scikit-learn, в которой реализовано большое количество алгоритмов машинного обучения. Кроме scikit-learn, популярны также библиотеки TensorFlow и Theano (эти библиотеки также реализуют различные методы анализа данных, но выигрывают у scikit-learn только в количестве реализованных техник работы с нейронными сетями). Для использования аппарата детерминированного, нечеткого и байесовского логического вы-

вода в Python применяется библиотека pyinference. Основное преимущество Python перед R – более высокая скорость выполнения скриптов.

В последние годы появились облачные платформы машинного обучения. Основным преимуществом таких систем является гибкая масштабируемость – выделениеи высвобождение вычислительных ресурсов происходит мгновенно в соответствии с решаемыми задачами. Amazon Machine Learning реализует только базовые алгоритмы бинарной и мультиклассовой классификации, а также регрессии. Google Machine Learning Engine предоставляет возможность запуска моделей TensorFlow в облачной среде.

Однако несмотря на все преимущества облачных средств анализа данных, в банках они практически не используются в связи с опасениями по поводу безопасности передачи в облачные хранилища конфиденциальных данных о клиентах.

2.2 Алгоритмы машинного обучения

Как область машинное обучение тесно связано с вычислительной статистикой, поэтому наличие базовых знаний в области статистики полезно для понимания и использования алгоритмов машинного обучения.

Для тех, кто, возможно, не изучал статистику, может быть полезно сначала определить корреляцию и регрессию, поскольку они обычно используются для исследования взаимосвязи между количественными переменными. Корреляция – это мера связи между двумя переменными, которые не обозначены как зависимые или независимые. Регрессия на базовом уровне используется для изучения взаимосвязи между одной зависимой и одной независимой переменной. Поскольку статистика регрессии может использоваться для прогнозирования зависимой переменной, когда независимая переменная известна, регрессия обеспечивает возможности прогнозирования.

Алгоритмы в машинном обучении постоянно развиваются. Для наших целей мы рассмотрим несколько популярных алгоритмов, которые используются в машинном обучении для решения такого рода проблем.

2.2.1 k Nearest Neighbor

kNN расшифровывается как k Nearest Neighbor или k Ближайших Соседей – это один из самых простых алгоритмов классификации, также иногда используемый в задачах регрессии. Благодаря своей простоте, он является хорошим примером, с которого можно начать знакомство с областью Machine Learning [24].

В машинном обучении задача классификации – это задача отнесения объекта к одному из заранее определенных классов на основании его формализованных признаков. Каждый из объектов в этой задаче представляется в виде вектора в N-мерном пространстве, каждое измерение в котором представляет собой описание одного из признаков объекта. Допустим нам нужно классифицировать мониторы: измерениями в нашем пространстве параметров будут величина диагонали в дюймах, соотношение сторон, максимальное разрешение, наличие HDMI-интерфейса, стоимость. Случай классификации текстов несколько сложнее, для них обычно используется матрица термин-документ.

Для обучения классификатора необходимо иметь набор объектов, для которых заранее определены классы. Это множество называется обучающей выборкой, её разметка производится вручную, с привлечением специалистов в исследуемой области. Например, в задаче Detecting Insults in Social Commentary для заранее собранных тестов комментариев человеком проставлено мнение, является ли этот комментарий оскорблением одного из участников дискуссии, само же задание является примером бинарной классификации. В задаче классификации может быть более двух классов (многоклассовая), каждый из объектов может принадлежать более чем к одному классу (пересекающаяся).

2.2.2 Случайный лес (Random forest)

Случайный лес – один из самых потрясающих алгоритмов машинного обучения, придуманные Лео Брейманом и Адель Катлер ещё в прошлом веке. Он дошёл до нас в «первозданном виде» (никакие эвристики не смогли его существенно улучшить) и является одним из немногих универсальных алгоритмов. Универсальность заключается, во-первых, в том, что он хорош во многих задачах (по оценкам, 70% из встречающихся на практике, если не учитывать задачи с изображениями), во-вторых, в том, что есть случайные леса для решения задач классификации, регрессии, кластеризации, поиска аномалий, селекции признаков.

RF (random forest) – это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо по следующей схеме:

1. Выбирается подвыборка обучающей выборки размера samplesize (м.б. с возвращением) – по ней строится дерево (для каждого дерева – своя подвыборка).

2. Для построения каждого расщепления в дереве просматриваем max_features случайных признаков (для каждого нового расщепления – свои случайные признаки).

3. Выбираем наилучшие признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

Понятно, что такая схема построения соответствует главному принципу ансамблирования (построению алгоритма машинного обучения на базе нескольких, в данном случае решающих деревьев): базовые алгоритмы должны быть хорошими и разнообразными (поэтому каждое дерево строится на своей обучающей выборке и при выборе расщеплений есть элемент случайности).

Метод RF хорош ещё тем, что при построении леса параллельно может вычисляться т.н. oob-оценка качества алгоритма (которая очень точная и получается не в ущерб разделения на обучение/тест), oob-ответы алгоритмы (ответы, которые выдавал бы алгоритм на обучающей выборке, если бы «обучался не на ней»), оцениваются важности признаков. Также не стоит забывать про полный перебор значений параметров (если объектов в задаче не очень много)[18].

2.2.3 XGBoost

XGBoost – это контролируемый алгоритм обучения, который реализует процесс, называемый boosting, чтобы дать точные модели. Boosting относится к методу обучения ансамблю для построения многих моделей последовательно, причем каждая новая модель пытается исправить недостатки предыдущей модели. В повышении дерева каждая новая модель, добавленная в ансамбль, является деревом решений. XGBoost обеспечивает параллельное наращивание дерева (также известное как GBDT, GBM), которое быстро и точно решает многие проблемы с наукой о данных. Для многих проблем XGBoost – одна из лучших рамок ускорителя градиента (GBM) сегодня.

Возможности XGBoost – особенности модели и системные функции

Реализация модели поддерживает особенности реализации scikit-learn и R с новыми дополнениями, такими как регуляризация. Поддерживаются три основные формы повышения градиента:

1. Алгоритм Gradient Boosting также называется градиентной машиной повышения, включая скорость обучения;

2. Stochastic Gradient Boosting с суб-выборкой в строке, столбце и столбце на каждый уровень разделения;

3. Регулярное усиление градиента с регуляцией L1 и L2.

Библиотека предоставляет систему для использования в различных вычислительных средах, не в последнюю очередь:

1. Параллелизация построения дерева с использованием всех ваших ядер процессора во время обучения;

2. Распределенные вычисления для обучения очень крупных моделей с использованием кластера машин;

3. Вне корпоративного вычисления для очень больших наборов данных, которые не вписываются в память;

4. Кэш Оптимизация структуры данных и алгоритма для наилучшего использования аппаратного обеспечения.

Реализация алгоритма была разработана для эффективности вычислительных ресурсов времени и памяти. Цель проекта заключалась в том, чтобы наилучшим образом использовать имеющиеся ресурсы для обучения модели. Некоторые ключевые функции реализации алгоритма включают:

1. Редкая реализация Aware с автоматической обработкой отсутствующих значений данных;

2. Блочная структура для поддержки распараллеливания конструкции дерева;

3. Продолжение обучения, чтобы вы могли еще больше повысить уже установленную модель для новых данных.

После того как мы изучили алгоритмы машинного обучения стоит сделать выбор в пользу того алгоритма, который будет наилучшим выбором для решения нашей задачи. Несомненно, для начала следует разобраться в тех данных и задачах, которые перед нами поставлены.

Выводы по главе 2

Методы машинного обучения постоянно совершенствуются. Рассмотрены несколько популярных алгоритмов, которые используются в машинном обучении:

– kNN расшифровывается как k Nearest Neighbor или k Ближайших Соседей;

– Случайный лес – RF (random forest);

– XGBoost;

– Light GBM.

Были рассмотрены метрики качества, такие как:

– Кривая ROCAUC – это измерение производительности для задачи классификации при различных настройках порогов.

Также были рассмотрены метрики качества алгоритмов, которые возможно применять для оценки эффективности.

3 ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ РЕШЕНИЯ ЗАДАЧИ КРЕДИТНОГО СКОРИНГА НА ПРИМЕРЕ БАНКА HOME CREDIT

3.1 Описание набора данных

Обучающая выборка состоит из более чем 300000 записей, признаков достаточно много – 122, среди них много категориальных (не числовых). Признаки довольно подробно описывают заемщика. Часть данных содержатся в 6 дополнительных таблицах (данные по кредитному бюро, балансу кредитной карты и предыдущим кредитам.

Это стандартная задача классификации (1 в поле TARGET означает любые сложности с платежами, 0 – отсутствие сложностей). Однако следует предсказывать не 0/1, а вероятность возникновения проблем (что, впрочем, довольно легко решают методы предсказания вероятностей predict_proba, которые есть у всех сложных моделей).

Есть 8 таблиц с данными:

HomeCredit_columns_description.csv: Описание полей;

application_train/application_test: Основные данные, заемщик идентифицируется по полю SK_ID_CURR;

bureau: Данные по предыдущим займам в других кредитных организациях из кредитного бюро;

bureau_balance: Ежемесячные данные по предыдущим кредитам по бюро;

previous_application: Предыдущие заявки по кредитам в Home Credit, каждая имеет уникальное поле SK_ID_PREV;

POS_CASH_BALANCE: Ежемесячные данные по кредитам в Home Creditc выдачей наличными и кредитам на покупки товаров;

credit_card_balance: Ежемесячные данные по балансу кредитных карт в Home Credit;

installments_payment: Платежная история предыдущих займов в Home Credit.

На рисунке 1 показаны взаимосвязи между данными таблицами, а на рисунке 2 показан пример загрузки данных.

