ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ОБРАЗОВАНИЯ «НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Высшая школа бизнеса
Бизнес-информатика
Проектный семинар «ИТ-консалтинг»

Курсовой проект

Тема: «Внедрение системы предиктивной аналитики для управления рисками коммерческого банка»

Выполнили:
Губин Д.М.
Лебедев А.А.
Мамедов А.А.
Сухоруков Г.В.

29 мая 2021

Москва

Оглавление

Введение ... 3
Описание объекта исследования ... 3
Поставленные задачи ... 3
Ожидаемые результаты ... 4
Команда проекта ... 5
Стейкхолдеры проекта ... 6
Экспресс-анализ рынка банковского кредитования в РФ ... 6
Первый раздел – Обследование ... 10
Кредитный конвейер as-is ... 10
Кредитный конвейер to-be ... 11
Исходный набор переменных ... 13
Предполагаемые наиболее значимые предикторы и целевые переменные ... 25
Качество данных ... 25
Ожидания от модели ... 26
Второй раздел – Формализация требований ... 27
Требования к архитектуре приложения BI ... 27
Требования к модели ... 28
Требования к хранилищу данных ... 29
Третий раздел – Выбор ИТ-решения ... 30
Модель прогнозирования ... 30
BI-приложение ... 30
Хранилище данных ... 32
Архитектура системы ... 33
Четвёртый раздел – Построение модели ... 38
Модель прогнозирования ... 38
Оценка качества работы модели ... 42
Заключение ... 44
Библиографический список ... 46
Приложения ... 48

Введение

Цель проекта: Разработка и внедрение системы предиктивной аналитики для совершенствования кредитного скоринга коммерческого банка.

Объект исследования: Объектом исследования в данном проекте является коммерческий банк, для которого разрабатывается система предиктивной аналитики.

Предмет исследования: Бизнес-процесс “Управление рисками”, а также поддерживающие его аналитические технологии.

Описание объекта исследования

Объект исследования - крупный коммерческий банк, входящий в пятёрку крупнейших банков РФ по объёму активов. Банк имеет около 800 отделений и офисов более чем в 100 городах России. Активная клиентская база составляет 543 тыс. корпоративных клиентов и 5,7 млн физических лиц. Стратегическими приоритетами банка в настоящее время являются поддержание статуса одного из лидирующих банков в России с акцентом на надежность и качество активов, а также ориентированность на лучшие в отрасли качество обслуживания клиентов, технологии, эффективность и интеграцию бизнеса.

С целью повышения качества и надежности выдаваемых кредитов, руководством кредитного департамента банка принято решение внедрить новую систему кредитного скоринга. Предполагается, что новая система будет работать не с личными данными клиентов, а с обезличенными транзакционными данными. Это должно упростить и улучшить процесс определения надёжных заёмщиков.

Поставленные задачи

1. Провести экспресс-анализ рынка банковского кредитования
2. Изучить и описать основные подходы к оценке кредитных рисков на рынке
3. Построить и описать схемы кредитного конвейера as-is и to-be
4. Выявить и проанализировать требования банка к прогнозной модели

5. Разработать требования к системе предиктивной аналитики для управления рисками коммерческого банка
6. Провести исследование и подготовку данных о транзакциях заёмщиков
7. Выделить значимые факторы для скоринговой модели
8. Выбрать метрики качества модели
9. Выбрать и обосновать методы прогнозирования
10. Построить модели машинного обучения для прогнозирования кредитного риска, а затем создать итоговую модель как ансамбль исходных
11. Провести предварительную оценку качества итоговой модели
12. Спроектировать и разработать хранилище данных для системы предиктивной аналитики
13. Спроектировать и разработать BI-интерфейс для системы предиктивной аналитики
14. Построить систему предиктивной аналитики путем интеграции разработанного хранилища данных, модели машинного обучения и BI-интерфейса
15. Провести предварительное тестирование системы предиктивной аналитики
16. Описать ожидаемые бизнес-эффекты от внедрения системы предиктивной аналитики

Ожидаемые результаты

- Построена модель машинного обучения для прогнозирования кредитного риска для полностью обезличенных транзакционных данных
- Разработано хранилище данных для системы предиктивной аналитики
- Разработан BI-интерфейс для системы предиктивной аналитики
- Создана система предиктивной аналитики для управления рисками коммерческого банка
- Описаны ожидаемые бизнес-эффекты от внедрения системы предиктивной аналитики

Команда проекта

Таблица 1 «Команда проекта»

Команда и роли в проекте		Описание
Лебедев Андрей	Team leader и разработчик модели, бизнес-аналитик	Координация проекта; выявление и анализ требований к модели; выделение значимых факторов, выбор метрики качества и методов прогнозирования для моделей; построение моделей машинного обучения; интеграция итоговой модели в систему предиктивной аналитики
Губин Даниил	Разработчик BI-интерфейса, бизнес-аналитик	Разработка требований к интерфейсу BI; анализ рисков проекта; создание устава проекта; проектирование и разработка BI-интерфейса; Интеграция BI-интерфейса в систему предиктивной аналитики; связь BI-интерфейса с хранилищем данных
Сухоруков Георгий	Рыночный аналитик, бизнес-аналитик	Проведение экспресс-анализа рынка; изучение и описание подходов к оценке кредитных рисков; создание схем кредитного конвейера as-is и to-be; составление плана проекта; оценка ожидаемых бизнес-эффектов от внедрения системы
Мамедов Артём	Data-engineer, data-scientist, бизнес-аналитик	Разработка требований к хранилищу данных; проведение исследования и подготовки данных о транзакциях заёмщиков; проектирование и разработка хранилища данных; интеграция хранилища данных в систему предиктивной аналитики

Стейкхолдеры проекта[1]

Таблица 2 «Стейкхолдеры проекта»

Стейкхолдеры	Значение	Вид	Участие в проекте и влияние
Акционеры	Повышение прибыли в результате усовершенствования механизма кредитного скоринга	Внутренний	Первичный (инвесторы проекта); влияние высокое
Кредитный департамент банка	Повышение эффективности и качества работы департамента	Внутренний	Первичный (заказчик проекта); влияние очень высокое
Контролирующие органы (ЦБ РФ)	Повышение надежности банка, его лучшее соответствие установленным нормам	Внешний	Вторичный (косвенно заинтересован в проекте); влияние среднее
Заёмщики	Снижение риска отказа “хорошим” заёмщикам; высокая вероятность отказа “плохим” заёмщикам	Внешний	Вторичный (косвенно заинтересован в проекте); влияние низкое
Вкладчики	Снижение риска банкротства банка и потери вклада	Внешний	Вторичный (косвенно заинтересован в проекте); влияние низкое

Экспресс-анализ рынка банковского кредитования в РФ[2]

В целом, в 2018–2019 годах и 1-м квартале 2020 наблюдался постепенный рост числа заёмщиков на рынке банковского кредитования. Однако начало пандемии коронавируса и введение ограничений на ее фоне привели к краткосрочному сокращению розничного кредитования во 2-м квартале 2020 года. В итоге, в 1-м полугодии 2020 года количество заёмщиков в среднем не изменилось в сравнении со 2-м полугодием 2019 года. Но уже в 3-м квартале имел место небольшой восстановительный рост во всех сегментах кредитования, во многом обусловленный реструктуризацией задолженностей. В итоге, во 2-м полугодии 2020 года общее количество заемщиков, имеющих задолженность по банковскому

[1] Под стейкхолдерами данного проекта понимаются лица и организации, которые заинтересованы в реализации проекта (т.е. на них влияет результат проекта) и/или оказывают влияние на проект.