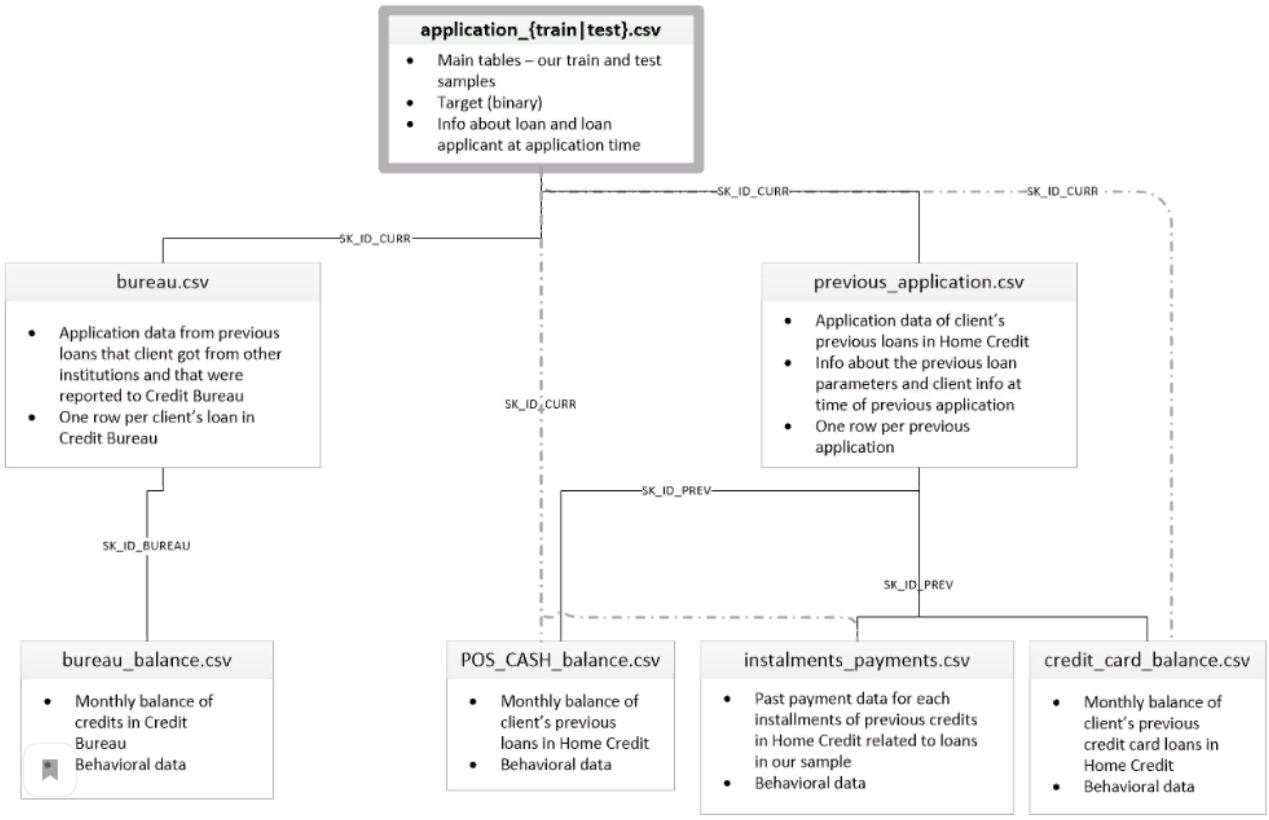

Рисунок 1 – Взаимосвязи между таблицами с исходными данными

```
In [17]: app_train = pd.read_csv(r"C:\input\application_train.csv",)

In [18]: app_test = pd.read_csv(r"C:\input\application_test.csv",)

In [20]: print ('формат обучающей выборки:', app_train.shape)
         print ("формат тестовой выборки:", app_test.shape)
```

Рисунок 2 – Загрузка основных данных

Итого у нас есть 307 тысяч записей и 122 признака в обучающей выборке и 49 тысяч записей и 121 признак в тестовой. На рисунке 3 показан небольшой пакет данных из представленных банком записей. Расхождение, очевидно, вызвано тем, что целевого признака TARGET в тестовой выборке нет, его-то мы и будем предсказывать.

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL
0	100002	1	Cash loans	M	N	Y	0	202500.0
1	100003	0	Cash loans	F	N	N	0	270000.0
2	100004	0	Revolving loans	M	Y	Y	0	67500.0
3	100006	0	Cash loans	F	N	Y	0	135000.0
4	100007	0	Cash loans	M	N	Y	0	121500.0

Рисунок 3 – Таблица представленная данных

3.2 Предварительная обработка данных

В процессе Exploratory Data Analysis (далее – EDA) считаем основные статистики и рисуем графики, чтобы найти тренды, аномалии, паттерны и связи внутри данных. Цель EDA – узнать, что могут рассказать данные. Обычно анализ идет сверху вниз – от общего обзора к исследованию отдельных зон, которые привлекают внимание и могут представлять интерес. Впоследствии эти находки можно использовать в построении модели, выборе признаков для нее и в её интерпретации. В исходных данных 67 столбцов с неполными данными, что отображено на рисунках 4 и 5.

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4

Рисунок 4 – Таблица отображения неполных данных

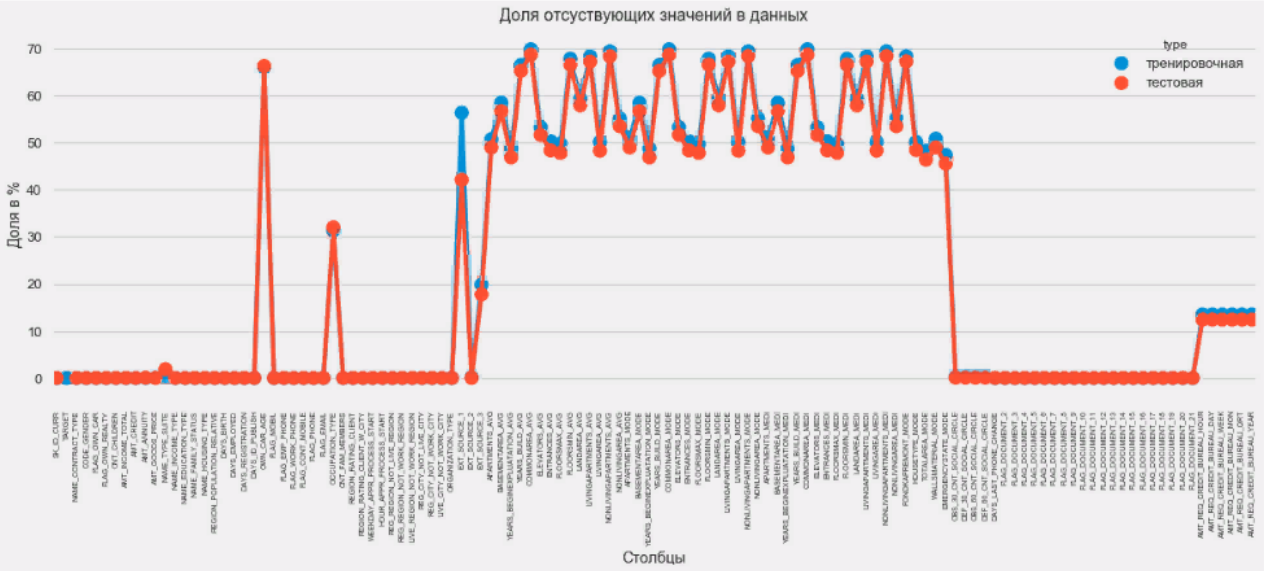

Рисунок 5 – Доля отсутствующих значений в данных

Часть столбцов имеет тип object, то есть имеет не числовое значение, а отражает какую-либо категорию. Данных столбцов – 16, в каждом из которых от 2 до 58 разных вариантов значений. При возникновении аналогичных ситуаций с данными пользуются в основном двумя подходами:

1. Label Encoding – категориям присваиваются цифры 0, 1, 2 и так далее и записываются в тот же столбец;

2. One-Hot-Encoding – один столбец раскладывается на несколько по количеству вариантов и в этих столбцах отмечается, какой вариант у данной записи.

Так как количество вариантов в столбцах выборок не равное, количество столбцов теперь не совпадает. Требуется выравнивание – необходимо убрать из тренировочной выборки столбцы, которых нет в тестовой. Это делает метод align, значение которого необходимо указать axis=1 (для столбцов) как показано на рисунке 6.

```
#сохраним лейблы, их же нет в тестовой выборке и при выравнивании они потеряются.
train_labels = app_train['TARGET']

# Выравнивание - сохранятся только столбцы. имеющиеся в обоих датафреймах
app_train, app_test = app_train.align(app_test, join = 'inner', axis = 1)

print('Формат тренировочной выборки: ', app_train.shape)
print('Формат тестовой выборки: ', app_test.shape)

# Add target back in to the data
app_train['TARGET'] = train_labels
```

```
Формат тренировочной выборки: (307511, 242)
Формат тестовой выборки: (48744, 242)
```

Рисунок 6 – Выравнивание выборок

Влияние параметров на вероятность возвратов

Возраст

Чем старше клиент, тем выше вероятность возврата (до определенного предела). Влияние возраста отображено графически на рисунке 7;

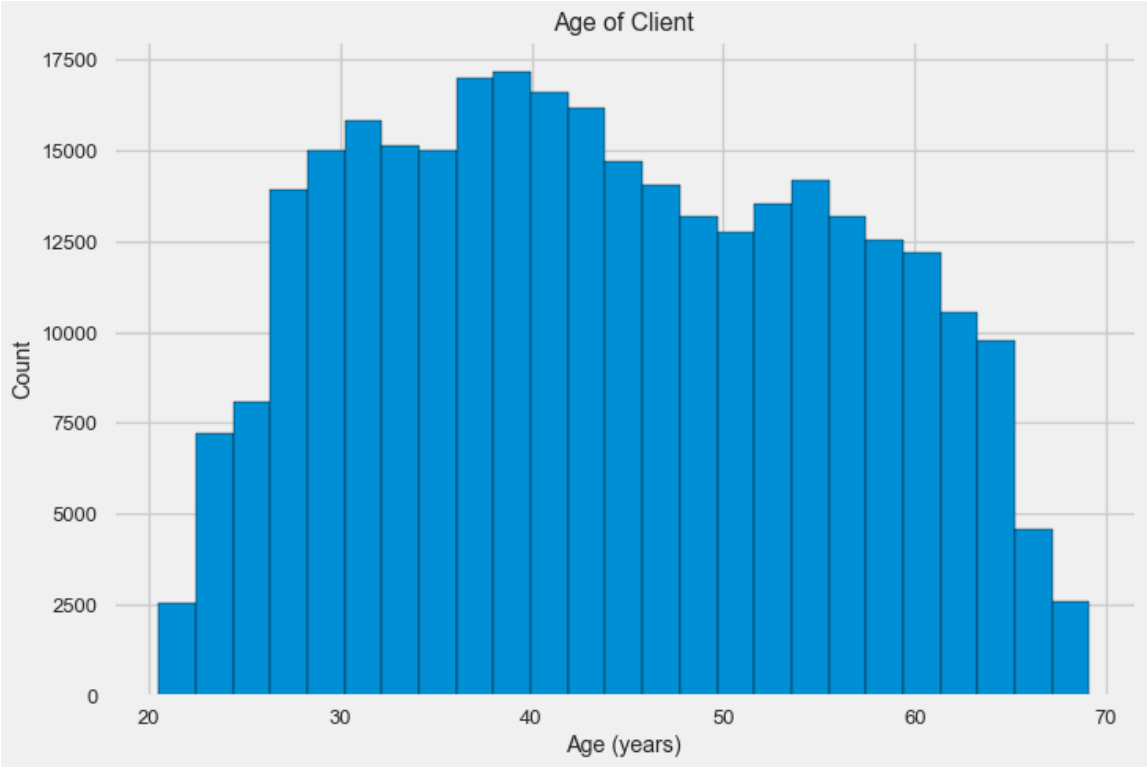

Рисунок 7 – Влияние возраста на вероятность возврата

Для наглядной демонстрации эффекта влияния возраста на результат, можно построить график Kernel density estimation (KDE) — распределение ядерной плотности, раскрашенный в цвета целевого признака. Данный график показан на рисунке 8. Он демонстрирует распределение одной переменной и может быть истолкован как сглаженная гистограмма.