[2] На основе отчёта ЦБ РФ “АНАЛИЗ ДИНАМИКИ ДОЛГОВОЙ НАГРУЗКИ НАСЕЛЕНИЯ РОССИИ В II–III КВАРТАЛАХ 2020 ГОДА НА ОСНОВЕ ДАННЫХ БЮРО КРЕДИТНЫХ ИСТОРИЙ”. www.cbr.ru/collection/collection/file/31945/review_03022021.pdf

кредиту, в среднем составило 36,1 млн человек, что практически соответствует значениям до начала пандемии.

Таблица 3 «Динамика количества заёмщиков на рынке банковского кредитования»

Год	2018, 1-е полугодие	2018, 2-е полугодие	2019, 1-е полугодие	2019, 2-е полугодие	2020, 1-е полугодие	2020, 2-е полугодие
Кол-во заёмщиков, млн. чел	34,37	35,07	35,8	36,13	36,05	36,1
Изменение, %	-	2,04	2,08	0,92	-0,22	0,14

Качество кредитных портфелей банков, в целом, остается устойчивым, в том числе благодаря проведенной реструктуризации задолженностей в период ограничений из-за пандемии коронавируса.

Заёмщики в среднем имеют задолженность более чем по одному кредиту, что указывает на **важность расчета** показателя долговой нагрузки при выдаче кредита. Следует отметить, что среднее количество кредитов, приходящихся на одного банковского заемщика, достаточно стабильно увеличивалось в 2018 – 1-й половине 2020 года. Во 2-й половине 2020 года рост продолжился, хотя и значительно замедлился.

Таблица 4 «Динамика среднего количества кредитов на одного заемщика»

Год	2018, 1-е полугодие	2018, 2-е полугодие	2019, 1-е полугодие	2019, 2-е полугодие	2020, 1-е полугодие	2020, 2-е полугодие
Кол-во кредитов на заёмщика, шт.	1,707	1,75	1,803	1,853	1,893	1,90
Изменение, %	-	2,51	3,03	2,77	2,16	0,37

В 2018 – 1-м квартале 2021 года наблюдается непрерывный рост объёма задолженности по кредитам. В настоящее время темпы этого роста сильно снизились, что, скорее всего, обусловлено значительным замедлением увеличения числа заёмщиков на рынке. Тем не менее, суммарный объём задолженности составил рекордные 19,83 трлн. руб. по итогам 1-го квартала 2021 года.

Таблица 5 «Динамика объема задолженности по банковским кредитам»

Год	2018, 1-е полугодие	2018, 2-е полугодие	2019, 1-е полугодие	2019, 2-е полугодие	2020, 1-е полугодие	2020, 2-е полугодие	2021, 1-й квартал
Объём задолженности, трлн. руб.	13,33	14,73	16,1	17,17	18,13	19,3	19,83
Изменение, %	-	10,5	9,3	6,65	5,59	6,45	2,75

Следует отметить, что в 2018–2020 годах суммарный объём задолженности по банковским кредитам и среднее количество кредитов на одного заёмщика увеличивались более быстрыми темпами, чем общее количество заёмщиков на рынке банковского кредитования. Несомненно, это должно было приводить к постепенному росту показателя долговой нагрузки[3]. Так и происходило в 2018–2020 годах. В настоящее время (1-й квартал 2021 года) долговая нагрузка населения РФ достигла отметки в 11,7%, что является историческим рекордом.

Таблица 6 «Динамика долговой нагрузки населения РФ»

Дата	01.04.2018	01.04.2019	01.04.2020	01.04.2021
Долговая нагрузка, %	9,7	10,6	10,9	11,7

Безусловно, значительно возросшая долговая нагрузка населения требует вмешательства ЦБ РФ (Центральный банк Российской Федерации), т.е. осуществления более жёсткого контроля за выдачей кредитов. Таким образом, банки фактически будут вынуждены разработать **более тщательные и совершенные подходы к рассмотрению кредитных заявок**. Следует отметить, что разрабатываемая в рамках данного проекта система предиктивной аналитики как раз предлагает оценку кредитоспособности заёмщика на полностью обезличенных данных (для снижения влияния человеческого фактора на принятие решений по заявке и исключения нерелевантных признаков) с использованием передовых методов машинного обучения и, соответственно, является весьма **актуальной**.

[3] Долговая нагрузка - процентное отношение обязательных платежей по кредитам к располагаемому доходу домохозяйства.

Основные подходы к оценке кредитных рисков на рынке[4]

На сегодняшний день в кредитном скоринге активно применяются методы регрессионного анализа, деревья решений, а также нейронные сети.

Линейная и логистическая регрессия используются в задачах ранжирования заёмщиков. Преимуществами методов регрессионного анализа является их меньшая, в сравнении с другими методами классификации, чувствительность к размеру обучающей выборки, а также к соотношению плохих и хороших классов (рисков) в обучающей выборке. Однако регрессионный анализ уступает некоторым другим методам в эффективности.

Деревья решений также широко применяются в задачах кредитного скоринга. Достоинствами данного метода являются хорошая интерпретируемость, а также возможность уделить меньшее внимание подготовке данных для модели (необязательно заполнять пропуски в данных и нормировать признаки). Среди недостатков решающих деревьев можно выделить высокую вычислительную сложность алгоритма, а также склонность к переобучению.

Нейронные сети все активнее применяются банками при оценке кредитного риска заемщика, поскольку часто превосходят традиционные статистические модели в качестве и скорости прогнозирования. Однако такие модели требуют тщательно подготовленных и качественных данных, поскольку эффективность обучения модели заметно снижается, если в данных присутствуют нерелевантные атрибуты или датасет имеет недостаточный размер. Кроме того, модели, построенные с использованием нейросетей, долго обучаются и их достаточно сложно интерпретировать.

[4] На основе информации с интернет-ресурса: https://craftappmobile.com/obzor-metodov-kreditnogo-skoringa/#i-3

Первый раздел – Обследование

Кредитный конвейер as-is

Кредитный конвейер as-is исследуемого коммерческого банка включает три основные взаимосвязанные системы: BI-система (система Business Intelligence), Система принятия решений и Интеграционная система. BI-система отвечает за наглядное и оптимальное для бизнеса представление данных. С данной системой работают специалисты кредитного департамента банка. Система принятия решений непосредственно осуществляет процедуру кредитного скоринга на основе полученной личной информации (возраст, пол, доход, адрес, имущество и другие характеристики) о заёмщиках с использованием алгоритмов машинного обучения. Интеграционная система служит для оптимального хранения и эффективного извлечения полученных данных о заёмщиках.

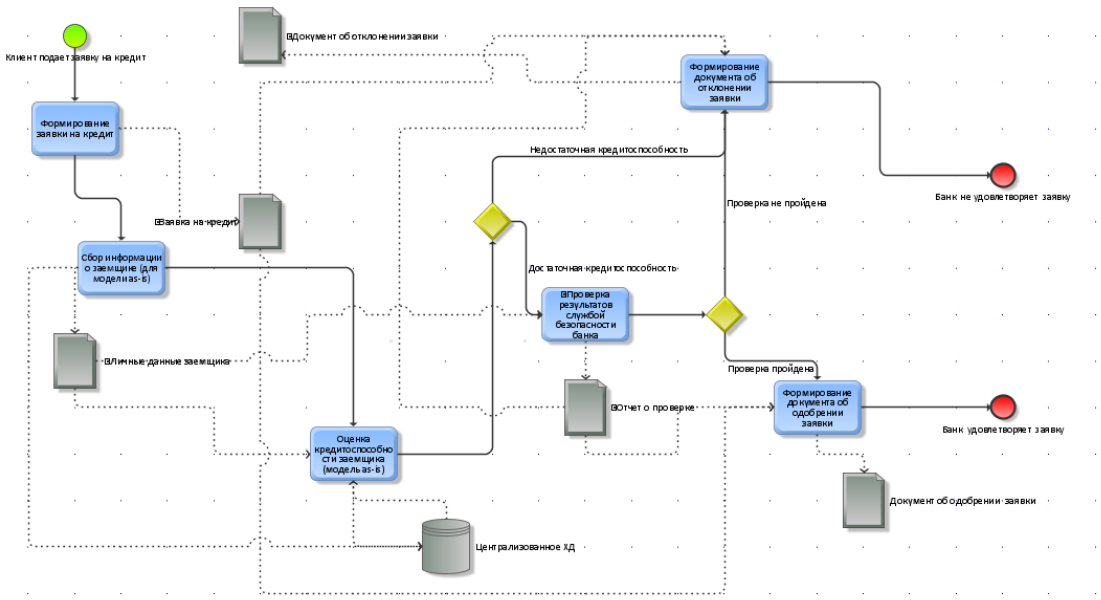


Схема 1 «Кредитный конвейер as-is»

Работа кредитного конвейера as-is начинается с подачи клиентом заявки на получение кредита. Далее кредитный специалист банка формирует заявку на кредит. Затем происходит сбор личной информации о заёмщике, которая необходима для корректной работы модели машинного обучения as-is. Собранные данные вносятся в централизованное хранилище данных кредитного департамента.

Далее происходит автоматическая оценка кредитоспособности заёмщика с использованием модели машинного обучения as-is. Данные для модели берутся из централизованного хранилища данных (ЦХД). Прогноз модели, соответственно, также поступает в ЦХД.

Если кредитоспособность заемщика оказывается достаточной, то результаты оценки кредитоспособности проверяются сотрудниками службы безопасности банка. Если проверка проходит успешно, то сотрудники кредитного департамента банка формируют документ об одобрении заявки, т.е. банк удовлетворяет заявку. Если проверка проходит неудачно, то сотрудники кредитного департамента банка формируют документ об отклонении заявки, т.е. банк не удовлетворяет заявку.

Если же по результатам оценки кредитоспособность заемщика окажется недостаточной, сотрудники кредитного департамента банка формируют документ об отклонении заявки, т.е. банк не удовлетворяет заявку.

К недостаткам кредитного конвейера as-is является малоэффективная система принятия решений: построенные модели кредитного скоринга на основе имеющихся личных данных о заёмщиках имеют достаточно среднее качество. То есть риски ошибочной классификации заёмщиков, а, следовательно, несения убытков в дальнейшем, остаются значительными.

Кредитный конвейер to-be

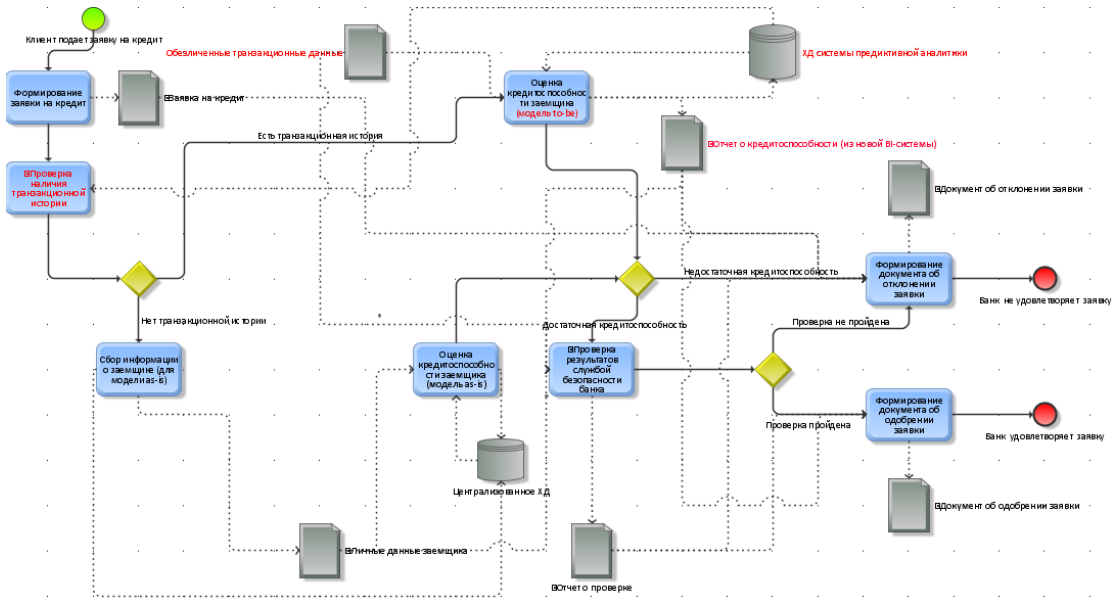


Схема 2 «Кредитный конвейер to-be»

Работа кредитного конвейера to-be начинается с подачи клиентом заявки на получение кредита. Далее кредитный специалист банка формирует заявку на кредит. Затем происходит проверка наличия транзакционных данных заёмщика (необходимых для корректной работы модели to-be) в хранилище данных (ХД) системы предиктивной аналитики.

Если транзакционная история присутствует, то далее происходит автоматическая оценка кредитоспособности заёмщика с использованием модели машинного обучения to-be. Транзакционные данные берутся из ХД системы предиктивной аналитики. Прогноз также поступает в ХД системы предиктивной аналитики. Кроме того, происходит формирование отчёта о кредитоспособности заёмщика посредством нового BI-интерфейса.

Если кредитоспособность заемщика оказывается достаточной, то результаты оценки кредитоспособности проверяются сотрудниками службы безопасности банка. Если проверка проходит успешно, то сотрудники кредитного департамента банка формируют документ об одобрении заявки, т.е. банк удовлетворяет заявку. Если проверка проходит неудачно, то сотрудники кредитного департамента банка формируют документ об отклонении заявки, т.е. банк не удовлетворяет заявку.

Если же по результатам оценки кредитоспособность заемщика окажется недостаточной, сотрудники кредитного департамента банка формируют документ об отклонении заявки, т.е. банк не удовлетворяет заявку.

Если транзакционная история отсутствует, то далее происходит автоматическая оценка кредитоспособности заёмщика с использованием модели машинного обучения as-is. Данные для модели берутся из ЦХД. Прогноз модели, соответственно, также поступает в централизованное ЦХД.

Если кредитоспособность заемщика оказывается достаточной, то результаты оценки кредитоспособности проверяются сотрудниками службы безопасности банка. Если проверка проходит успешно, то сотрудники кредитного департамента банка формируют документ об одобрении заявки, т.е. банк удовлетворяет заявку. Если проверка проходит неудачно, то сотрудники кредитного департамента банка формируют документ об отклонении заявки, т.е. банк не удовлетворяет заявку.

Если же по результатам оценки кредитоспособность заемщика окажется недостаточной, сотрудники кредитного департамента банка формируют документ об отклонении заявки, т.е. банк не удовлетворяет заявку.

Таким образом, необходимо разработать новую систему принятия решений. Разрабатываемая система будет работать с обезличенными транзакционными данными заёмщиков, что поможет получить меняющееся с течением времени представление о благосостоянии заёмщиков и, следовательно, позволит повысить качество построенной модели. Поскольку характеристики нового датасета с транзакционными данными будут значительно отличаться от прежнего набора данных с личными данными, предполагается также разработать новую интеграционную систему (оптимальное хранение и извлечение данных) и BI-систему (оптимальное представление данных).