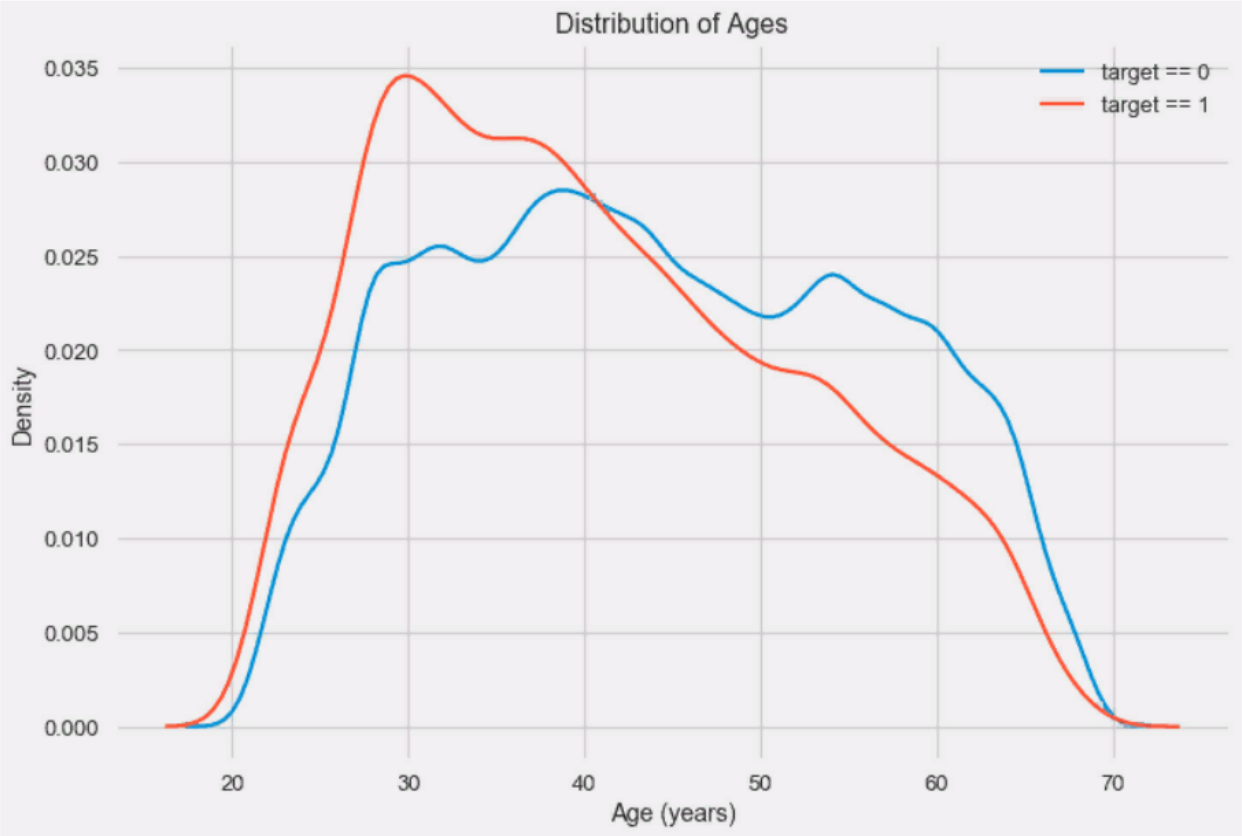

Рисунок 8 – KDE для определения влияния возраста

Доля невозвратов выше для молодых людей и снижается с ростом возраста. Это не повод отказывать молодым людям в кредите всегда, такая «рекомендация» приведет лишь к потере доходов и рынка для банка. Это повод задуматься о более тщательном отслеживании таких кредитов, оценке и, возможно, даже каком-то финансовом образовании для молодых заемщиков.

Влияние параметра «Тип займа» отображено на рисунке 9.

```
plot_stats('NAME_CONTRACT_TYPE')
```

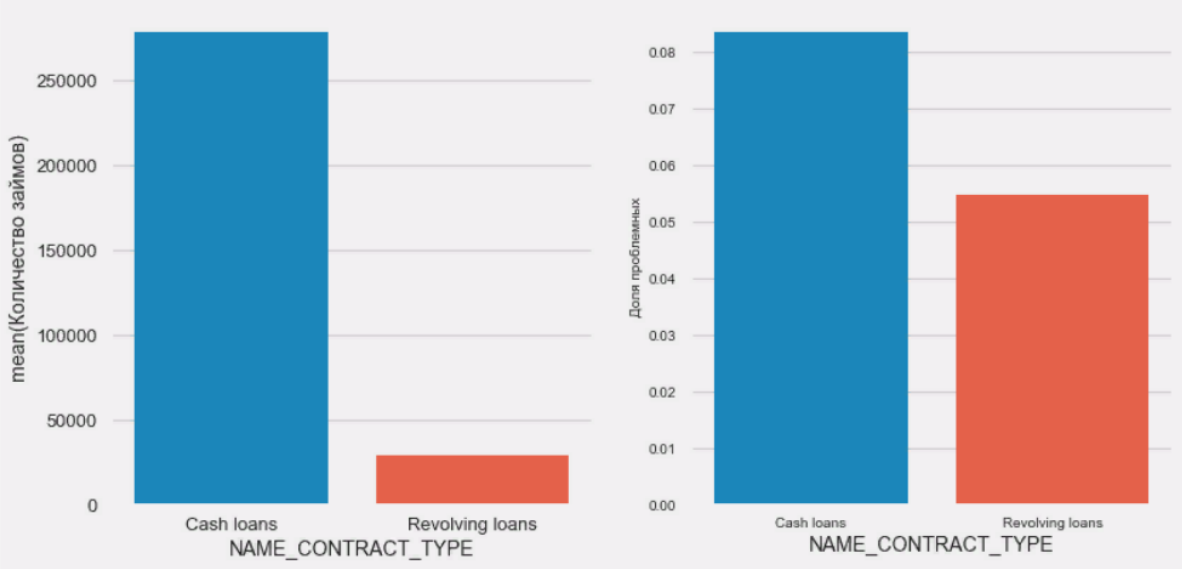

Рисунок 9 – Распределение влияния параметра «Тип займа»

Пол клиента

Женщин-клиентов почти вдвое больше мужчин, при этом мужчины показывают гораздо более высокий риск. Распределение данного параметра показывается на рисунке 10.

```
plot_stats('CODE_GENDER')
```

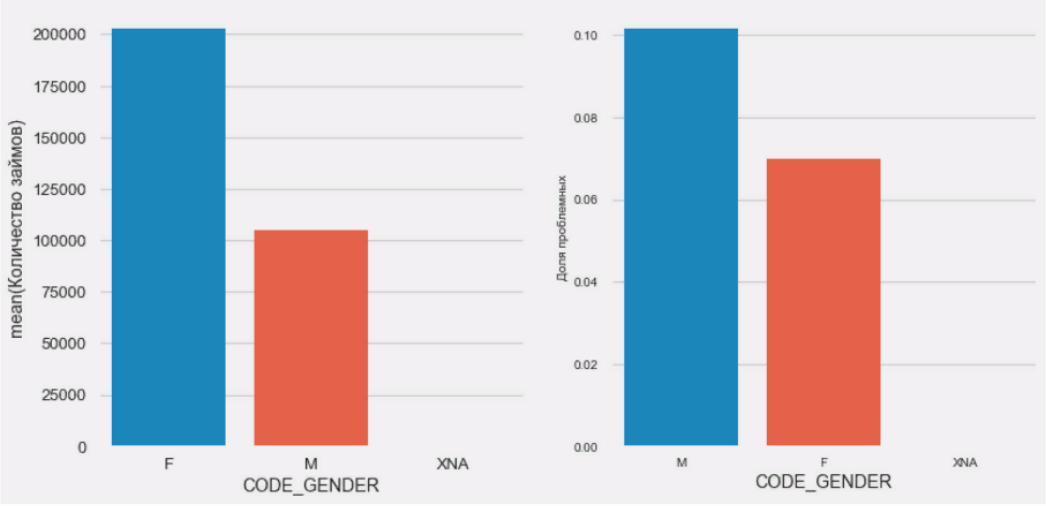

Рисунок 10 – Распределение влияния параметра «Пол клиента»

Семейный статус

В то время как большинство клиентов состоит в браке, наиболее рискованны клиенты в гражданском браке и одинокие. Вдовцы показывают минимальный риск. Данную информацию можно увидеть на рисунке 11.

```
plot_stats('NAME_FAMILY_STATUS',True, True)
```

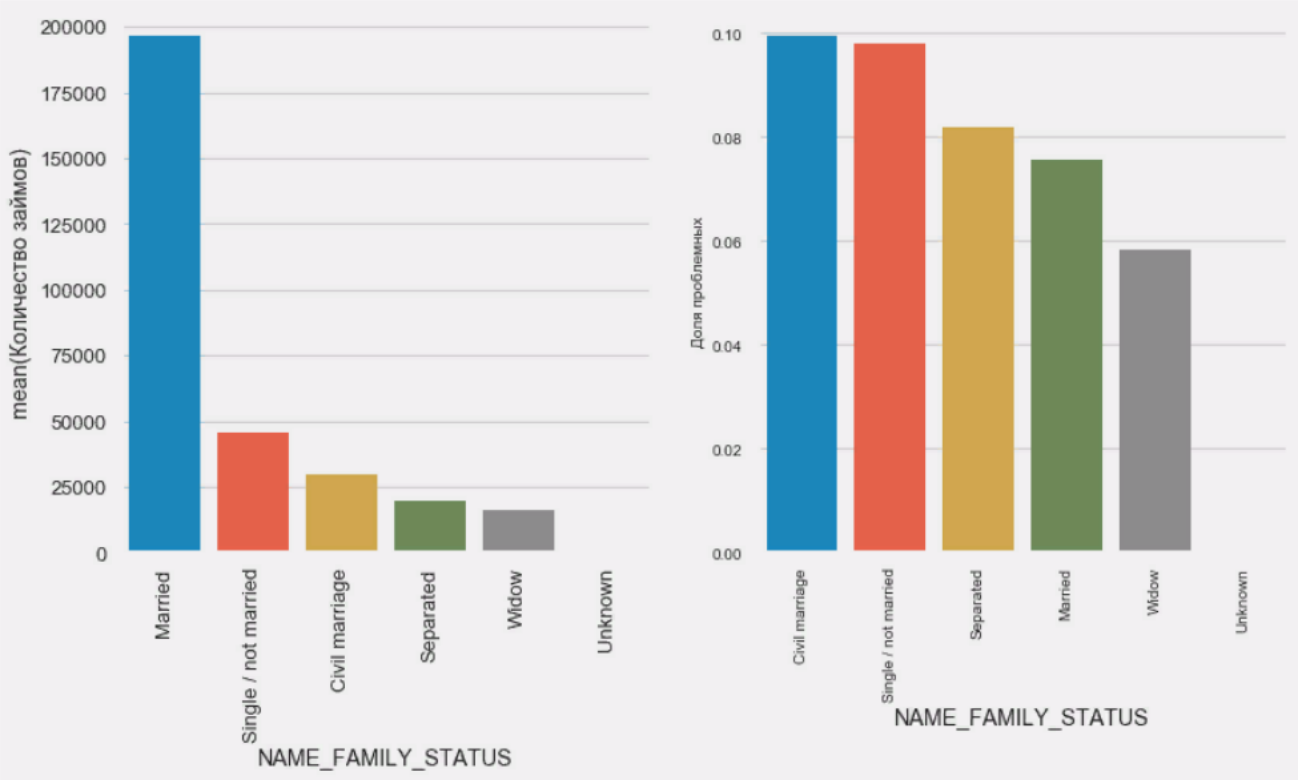

Рисунок 11 – Распределение влияния параметра «Семейный статус»

Количество детей

Большинство клиентов бездетны. При этом клиенты с 9 и 11 детьми показывают полный невозврат. Это особенно ярко видно на рисунке 12.

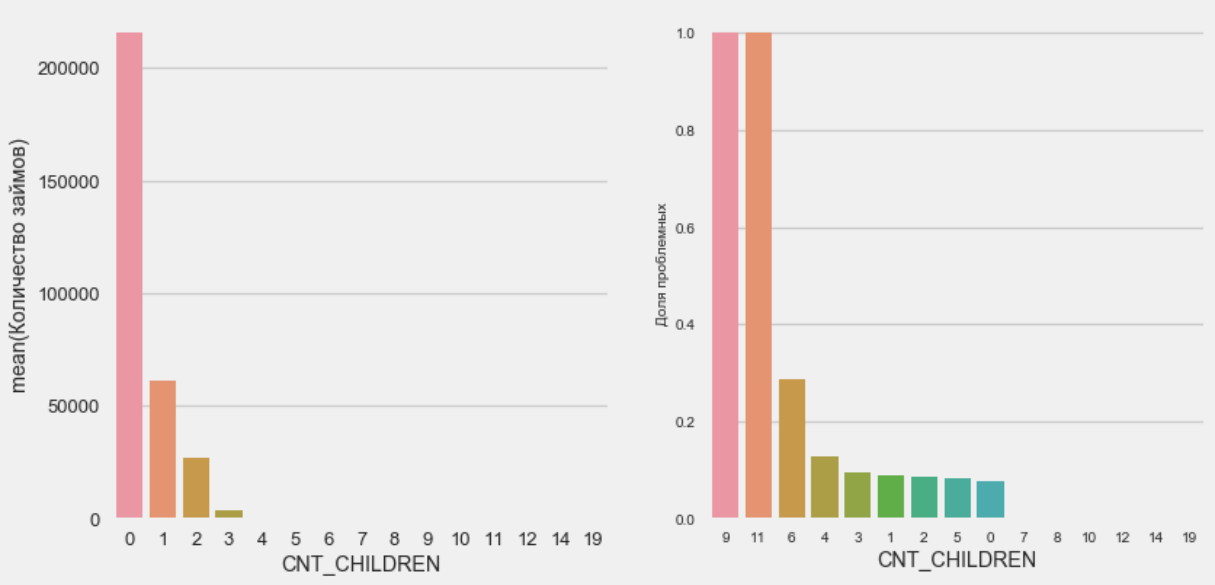

Рисунок 12 – Влияние количества детей на вероятность возврата

Влияние параметра «Тип дохода» отображено на рисунке 13.

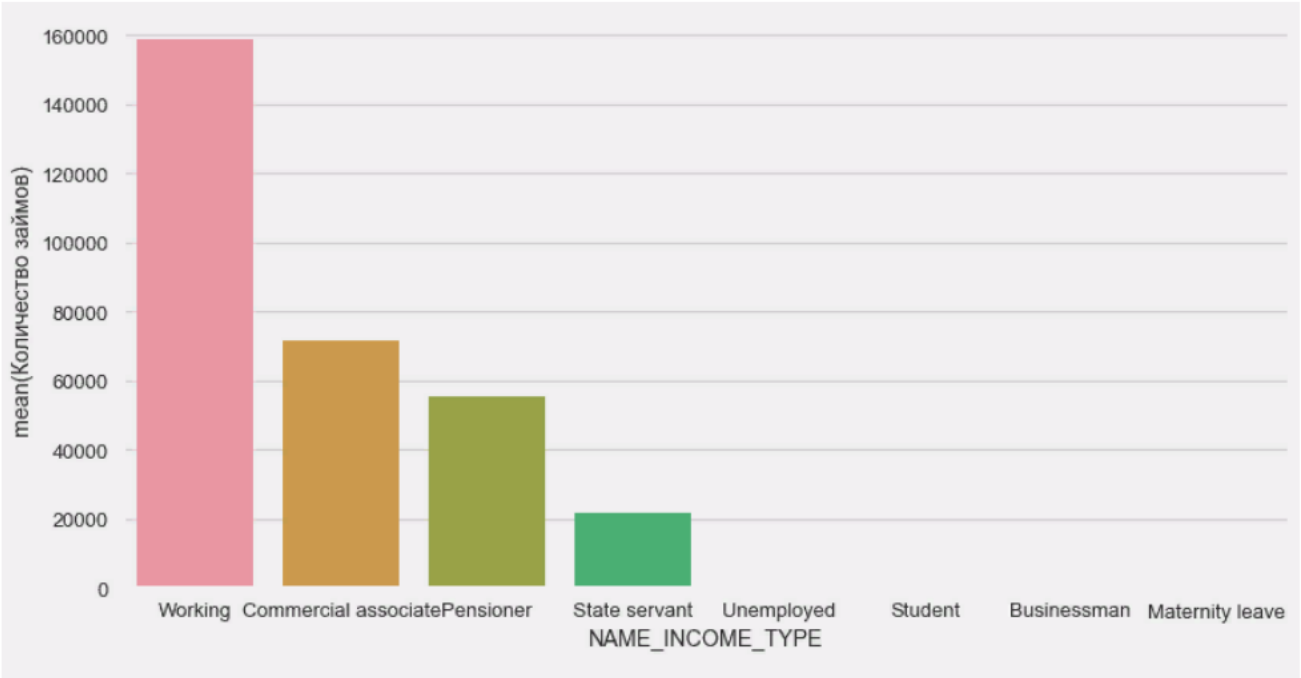

Рисунок 13 – Влияние типа дохода на вероятность возврата

Влияние параметра «Образование» продемонстрировано на рисунке 14.

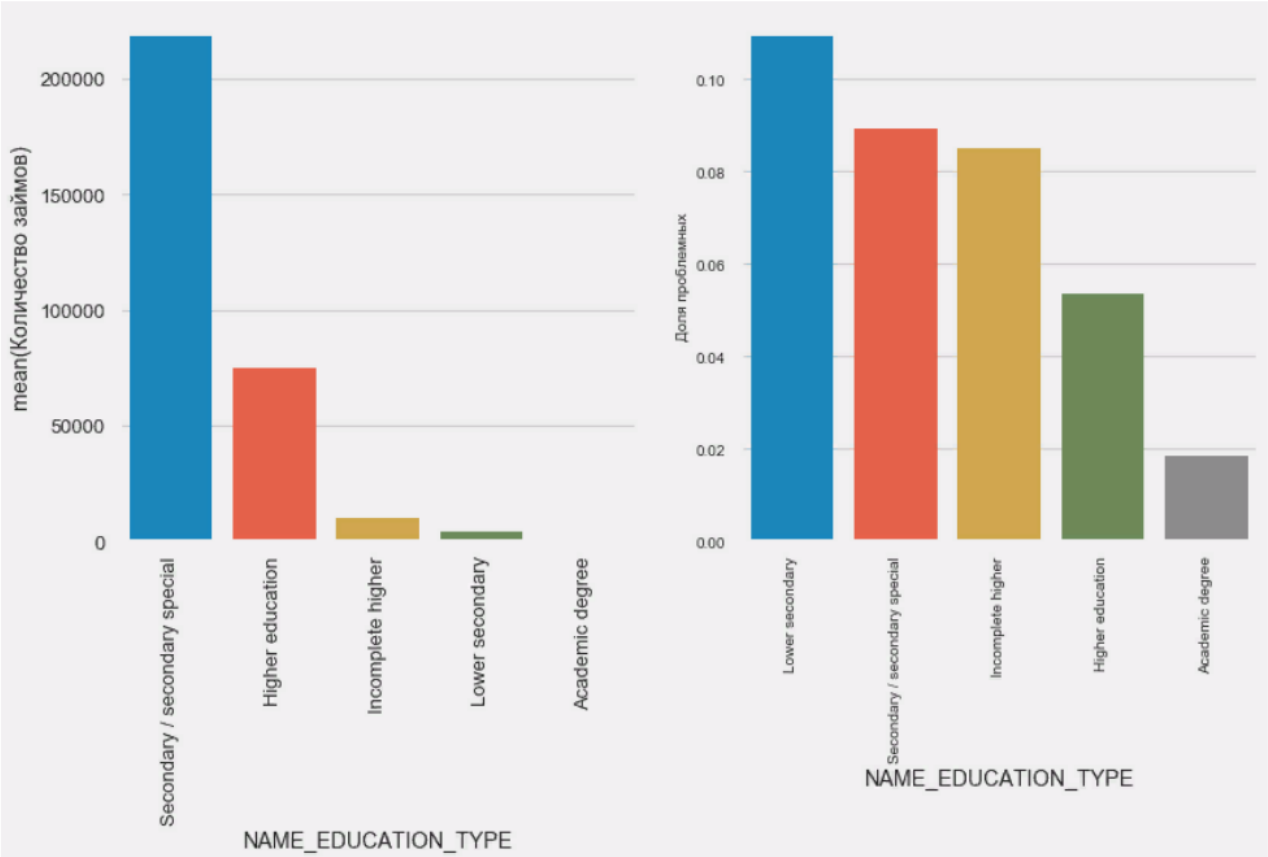

Рисунок 14 – Влияние типа образования на вероятность возврата

Распределение по плотности проживания

Клиенты из более населенных регионов склонны лучше выплачивать кредит, что видно из рисунка 15.