Исходный набор переменных

Важной составляющей работы любого современного банка является ответственное кредитование, основанное на модельной оценке вероятности того, что заёмщик перестанет выполнять взятые обязательства (выйдет в дефолт). С этой целью проводится анализ больших массивов данных, в ходе которого исследователи пытаются выявить возможные закономерности, предсказывающие позитивные и негативные исходы.

Исходный датасет предлагает оценить вероятность того, что клиент выйдет в дефолт, основываясь на истории потребительского поведения по карточным транзакциям.

Каждая такая транзакция содержит информацию о сумме покупки, месте, дате, mcc-категории, валюте и признаки от платежной системы. Все графики построены на основе случайной выборки, представляющей 10% от исходных данных.

app_id - идентификатор заявки. заявки пронумерованы так, что более поздним заявкам соответствует более поздняя дата, количественная переменная

amnt - нормированная сумма транзакции. 0.0 - соответствует пропущенным значениям, количественная переменная

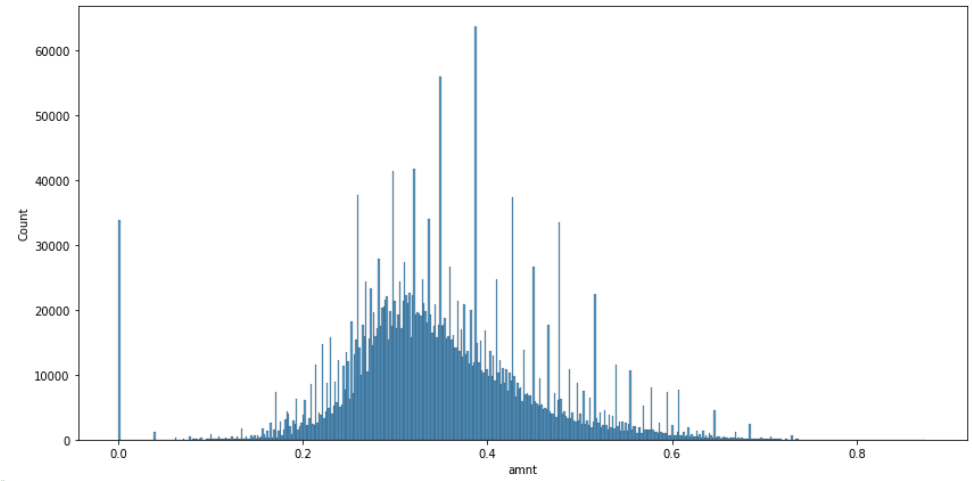


График 1 «Распределение amnt»

Для данной переменной чётко видно, что пик её распределения смещён сильно левее 0.5, что потенциально может создать проблемы для тех желающих получить кредит, кто в своих операциях управляет достаточно большими объёмами денежных средств. При этом стоит выделить большое количество отдельных частотных пиков, которые привязаны к стандартным «круглым» суммам. Эти пики тоже складываются в своеобразное распределение, вершина которого находится правее, чем пик распределения основных данных.

currency - идентификатор валюты транзакции, категориальная переменная

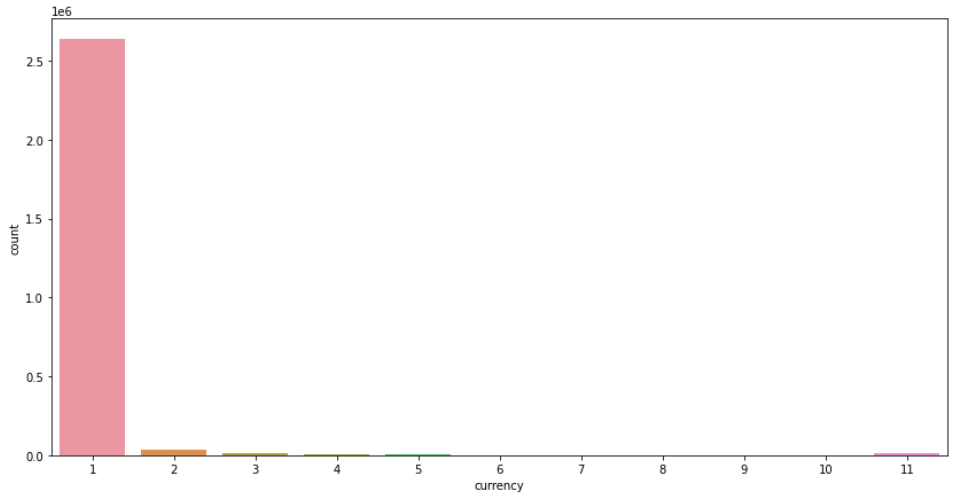


График 2 «Распределение currency»

Т.к. объектом исследования в этой работе является крупный российский банк, абсолютное большинство операций в нём идут в рублях, из-за чего прогнозирование кредитных продуктов в иных валютах (особенно 6-10) может быть менее точным и значительно более рискованным.

operation_kind - Идентификатор типа транзакции, категориальная переменная

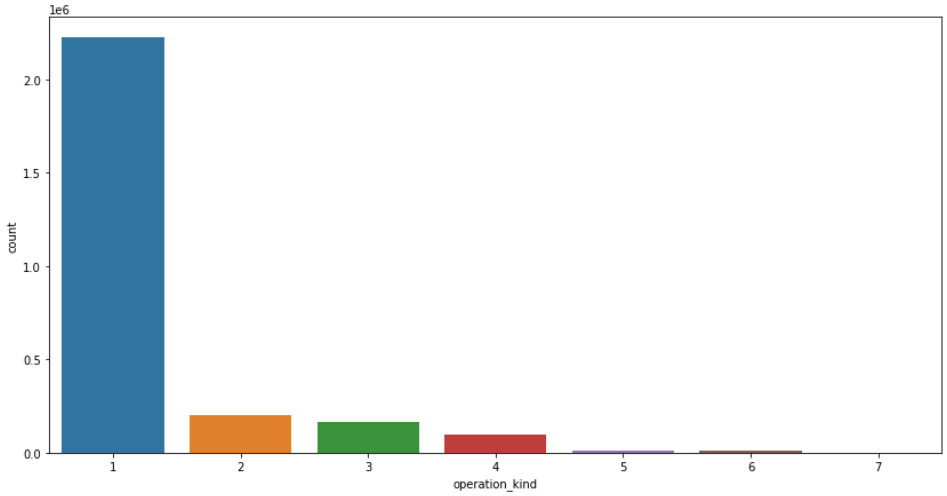


График 3 «Распределение operation_kind»

Наиболее частым типом транзакции является списание денежных средств, что может сильно помочь в построении модели, т.к. именно расходы могут дать нам понимание того, как человек привык тратить свои деньги. При этом количество ещё 3 других типов транзакций тоже достаточно позитивно может сказаться на моделировании, т.к. они представлены в достаточном числе. С 5-7 при этом скорее всего возникнут проблемы.

card_type - Уникальный идентификатор типа карты, категориальная переменная

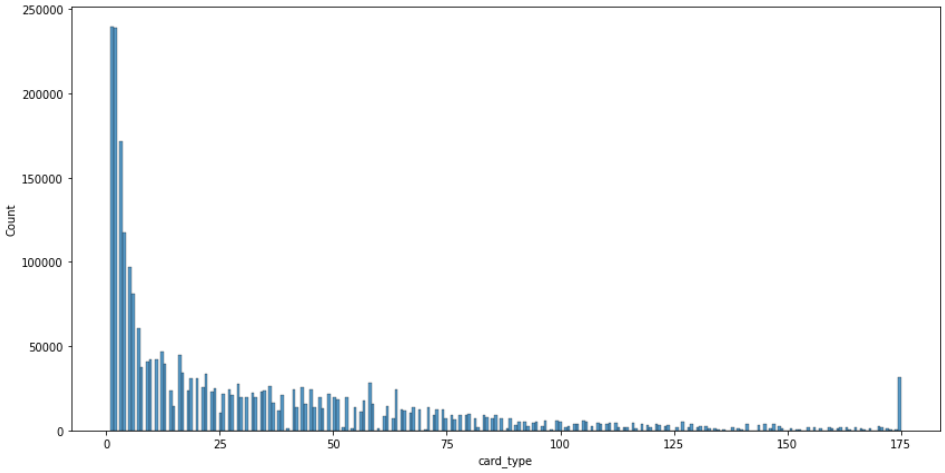


График 4 «Распределение card_type»

Хорошо видно, что в операциях участвует большое количество карт исследуемого банка, а также других банков нашей страны и мира. Хвост за 75-й транзакцией способен создать проблемы при моделировании, однако типы до неё должны помочь в подготовке хорошего прогноза.

operation_type - Идентификатор типа операции по пластиковой карте, категориальная переменная

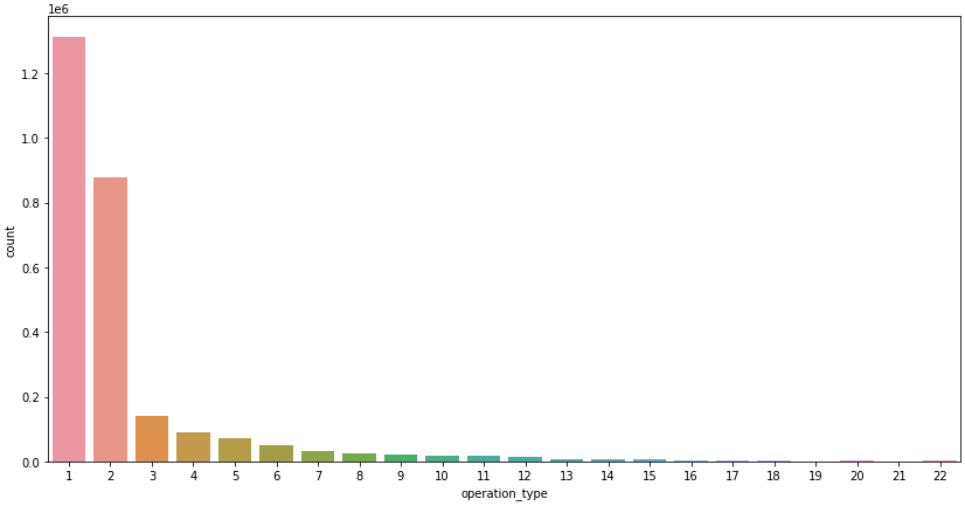


График 5 «Распределение operation_type»

Чётко видна основная доля операций в первый двух типах, остальные 20 могут создать проблемы при предсказании вероятности дефолта.

operation_type_group - Идентификатор группы карточных операций, например, дебетовая карта или кредитная карта, категориальная переменная

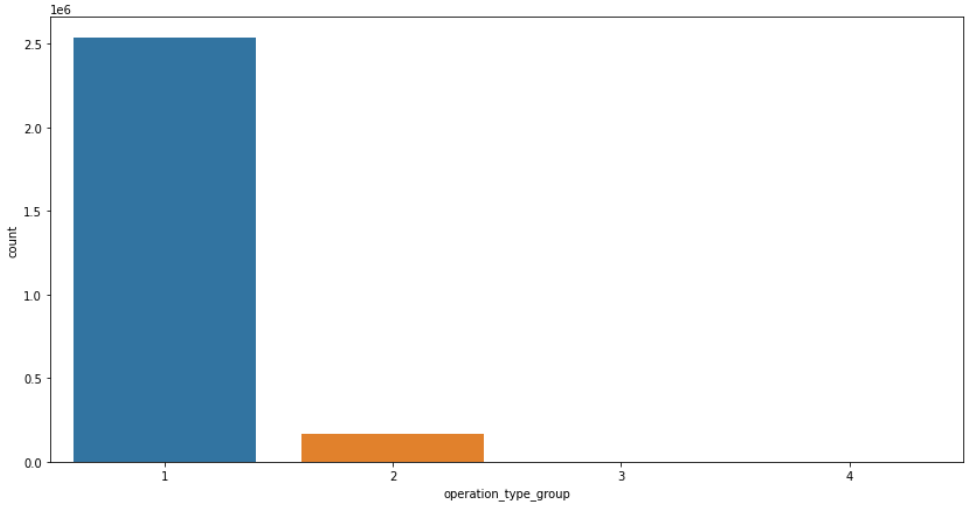


График 6 «Распределение operation_type_group»

Абсолютное большинство операций проводится через дебетовые карты, что абсолютно характерно для современного банковского рынка. При этом операций по кредитным картам менее 10 процентов, что не очень хорошо для нашей системы, т.к. было бы несколько проще её строить, обладая большими знаниями именно по данной группе.

ecommerce_flag - Признак электронной коммерции, бинарная переменная

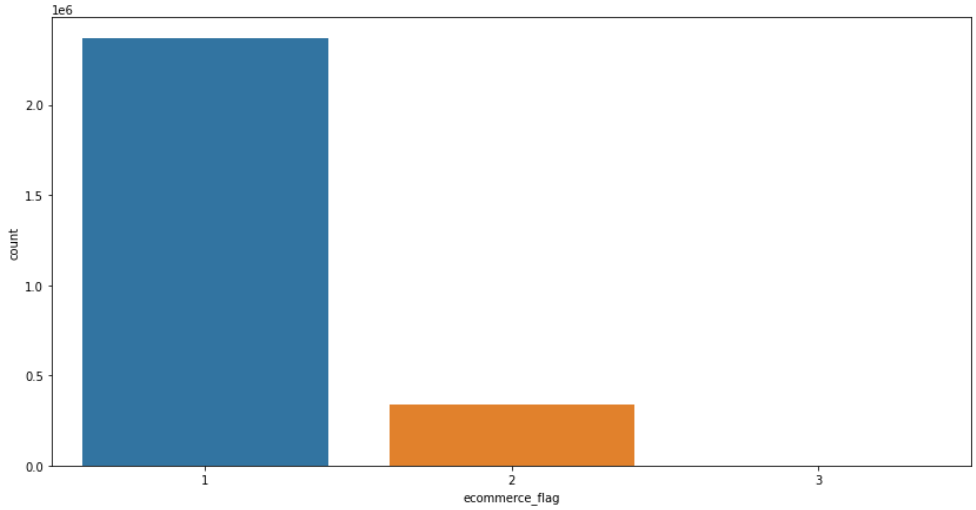


График 7 «Распределение ecommerce_flag»

Большая часть операций проводится через электронную коммерцию, однако не через неё идёт всё ещё достаточно большая часть транзакций, что не должно создать серьёзного перекоса в модели.