Рисунок 15 – Влияние суммы кредитования на вероятность возврата

3.3 Метрика качества (ROC-AUC)

Кривая ROC. В том случае, если скоринговая система на выходе выдаёт непрерывное значение счёта, точность классификации зависит не только от самой модели, но и от порогового значения счёта, начиная с которого принимается положительное решение по выдаче кредита. Для сравнения различных моделей в этом случае применяется кривая ROC (receiver operating characteristic), показывающая зависимость (1 − ER1) от ER2. Чем выше проходит такая кривая, тем точнее классификация независимо от порогового значения. Пример отображения кривой ROC продемонстрировано на рисунке 16. Применяется также численный показатель, обозначаемый AUROC или AUC (area under ROC) и равный площади фигуры между кривой ROC и прямой 1 − ER1 = ER2. Этот показатель изменяется от 0 (бесполезный классификатор) до 1/2 (абсолютно точный классификатор).

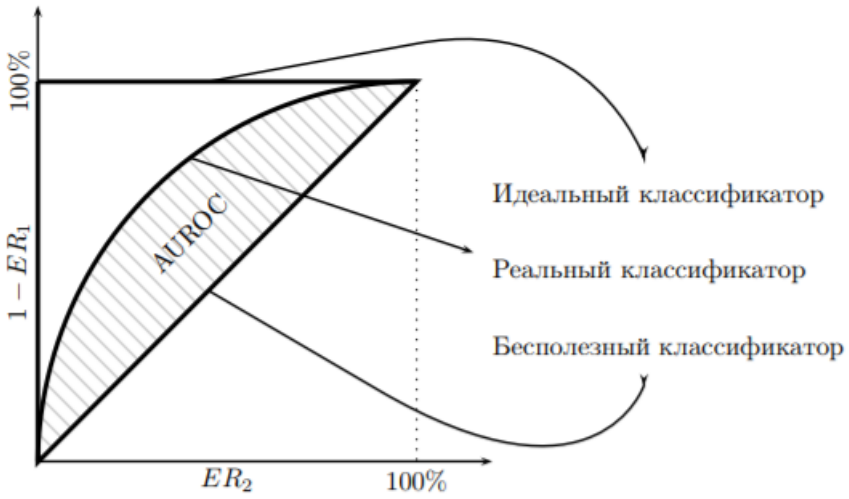

Рисунок 16 – Кривая ROC

Для оценки показателей, обсуждаемых в этом разделе, необходим набор справочных данных, для которого он известен, когда произошли сбои. В машинном обучении это называется «помеченным набором данных». Поскольку метрики оценки определяются с использованием статистических оценок, набор данных должен быть максимально большим. Однако сбои – это, как правило, редкие

события, которые обычно устанавливают естественное ограничение на количество сбоев в наборе данных.

Если метод онлайн-прогнозирования предполагает оценку параметров на основе данных, набор данных должен быть разделен на три части:

1. Набор обучающих данных: данные, по которым выполняется оптимизация параметров.

2. Набор данных проверки. В случае, если алгоритм оптимизации параметров может привести к локальным, а не глобальным оптимам, или для контроля так называемого компромисса с биасвариацией, данные проверки используются для выбора наилучшего параметра.

3. Набор тестовых данных: оценка эффективности прогнозирования отказов выполняется на данных, которые не использовались для определения параметров метода прогнозирования. Такая оценка также называется оценкой вне выборки.

3.4 Обсуждение полученных результатов

Для расчета логистической регрессии необходимо взять таблицы с закодированными категориальными признаками, заполнить недостающие данные и нормализовать их (привести к значениям от 0 до 1).

Использование логистической регрессии из Scikit-Learn как первой модели. Синтаксис – создание модели, ее тренировка и предсказание вероятности при помощи predict_proba. Результат: 0.673.

```
from sklearn.linear_model import LogisticRegression

# Создаем модель
log_reg = LogisticRegression(C = 0.0001)

# Тренируем модель
log_reg.fit(train, train_labels)
LogisticRegression(C=0.0001, class_weight=None, dual=False,
          fit_intercept=True, intercept_scaling=1, max_iter=100,
          multi_class='ovr', n_jobs=1, penalty='l2', random_state=None,
          solver='liblinear', tol=0.0001, verbose=0, warm_start=False)
Теперь модель можно использовать для предсказаний. Метод prdict_proba даст на выходе ма
ссив m x 2, где m - количество наблюдений, первый столбец - вероятность 0, второй - вер
оятность 1. Нам нужен второй (вероятность невозврата).


log_reg_pred = log_reg.predict_proba(test)[:, 1]
```

Использовать улучшенную модель – Random Forest Classifier . Это гораздо более мощная модель, которая может строить сотни деревьев и выдавать куда более точный результат. Используется 100 деревьев. Схема работы с моделью все та же, совершенно стандартная – загрузка классификатора, тренировка. предсказание. В таблице 1 представлено сравнение алгоритмов.

```
from sklearn.ensemble import RandomForestClassifier

# Создадим классификатор
random_forest = RandomForestClassifier(n_estimators = 100, random_state = 50)

# Тренировка на тернировочных данных
random_forest.fit(train, train_labels)

# Предсказание на тестовых данных
predictions = random_forest.predict_proba(test)[:, 1]

# Создание датафрейма для загрузки
submit = app_test[['SK_ID_CURR']]
submit['TARGET'] = predictions

# Сохранение
submit.to_csv('random_forest_baseline.csv', index = False)
```

Таблица 1 – Сравнение различных алгоритмов в обучающем наборе данных

Используемая модель классификации	Результат
Logistic Regression	0.673 ± 0.004
Random Forest Classifier	0.683 ± 0.003
Light Gradient Boosting	0.735 ± 0.001

Выводы по главе 3

Была поставлена цель – спрогнозировать, будет ли плтежеспособен клиент банка выплатить запрашиваемый кредит Данные предоставлены банком Home Credit.

Чтобы приступить к прогнозированию, для начала следует тщательно изучить и проанализировать набор данных, который очень большой, с большим количеством признаков и иногда неполным набором информации о потенциальном заемщике. Поскольку никакой предварительной информации о признаках не было известно, была попытка найти соответствующие связи между признаками после изучения информации в наборе данных. Получив представление о наборе данных, были выбраны лучшие наборы признаков, которые объясняют набор данных лучше всего. Следующим шагом были исследованы параметры методов:

– k ближайших соседей;
– случайный лес;
– градиентный бустинга;
– легкого градиентного бустинга.

В табличном виде продемонстрировано сравнение всех алгоритмов в обучающем наборе данных Bosch с 3-кратной перекрестной проверкой. Была проработана метрика эффективности. Методы градиентного бустинга и легкого градиентного бустинга продемонстрировали наилучший результат, чем методы простых классификаторов (k Ближайших соседей, случайные деревья).

4 КОММЕРЦИАЛИЗАЦИЯ ПРОЕКТА

Коммерциализация проекта – это привлечение инвесторов для финансирования деятельности по реализации этого новшества из расчета участия в будущей прибыли в случае успеха. В тоже время процесс выведения инновационного проекта на рынок является ключевым этапом инновационной деятельности после чего (выведения на рынок) происходит возмещение затрат разработчика (или владельца) инновационного продукта и получение им прибыли от своей деятельности. Процесс выведения инновационного проекта на рынок содержит несколько этапов:

В процессе коммерциализации очень важно выбрать метод. На рисунке 17 представлены основные способы коммерциализации инноваций. У предприятия есть выбор: самостоятельно коммерциализировать проект и пройти все перечисленные выше этапы, либо можно продать лицензию, либо полностью все права. Каждый метод предоставляет разработчикам широкие возможности по реализации. Варианты получения прибыли от проекта так же зависят от самого проекта. Иногда возможно применение сразу нескольких методов коммерциализации инноваций.

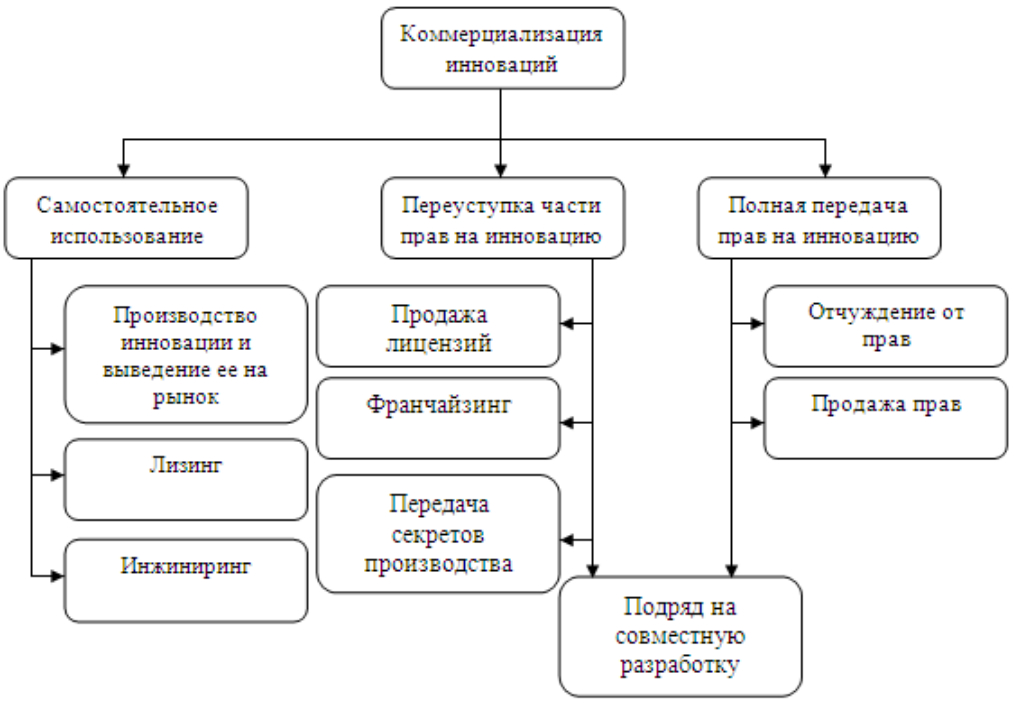

Рисунок 17 – Методы коммерциализации

Перед выбором метода коммерциализации, нужно рассмотреть каждый и выбрать тот, который лучше всего подходит для данной ситуации и для данного проекта. В Таблице 2 приведены основные достоинства и недостатки каждого метода.