payment_system - Идентификатор типа платежной системы, категориальная переменная

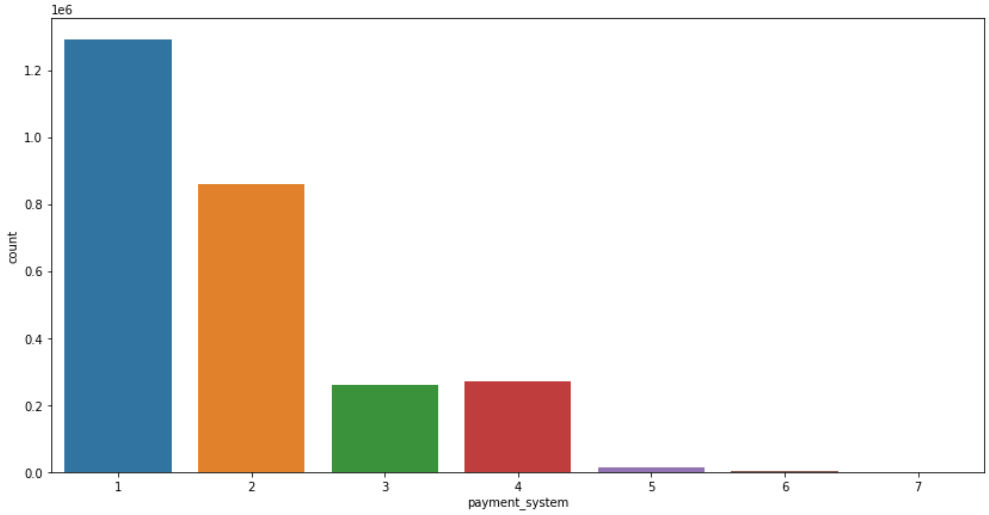


График 8 «Распределение payment_system»

Большая часть операций банка идёт через платёжную систему Visa, т.к. она является базовой для большинства карт в нём. При этом операций через Mastercard и иные основные системы так же представлены в достаточно удобной пропорции, которая позволит адекватно обучить модель на представленных данных.

income_flag - Признак списания/внесения денежных средств на карту, категориальная переменная

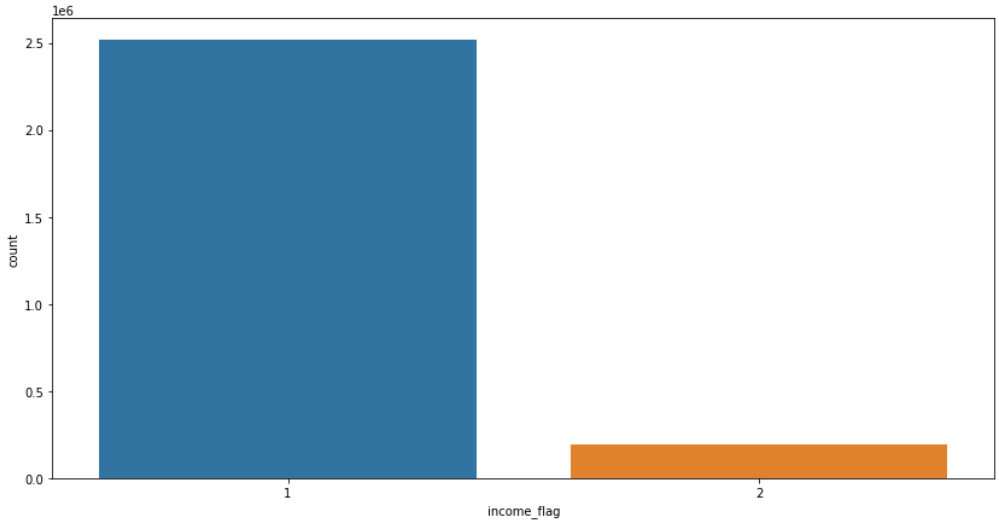


График 9 «Распределение income_flag»

Более укрупнённая версия Графика 3 по типам транзакции, где все имеющиеся типы были собраны в две по факту списания или внесения денежных средств. Здесь подтверждается предыдущий вывод о доминировании списаний, т.к. обычно они затрагивают значительно меньшую сумму, чем зачисления, когда мы говорим об операциях обычных банковских клиентов.

mcc - Уникальный идентификатор типа торговой точки, категориальная переменная

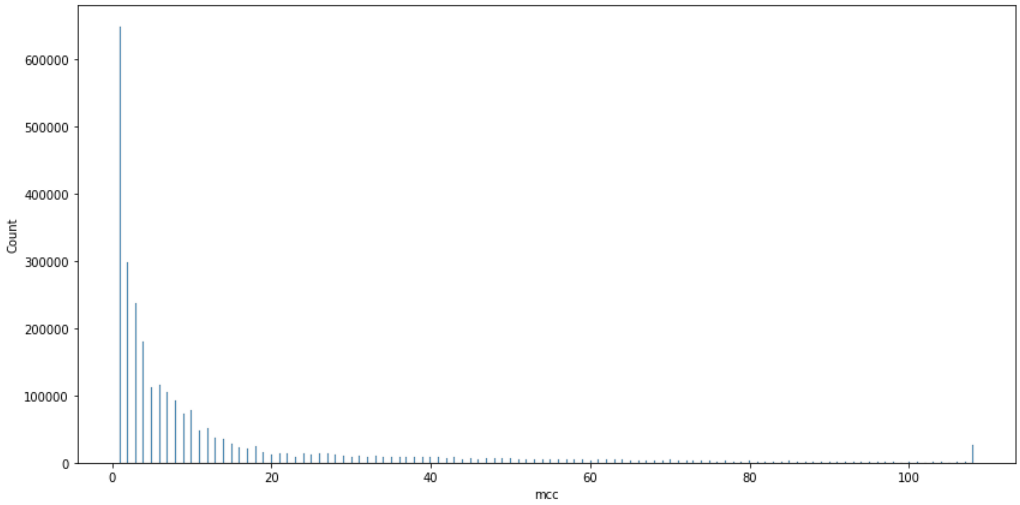


График 10 «Распределение mcc»

Сборщики данных подобрали достаточно большое количество различных типов торговых точек, что с одной стороны должно помочь разобраться с уникальными случаями, а с другой стороны создаст значительные проблемы при работе с типами, у которых очень малое представительство в датасете.

country - Идентификатор страны транзакции, категориальная переменная

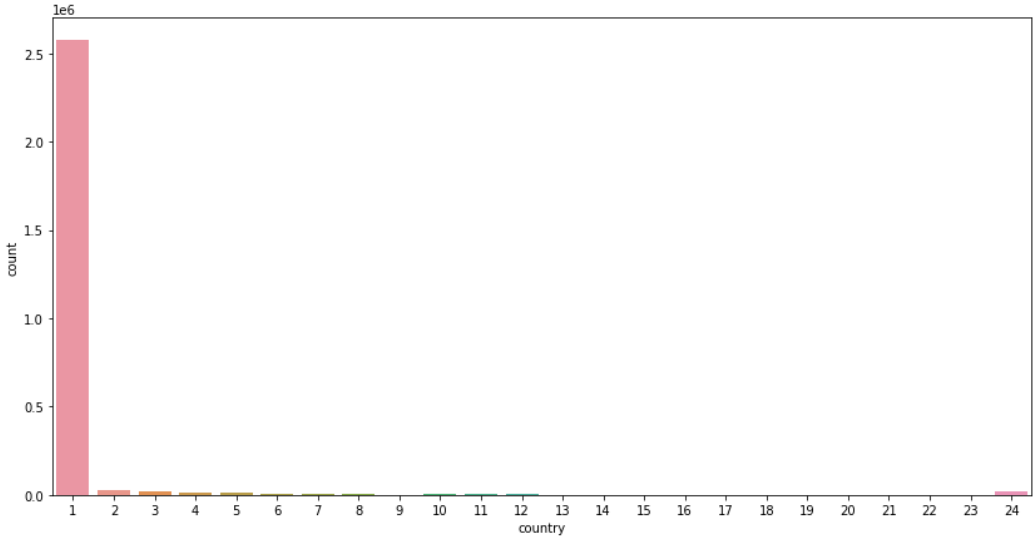


График 11 «Распределение country»

Вместе с участием большого количества валют мы можем наблюдать и большое разнообразие в странах транзакций, однако их количество за пределами РФ критически мало, что может не лучшим образом сказаться на качестве модели. Но может получиться и так, что в исходных данных есть страны, транзакции из которых будут исключительно надёжны (дорогие страны, до которых могут добраться только богатые люди, значительно реже уходящие в дефолт).

city - Идентификатор города транзакции, категориальная переменная

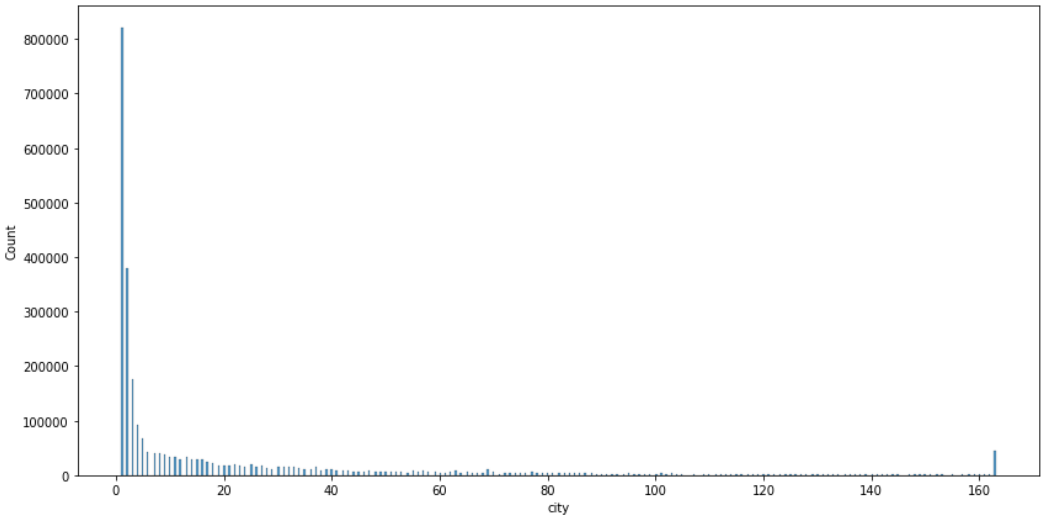


График 12 «Распределение city»

Здесь наблюдается картина, которая очень похожа на График 10, однако в этом случае появляется ещё большее число наблюдения, которые

оказываются в хвосте распределения. При этом часть из них явно находится за пределами РФ, что проистекает из знаний, полученных при помощи Графика 11.

mcc_category - Идентификатор категории магазина транзакции, категориальная переменная

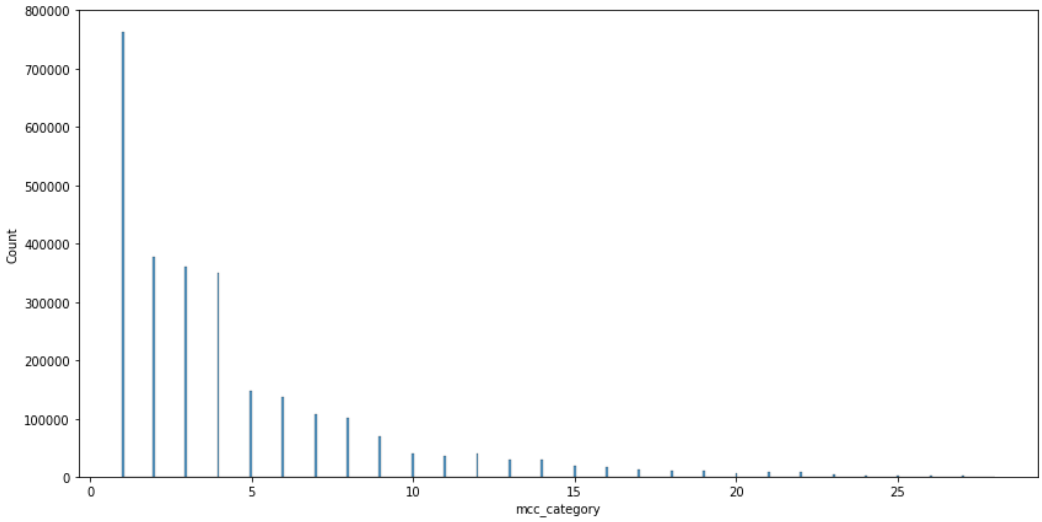


График 13 «Распределение mcc_category»

Более укрупнённая версия Графика 10, где торговые точки транзакций были перераспределены по большим категориям, что дало некоторое сокращение длины хвоста с малым числом наблюдений. При этом пиковые наблюдения удалось несколько сгладить, что должно упростить построение модели.

day_of_week - День недели, когда транзакция была совершена, категориальная переменная

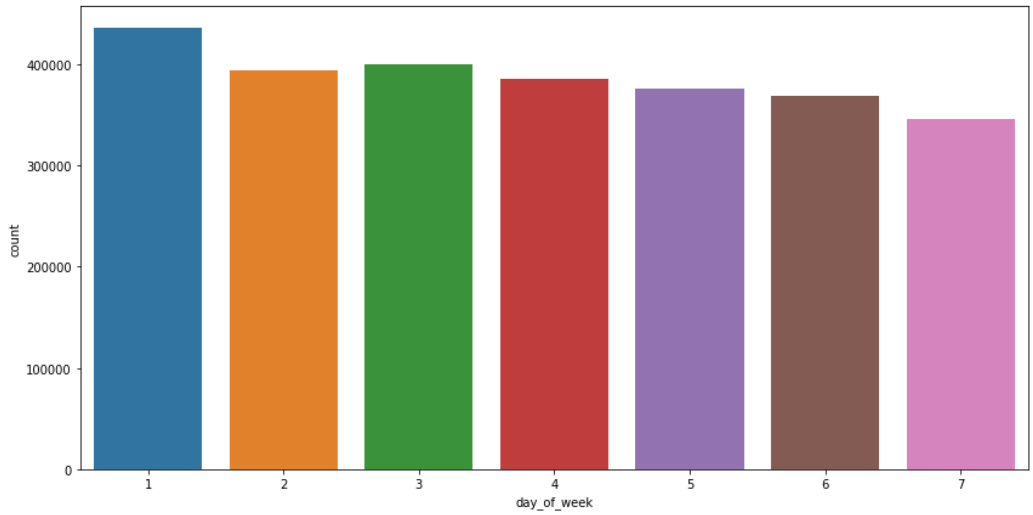


График 14 «Распределение day_of_week»

В течение недели наиболее транзакционным днём является понедельник, после чего начинается постепенное убывание числа транзакций, которое заканчивается в воскресенье сокращением более чем на 10% по отношению к началу недели.

hour - Час, когда транзакция была совершена, количественная переменная

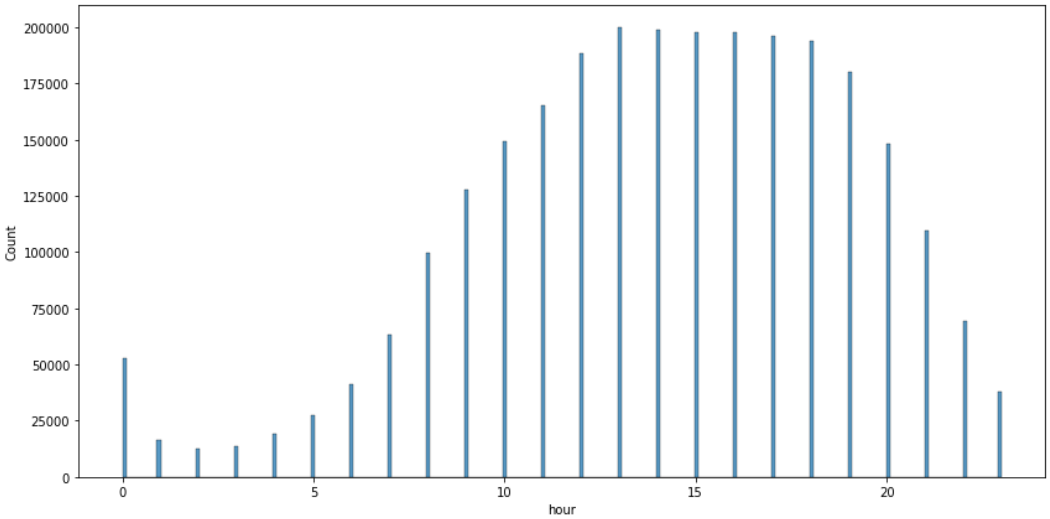


График 15 «Распределение hour»