Таблица 2 – Достоинства и недостатки способов коммерциализации инноваций

Способы коммерциализации	Достоинства	Недостатки
Самостоятельное использование	При успешной организации производства и «захвате» ниши на рынке, очень высокие доходы; Постоянный контроль предприятия и производства; полное распоряжение правами на интеллектуальную собственность (инновации).	Высокие риски; Большой срок окупаемости; Требуется наличие значительных финансовых ресурсов.

Окончание таблицы 2

Переуступка части прав на инновацию	Минимальные риски; Небольшие затраты; Достаточно короткий срок окупаемости; Выход на новые рынки за счет других компаний; Возможность формирования собственного товарного знака; Получение финансирования от заказчика при заключении подрядного договора.	Значительно меньше доходы по сравнению с другими способами коммерциализации; Риск нарушения лицензии патентных прав; Риск появления контрафактной продукции.
Полная передача прав на инновацию	Минимальные риски; Небольшие затраты; Минимальный срок окупаемости; Возможность получения очень высокого дохода, в зависимости от значимости разработанной инновации.	Риск не дополучения потенциального дохода; Из-за усиления позиций конкурентов вероятна вынужденная смена области деятельности.

Для реализации первого метода потребуются существенные трудовые, временные и финансовые ресурсы. Завоевание рынка и окупаемость скорее всего станут возможны в средне- или долгосрочной перспективе. Но даже если все хорошо организовано, остается риск, что спроса на продукцию не будет.

При выборе второго или третьего метода инвестиции в проект можно вернуть в краткосрочном периоде. Если предприятие продает лицензию, то вместе с ней и часть рынка переходит к лицензиату, но предприятие также может приобрести часть рынка лицензиата. В случае продажи лицензии разработчик получа-

ет стабильный доход в виде роялти. При продаже прав предприятие теряет все свои права на разработку, но зато получает значительный доход (в зависимости от значимости инновации).

Так как получение прибыли является главной целью, то наиболее перспективным методом для коммерциализации данного проекта является Самостоятельное использование. Это также обосновано тем, что при данном методе коммерциализации у предприятия присутствует полное распоряжение правами на интеллектуальную собственность (инновации), а это является довольно важным и существенным аспектом конкуренции рынка.

4.1 Дорожная карта коммерциализации проекта

Дорожная карта – это развернутый пошаговый план развития проекта, сформированный с учетом особенностей рынка и существующих технологий. Грамотно составленная «дорожная карта» помогает спрогнозировать пути развития проекта и выбрать наиболее эффективную стратегию. Таким образом, «дорожная карта» является мощным инструментом стратегического развития, планирования и принятия управленческих решений.

Составление «дорожной карты» — важный этап в создании инновационного продукта. Для формирования «дорожной карты» требуется провести тщательный анализ рынка, изучить технологии, оценить продукт, учесть особенности отрасли.

Дорожная карта – это наглядное представление пошагового сценария развития определенного объекта. Процесс формирования дорожной карты

коллективная ревизия имеющегося потенциала развития;

обнаружение возможностей роста;

обнаружение рисков;

выявление потребности в ресурсном обеспечении.

Планирование стратегии

Основные цели проекта

4.1.1 Планирование стратегии: основные цели и источники доходов проекта

Любое производство нуждается в определенном наборе услуг. Таким образом наше предприятие осуществляет свою деятельность в самой динамичной сфере - сфере услуг.

Основная цель проекта заключается в создании веб-представительства компании, деятельность которого будет направлена на предоставление услуг по оценке платежеспособности клиентов (заёмщиков) банков и кредитных организаций.

Наличие веб-представительства даст компании следующие преимущества и решение таких задач, как:

- создание веб-представительства для получения высоких доходов;
- организация получения стабильной прибыли;
- формирование и продвижение имиджа компании;
- увеличение спроса на предоставляемую услугу;
- улучшение системы связей с общественностью;
- обеспечение потребителей, партнеров, рекламных агентов полной и актуальной информацией о товаре и фирме;
- обеспечение информационной поддержки потребителей посредством обратной связи;
- расширение каналов сбыта предприятия.

Однако основной сферой деятельности планируется разработка, реализация и сервис по предоставлению услуги, а также в дальнейшем разработка и доработка существующего программного решения на базе новых информационных технологий. Данная отрасль является достаточно молодой для российского рынка и поэтому большинство компаний испытывают нехватку в профессиональных, качественных услугах. Потребность в данной услуге весьма велика. Все это позволит предоставить необходимые решения для плодотворного функционирования различного рода компаний.

Доходность предприятия подразумевает распространение (продажу) предоставляемых услуг, а также размещение интернет рекламы, схожей тематики на разработанном интернет ресурсе. Планируется что реклама будет осуществляться по модели CPM – цена, устанавливаемая за тысячу показов. Доходность предприятия отображена в таблице 3.

Таблица 3 – Доходность веб-представительства компании

Период	Вид услуги	Объем реализации в месяц, шт.	Стоимость, руб.	Прибыль от реализации, руб.
1-3 месяц	Предоставление услуги	5-15 шт.	15 000	75 000 – 225 000
	Реклама	1 шт.	350 CPM (показов – 2 000)	700
4-6 месяц	Предоставление услуги	20-35 шт.	25 000	500 000 – 875 000
	Реклама	1-3 шт.	700 CPM (показов – 10 000)	7 000

4.1.2 Оценка потенциальных возможностей Интернета для бизнеса

Таблица 4 – Целевая аудитория, конкурентная среда и потенциальные партнеры

Целевая аудитория	Описание
1. Продажи сервиса банкам и кредитным организациям	Покупатели: банки и кредитные

Окончание таблицы 4

2. Продажи лицензии на использование сервиса банкам и кредитным организациям.	В данной схеме банк (кредитная организация) является инициатором покупки лицензии на использование сервиса
3. Продажи лицензии на использование сервиса потенциальным клиентам (заёмщикам кредитов)	В данной схеме заёмщик (а не банк) является инициатором покупки лицензии на использование сервиса

В настоящее время количество Интернет-ресурсов, реализующих те или иные услуги или программные средства достаточно велико. Однако Web-сайтов, предлагающих целенаправленную услугу по решению данной проблемы не так много, особенно те, кто предлагает эксклюзивные решения, чем и будет являться данная услуга.

Потенциальными партнерами следует выделить организации, которые могут предоставить рекламу или размещение данной услуги на своем Web-ресурсе.

4.2 Создание сайта

Варианты доменного имени для сайта

Как придумать доменное имя для своего интернет-ресурса?

Выбор домена для своего интернет-магазина аналогичен со схемой придумывания названия обычного магазина. Основными методами для этого являются:

«мозговой» штурм

оформление заказа на неймингна одной или нескольких бирж фриланса

Для успешной работы интернет-представительства компании, доменное имя должно соответствовать некоторым критериям:

быть созвучным тематике бизнеса, либо продаваемых товаров;

легкость написания, произнесения, и запоминания;

быть свободным, т.е. незарегистрированным кем-нибудь другим.

В настоящее время регистраторы доменных имен предлагают на выбор более 740 различных доменных зон, из которых фактически для бизнеса подойдет не более десятка. Топ-4 самых популярных зон – это .ru, .com, .net, org. В таблице 5 представлены варианты доменных имен, выделены их достоинства и недостатки.

Таблица 5 – достоинства и недостатки доменных имен

Домен	Достоинства	Недостатки
Prediction.com/.ru	Незарегистрированный домен; Созвучно с тематикой; Частично отражает суть услуги и сайта; Просто запомнить; Популярная зона.	Частично понятна суть предлагаемой услуги и наполнения сайта;
Failure.com/.ru	Незарегистрированный домен; Созвучно с тематикой; Частично отражает суть услуги и сайта; Просто запомнить; Популярная зона.	Частично понятна суть предлагаемой услуги и наполнения сайта;

Окончание таблицы 5

ailure-prediction.com/.ru	Незарегистрированный домен; Созвучно с тематикой; Понятна суть предлагаемой услуги и сайта в целом; Популярная зона.	Сложность в запоминании и написания ссылки сайта.
Poif.com/.ru	Незарегистрированный домен; Краткое запоминающееся название; Популярная зона.	Непонятен смысл сайта и предлагаемой услуги; В полной мере не отражает сущность наполнения сайта, предлагаемой услуги.

Тип сайта для веб-представительства компании

Landing page – это «легкий» сайт, созданный для привлечения целевой аудитории к товарам, услугам или акциям. Обычно на целевую страницу попадают благодаря переходу с контекстной рекламы или информации поисковиков. На подобных одностраничных сайтах расположена необходимая для посетителя информация в такой форме, чтобы он максимально сфокусировался на ней. Более того, правильный лендинг направлен на стимулирование желания совершить полезное действие: регистрация на сайте, оформление заказа, звонок в офис компании, подписка на рассылку. Благодаря такой направленности landing page обеспечивает повышение конверсии до 30% и более. Как правило, landing page имеют привлекательный и в меру лаконичный дизайн как показано на рисунке 18. Все делается для того, чтобы на странице отсутствовали факторы, отвлекающие от ее содержания.

Преимущества успешного лендинга:

риентируясь на конкретную целевую аудиторию при правильной раскрутке и рекламе, конверсия landing page будет намного больше, чем у обычных сайтов;

благодаря простоте создания страницы она может быть готова к работе и запущена за несколько часов, а изменение информации на ней происходит в считанные минуты;

посадочные страницы обычно быстро загружаются, даже на устройствах со слабым интернетом, посетителю не надо долго ждать;

landing page – это весьма действенный и результативный инструмент, ведь если даже посетитель сайта ничего не приобретет или не закажет, велика вероятность, что он оставит свои данные. Таким образом, сформируется база потенциальных клиентов, которым в дальнейшем можно напоминать о себе по средствам e-mail рассылки;

при помощи landing page можно успешно повысить эффект от контекстной рекламы

лендинг пейдж позволяет оценить и проанализировать объемы и целесообразность интернет-продаж

помогает увеличить продажи при некачественном основном сайте

низкая стоимость разработки.

Информационное наполнение сайта

Тип и формат представления информации: текст и картинки обозначающие программное обеспечение;

Структурирование информации:

логотип и заголовок;

демонстрация услуги;

преимущества данного решения, в дальнейшем возможные акции;

описание оффера;

коммуникация.

Форма подачи информации:

Призыв используется в описание заголовка услуги. Призывает пользователя к действию нажать на кнопку заказа данной услуги или получения консультации по данному решению.

Аргументация используется в описание преимуществ программного решения, а также в описание оффера. Из названия следует, что пост-аргументация приводит доказательства определенной точки зрения. Его отличительная черта – вопрос «почему?», с которого начинается заголовок (например, почему предприятия должны радоваться появлению данного программного продукта: четыре причины).

Наполнение, расширение и актуализация информации. Landing page будет разбит на блоки:

Главный экран – его функция – произвести нужное впечатление на человека, информировать о том, куда он попал, мотивировать остаться и проскроллить страницу вниз.

Целевое действие – Бизнесу нужны клиенты, поэтому на лендинге должны быть блоки, которые будут генерировать лиды: формы заказа, подписки, обратной связи или телефон.

Дизайн сайта (обложка), наполнение:

Главный экран сайта – первое впечатление от компании. Есть всего несколько секунд, чтобы убедить пользователя остаться на странице.

Набор инструментов для этого небольшой: заголовок, подзаголовок, кнопка или форма, логотип, фон или изображение на фоне, меню, стрелка вниз.

Фон обложки – хорошая фотография, атмосферное видео, просто цвет, градиент или иллюстрация. Стоит обратить внимание на сочетание фона с текстом: фотография может быть удачной сама по себе, но если она неоднородная, пестрая, то она будет плохо работать с текстом. Видео нужно снимать, во-первых, плавно, во-вторых, лучше брать увеличенный фокус, чтобы все объекты были крупноваты.

Логотип – компании или продукта можно расположить как на самой обложке, так и в меню.

Инструменты работы с аудиторией сайта

Анализ поведения пользователей на сайте – владельцы ресурса могут следить за посещаемостью сервера, за наиболее популярными маршрутами по сайту, точками входа и выхода посетителей, временем, проведенным на каждой из страниц и т.д. Данная информация используется и для определения эффективности рекламных направлений, и для оптимизации структуры и навигации сайта. Получать подобные данные можно с помощью анализатора логов сайта или продвинутых счетчиков.

Консультации – с помощью интернет-технологий можно эффективно осуществлять информационную поддержку своих клиентов. Специалисты компании с помощью on-line конференций, чата или по e-mail могут отвечать на вопросы, давать консультации. В случае с конференцией это будет не столь оперативно (хотя и конференции могут проводиться в реальном режиме времени), но наглядно и информативно. Конференции имеют удобную древовидную структуру, а отсутствие необходимости отвечать сразу позволяет более тщательно подготовить ответ.

Чат дает максимальную оперативность, ту же, что и телефонная линия, но при этом не надо платить за международные переговоры, а специалист службы поддержки может одновременно отвечать сразу на несколько вопросов. Самым же распространенным способом поддержки пользователей остаются консультации посредством электронной почты.

Мониторинг сайта

В таблице 6 отображена полная информация о возможностях сайта.

Таблица 6 – Возможности сайта

Наименование сайта	Prediction
Информационное наполнение	Информации четко структурирована (в дальнейшем планируется добавлять новости и обновления в блог), используются различные форматы представления информации, техническая поддержка, адекватная и структурированная информация сайта, имеется расстановка информационных акцентов.
Функциональность	Представление товара и формирование заказа осуществляется по нажатию кнопки и переходу на вспомогательный экраны, а также окно формирования заказа, присутствует связь при помощи e-mail или телефона (в дальнейшем предусмотрена разработка чат-бота).
Usability	Сайт эргономичен и удобен в использовании (присутствует простая и эффективная навигация, имеется карта местоположения, привычный вид полей и кнопок.

Окончание таблицы 6

Дизайн	Дизайн дополняет и усиливает заложенную в сайт информацию и функционал, простота использования сайта благодаря легкому дизайну, возможность изменения дизайн-решений (гибкость), уникальность и запоминаемость.
Техническая реализация	Сайт написан при помощи движка WordPress.
Маркетинг	На сайте присутствуют адреса, ссылки на сайт, средства сбора информации о посетителях сайта, посещаемость и поведенческая линия на сайте, работа с аудиторией сайта.

Медиаплан, ценовая политика

Медиапланирование – это планирование каналов и способов рекламы для составления медиаплана на основе прогнозов и полученных результатов.

Медиаплан для первого рекламного мероприятия по продвижению, созданного Web-ресурса, должен выполнять следующие условия:

Бюджет – 10 000-20 000 $ (по нынешнему курсу рубль-$ = 65,38);

Время рекламной компании – 4 недели;

Задача рекламной компании - привлечение посетителей (раскрутка нового ресурса).

Для рекламы конечно же будут использоваться самые распространенные и популярные рекламные площадки такие как Вконтакте и Яндекс.

Для начала, чтобы раскрутить бренд следует максимизировать сиюминутную прибыль. Иначе говоря - извлечь как можно больше денег из каждой продажи предоставляемой услуги, даже если это сокращает количество потенциальных

покупателей. В итоге, будет меньше клиентов, но и количество проблем по их обслуживанию также сократится. И, кроме того, каждый из клиентов принесет больший доход.

Расчет будет производиться по общим издержкам – сумма постоянных и переменных издержек.

Для любого сайта важны инструменты работы с аудиторией. Ниже приведен примерный список, который следует реализовать:

1) Анализ поведения пользователей на сайте – владельцы ресурса могут следить за посещаемостью сервера, за наиболее популярными маршрутами по сайту, точками входа и выхода посетителей, временем, проведенным на каждой из страниц и т.д. Данная информация используется и для определения эффективности рекламных направлений, и для оптимизации структуры и навигации сайта. Получать подобные данные можно с помощью анализатора логов сайта или продвинутых счетчиков.

2) Консультации – с помощью интернет-технологий можно эффективно осуществлять информационную поддержку своих клиентов. Специалисты компании с помощью on-line конференций, чата или по e-mail могут отвечать на вопросы, давать консультации. В случае с конференцией это будет не столь оперативно (хотя и конференции могут проводиться в реальном режиме времени), но наглядно и информативно. Конференции имеют удобную древовидную структуру, а отсутствие необходимости отвечать сразу позволяет более тщательно подготовить ответ.

3) Чат – в дальнейшем планируется разработка чата. Он дает максимальную оперативность, ту же, что и телефонная линия, но при этом не надо платить за международные переговоры, а специалист службы поддержки может одновременно отвечать сразу на несколько вопросов. Самым же распространенным способом поддержки пользователей остаются консультации посредством электронной почты.

4) Патчи, драйвера и обновления программ – продавцы программного обеспечения, помимо консультаций и инструкций, посредством Интернета могут распространять как непосредственно свою продукцию, так и патчи и обновления к ней. А производители высокотехнологического оборудования могут выкладывать на сайте для скачивания последние версии драйверов устройств.

В таблице 7 отображена полная информация о возможностях сайта.

Таблица 7 – Возможности сайта

Наименование	Prediction
Информационное наполнение	Информация четко структурирована (в дальнейшем планируется добавлять новости и обновления в блог), используются различные форматы представления информации, тех. поддержка, адекватная и структурированная информация сайта, имеется расстановка информационных акцентов.
Функциональность	Представление товара и формирование заказа осуществляется по нажатию кнопки и переходу на вспомогательный экраны, а также окно формирования заказа, присутствует связь при помощи e-mail или телефона (в дальнейшем предусмотрена разработка чат-бота).

Окончание таблицы 7

Usability	Сайт эргономичен и удобен в использовании (присутствует простая и эффективная навигация, имеется карта местоположения, привычный вид полей и кнопок.
Дизайн	Дизайн дополняет и усиливает заложенную в сайт информацию и функционал, простота использования сайта благодаря легкому дизайну, возможность изменения дизайн-решений (гибкость), уникальность и запоминаемость.
Техническая реализация	Сайт написан при помощи движка WordPress.
Маркетинг	На сайте присутствуют адреса, ссылки на сайт, средства сбора информации о посетителях сайта, посещаемость и поведенческая линия на сайте, работа с аудиторией сайта.

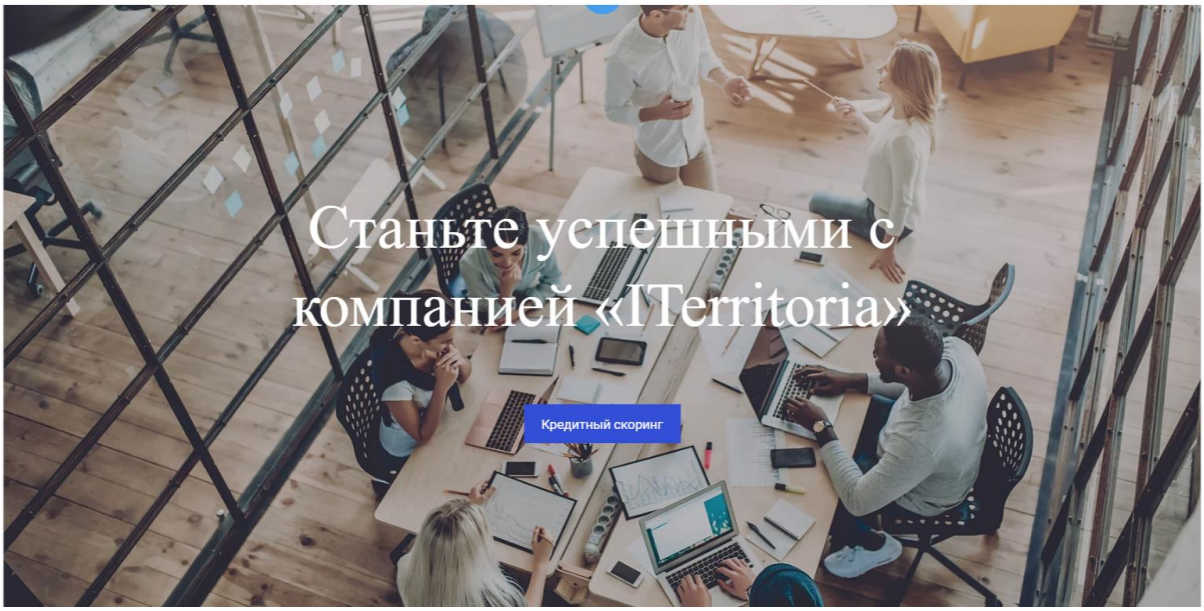

Рисунок 18 – Главная страница Landing page

4.3 Медиапланирование и ценовая политика сайта

Медиапланирование – это планирование каналов и способов рекламы для составления медиаплана на основе прогнозов и полученных результатов.

Медиаплан для первого рекламного мероприятия по продвижению, созданного Web-ресурса, должен выполнять следующие условия:

1) Бюджет – 10 000-30 000 $ (по нынешнему курсу рубль-$ = 64,60);
2) Время рекламной компании – 4 недели;
3) Задача рекламной компании – привлечение посетителей (раскрутка нового ресурса).

Для рекламы конечно же будут использоваться самые распространенные и популярные рекламные площадки такие как Вконтакте и Яндекс. В таблице 7 продемонстрирована полная информация по выбору той или иной площадки.

Для начала, чтобы раскрутить бренд следует максимизировать сиюминутную прибыль. Иначе говоря - извлечь как можно больше денег из каждой продажи предоставляемой услуги, даже если это сокращает количество потенциальных покупателей. В итоге, будет меньше клиентов, но и количество проблем по их обслуживанию также сократится. И, кроме того, каждый из клиентов принесет больший доход.

Расчет будет производиться по общим издержкам – сумма постоянных и переменных издержек.

"Снятие сливок". Предлагая новую революционную услугу, стоит изначально установить на нее высокую цену, так как тот, кто чувствителен к нововведениям – нечувствителен к цене. Затем снижаем цену и "снимаем сливки" со следующего слоя покупателей и так далее. В конечном счете, цена падает под воздействием того, что товар укрепляет свои позиции на рынке и конкуренты снижают цены. Так же стоит задуматься о скидках на повторное приобретение услуги если в таковой нуждаются.

Таблица 8 – Медиаплан

Рекламные каналы	Дополнительная информация	Посыл	Формат	Общая стоимость рекламы, руб.	Число публикаций	CPT, руб.	Частота	Бюджет
Контекстная реклама в Яндексе	Только горячие запросы	Инновационная разработка	Спец. размещение	120 000	1	1 500	1	120 000
РСЯ	Публикация рекламных постов	Инновационная разработка	Графические баннеры	450 000	3	40	4	1 350 000
Группа Вконтакте	Публикация рекламных постов	Инновационная разработка	Промо-пост	10 000	3	833	1,2	30 000
Таргетированная реклама Вконтакте	Публикация рекламных постов	Инновационная разработка	Лид-форма	100 000	2	1 500	1,2	200 000
Mail.Ru Group	Рекламный баннер	Инновационная разработка	Графические баннеры	50 000	2	900	1	100 000
Итого:				730 000	11			1 800 000

71

Выводы по главе 4

Составление «дорожной карты» – важный этап в создании инновационного продукта. Благодаря составлению которой, было спланировано веб-представительство будущей компании, деятельность которого направлена на предоставление услуг по предотвращению и сокращению технологических сбоев производственной линии на предприятии.

Была составлена таблица примерной доходности веб-ресурса. Рассмотрена целевая аудитория и проведен SWOT-анализ для определения внешних и внутренних факторов, влияющих на возможности работы в Интернете и формировании интернет стратегии.

Подходящим вариантом типизации сайта был выбран Landing page с доменным именем Prediction.com/.ru. После чего, продумано информационное наполнение, продемонстрирован первоначальный предполагаемый вид и раскрыты возможности сайта.

В дальнейшем планируется разработать инструменты работы с аудиторией, такие как:

– анализ поведения пользователей на сайте;
– консультации;
– чат;
– патчи, драйвера и обновления программ.

Разработан медиаплан и ценовая политика.

ЗАКЛЮЧЕНИЕ

1. В ходе исследования предметной области был изучен механизм взаимодействия банка с потенциальными клиентами на примере банка Хоум Кредит, механизм кредитного скоринга и методов его достижения.

2. В дипломной работе были изучены методы машинного обучения, а в частности контролируемое обучение и обучение без учителя. Рассмотрены популярные алгоритмы, которые используются в машинном обучении для решения данных проблем, такие как k Nearest Neighbor или k Ближайших Соседей, Random forest (случайный лес) и GBoost. Но для того чтобы понять в пользу какого алгоритма сделать выбор, который будет наилучшим для решения данной задачи, стоило разобраться в предоставленных данных и задачах, которые нужно решить, что и было продемонстрировано в 3-ей главе.

3. Была раскрыта тема проекта и ее цель, которая подразумевает прогнозирование платежеспособности клиентов банка. Показано описание набора данных предоставленных банком Хоум Кредит, продемонстрированных в виде табличных данных. Произведена оценка качества и показателей. В связи с большим объемом данных и иногда неполной картиной относительно какого-либо клиента банка (некоторая информация отсутствовала в предоставленных данных) необходимо было первостепенно произвести отчистку (предварительную обработку) данных. Все данные по итогу были разбиты на категориальные, числовые и особенности даты. Исходя из описания методов во второй главе и набора данных из третьей главы был сделан выбор в пользу метода Gradient Boosting с описанием преимуществ в сравнении с другими методами. Разработана коммерциализация проекта по этапам. Изначально была разработана дорожная карта коммерциализации данного проекта, которая подразумевает наглядное представление пошагового сценария развития, в которую входит – планирование стратегии, исходя из задач, для решения поставленной цели, описаны источники доходов по видам предоставляемых услуг и их стоимость в рублях. Проведена оценка потенциальных возмож-

ностей Интернета для бизнеса, в которой были рассмотрены: целевая аудитория, конкурентная среда и потенциальные партнеры. Также продемонстрирована таблица SWOT-анализа.

4. По потенциальным возможностям Интернета было выбрано создать сайт по предоставлению услуги прогнозирования сбоев технологических линий другим компаниям. Для решения данной задачи первостепенным было принято решение выбора доменного имени для сайта, а также был выбран тип и информационное наполнение сайта. Следующим шагом был выбор инструментов для работы с аудиторией сайта. В табличном виде представлен мониторинг сайта. Описано продвижение и ценовая политика сайта. Разработан медиаплан, также продемонстрированный в табличном виде.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Антипова О.Н. Банковское дело, М.: «Регулирование рыночных рисков», №4, 2005 г.

2. Батракова Л.Г. Экономический анализ деятельности коммерческого банка, М.; Логос, 2007г.

3. Барский А. Б. Нейронные сети: распознавание, управление, принятие решений. М.: Финансы и статистика, 2004. 176 с.

4. Бебрис А.О., Решетько Н.И. Формирование механизмов развития предпринимательских структур в условиях конкуренции. Вестник Университета (Государственный университет управления). 2011. № 17. С. 113-118.

5. Белоглазова Г.Н., Кроливецкая Л.П. Банковское дело, М.: «Финансы и статистика», 2004.

6. Букато В. И., Ю. И. Львов, Банки и банковские операции в России, М.: Финансы и статистика, 2006.

7. Глинкина, Е.В. Кредитный скоринг как инструмент эффективной оценки кредитоспособности // Финансы и кредит. – 2011. – № 16. – С. 43–47.

8. Кабушкин Н. И. Управление банковским кредитным риском / Кабушкин Н. И. – М: Новое знание, 2007. – 336 с.

9. Кирисюк Г. М ., Ляховский В. С. Оценка банком кредитоспособности заемщика // Деньги и кредит. 2010. № 4.

10. Лаврушин О. И. Банковские риски / Лаврушин О. И. – М.: КноРус, 2008. – 232 с.

11. Леонтьева Л.С., Кузнецов В.И., Конотопов М.Н., Орехов С.А., Башкатова Ю.И., Морева Е.Л., Орлова Л.Н. Теория менеджмента. Москва, 2013.

12. Лобанов А. А. Энциклопедия финансового риск-менеджмента / Лобанов А. А., Чугунов А. В. – М.: Альпина, 2009. – 936 с.

13. Любушин Н.И., Лещева В.Б., Дьякова В.Г. Анализ финансово-экономической деятельности предприятия. - М.; Юнити - Дана, 2005г.

14. Мальцев А.Э., Мальцев Э.В. Новые подходы к управлению розничным кредитным портфелем / Банковское кредитование, №2, 2005.

15. Министерство регионального развития [сайт]: URL: http://www.minregion.ru/activities/monitor/exec_evaluation

16. Положение «О порядке формирования кредитными организациями резерва на возможные потери по ссудам, поссудной и приравненной к ней задолженности» № 302-П от 26.03.2007

17. Положение ЦБР от 26 марта 2004 г. № 254-П «О порядке формирования кредитными организациями резервов на возможные потери по ссудам, по ссудной и приравненной к ней задолженности»// www.garant.ru›Все федеральные документы›102892

18. Помазанов М. В. Продвинутый подход к управлению кредитным риском в банке: методология, практика, рекомендации. Москва: Регламент, 2010.

19. Помазанов М. В., Колоколова О. В. Оценка вероятности банкротства предприятия по финансовым показателям// http://www.creditrisk.ru/publications/

files_attached/formula_preprint.pdf (дата обращения – 12.03.2019)

20. Решетько Н.И. Роль CRM-систем в разработке и реализации стратегии развития предприятия. Менеджмент в России и за рубежом. 2007. № 6. С. 138-141.

21. Соколов М.А. Аналитическая модель комплексной оценки эффективности интеграционных трансформаций организаций за счет слияний и поглощений. Транспортное дело России. 2010. № 6. С. 139-143.

22. Соколов М.А. Возможности использования зарубежного опыта в российской практике слияний и поглощений. Вопросы экономических наук. 2007. № 5. С. 199-201. Вопросы экономических наук. 2007. № 5. С. 199-201.

23. Соколов М.А. Организационно-экономическая модель управления стоимостью российских компаний в условиях проводимых процедур по слияниям и поглощениям. Проблемы экономики. 2007. № 5. С. 27-31.

24. Федеральный закон «О банках и банковской деятельности» № 395-1 от 02.12.1990 с изм. и доп. от 03.03.2008 Жуков Е. Ф. Банки и банковские операции, М.: ЮНИТИ., 2004.

25. Шорохов, Ю. Эффективность организаций: системные, организационные и психологические факторы (сайт Executive.ru, февраль 2008) [электронный ресурс]: URL: http://www.classs.ru/hrclub/practice/practice43.html

26. Штовба С. Д. Введение в теорию нечетких множеств и нечеткую логику. [Электронный ресурс], http://matlab.exponenta.ru/fuzzylogic/book1/

27. Abdou, H., et al. On the applicability of credit scoring models in Egyptian banks // Banks and Bank Systems. – 2007. – Vol. 2. – No. 1. – P. 4–20

28. Breiman L. Bagging Predictors. Machine Learning, 1996, vol. 24, iss. 2, pp. 123–140

29. Durand D. Risk Elements in Consumer Installment Financing. New York, National Bureau of Economic Research Books, 1941, 163 p

30. Friedman N., Geiger D., Goldszmidt M. Bayesian Network Classifiers. Machine Learning, 1997, vol. 29, iss. 2-3, pp. 131–163

31. Gurný, P., et al. Comparison of Credit Scoring models on probability of default estimation for US Banks // Prague Economic Papers. – 2013.– No. 2. – P. 163–181

32. RiskCalk for private companies: Moody's Default Model. http://www.creditrisk.ru/publications/files_attached/ Moodys_Default_Model.pdf

33. Thomas, L.C., Edelman, D.B., Crook, J.N. Credit scoring and its applications. – USA: SIAMP, 2002. – 248 p

34. West D. Neural network credit scoring models //Computers & Operations Research. – 2000. – Т. 27. – №. 11. – С. 1131-1152

35. Wolpert D.H. Stacked Generalization. Neural Networks, 1992, vol. 5, no. 2, pp. 241–259