График распределения по часам выглядит вполне в соответствии с ожиданиями команды аналитики, однако стоит выделить 0 часов, когда мы видим резкий скачок числа операций, которые превышают стоящие перед ними 23:00-23:59. Скорее всего это можно объяснить большим числом запланированных транзакций, которые срабатывают именно в 0:00.

days_before - Количество дней до даты выдачи кредита, количественная переменная

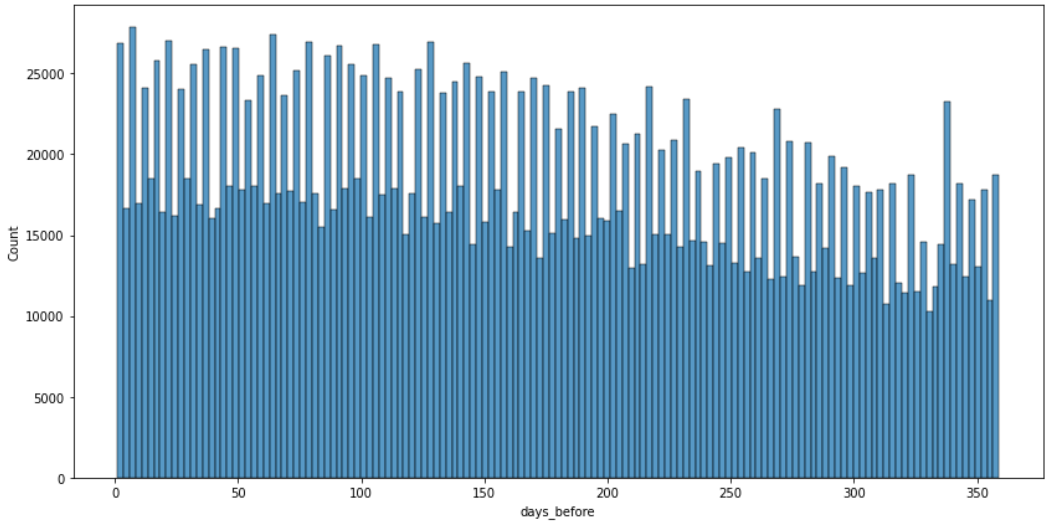


График 16 «Распределение days_before»

На графике числа дней, которые остались до момента подачи заявки явно видно, что ближе к дате выдачи количество транзакций начинает постепенно расти.

weekofyear - Номер недели в году, когда транзакция была совершена, количественная переменная

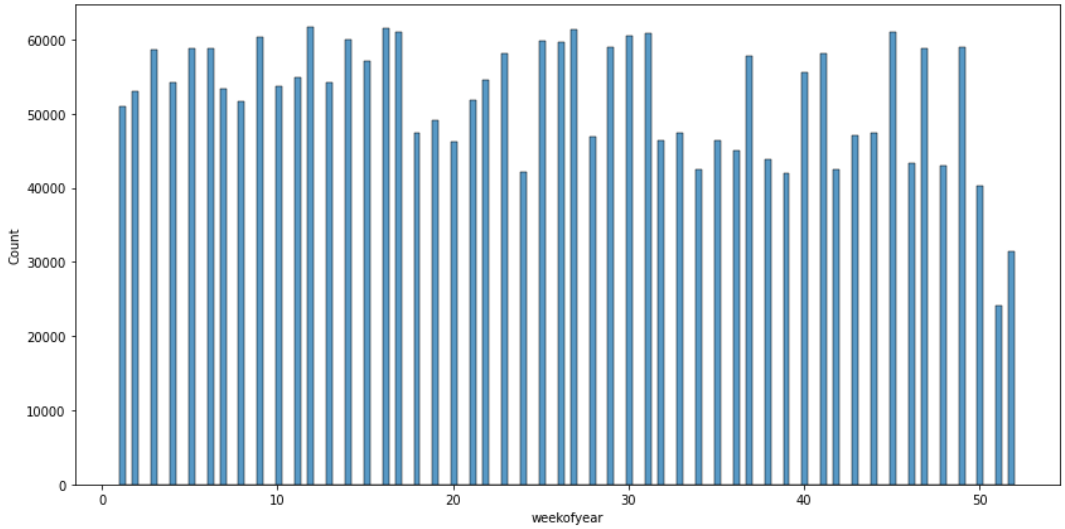


График 17 «Распределение weekofyear»

По этому графику можно заметить, что люди значительно более активно совершают транзакции в начале года и на отдельных неделях летом, чем ближе к концу года, когда активность начинает проседать на 10-15%.

hour_diff - Количество часов с момента прошлой транзакции для данного клиента, количественная переменная

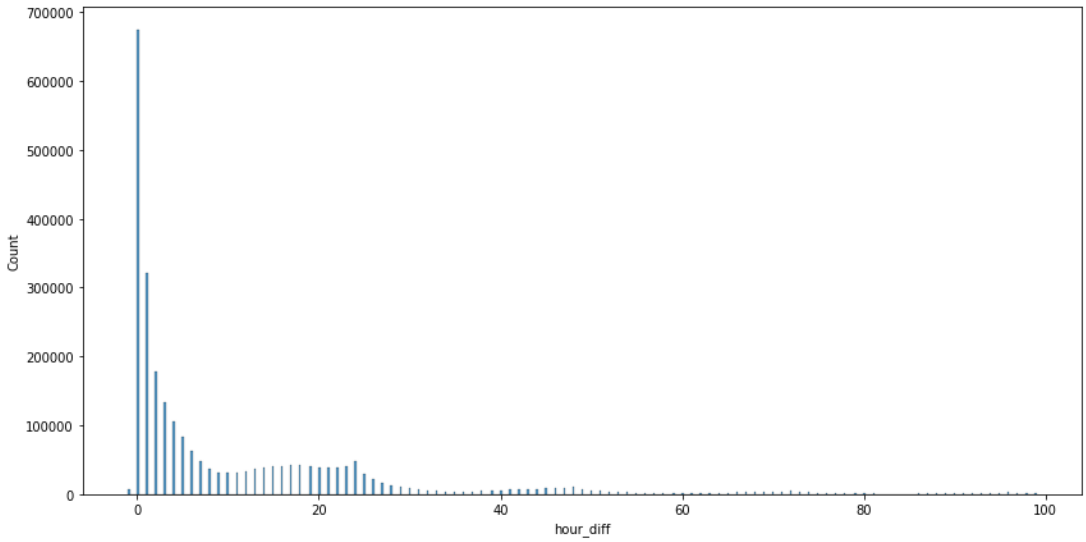


График 18 «Распределение hour_diff»

На представленном графике чётко видно две вещи: большинство операций совершаются потоком с разницей меньше часа, операции по отдельным картам проходят чётко через некоторое количество дней (24, 48, 72 или 96 часов), что может немного помочь с предсказанием риска по отдельным заявкам.

transaction_number - Порядковый номер транзакции клиента, порядковая переменная

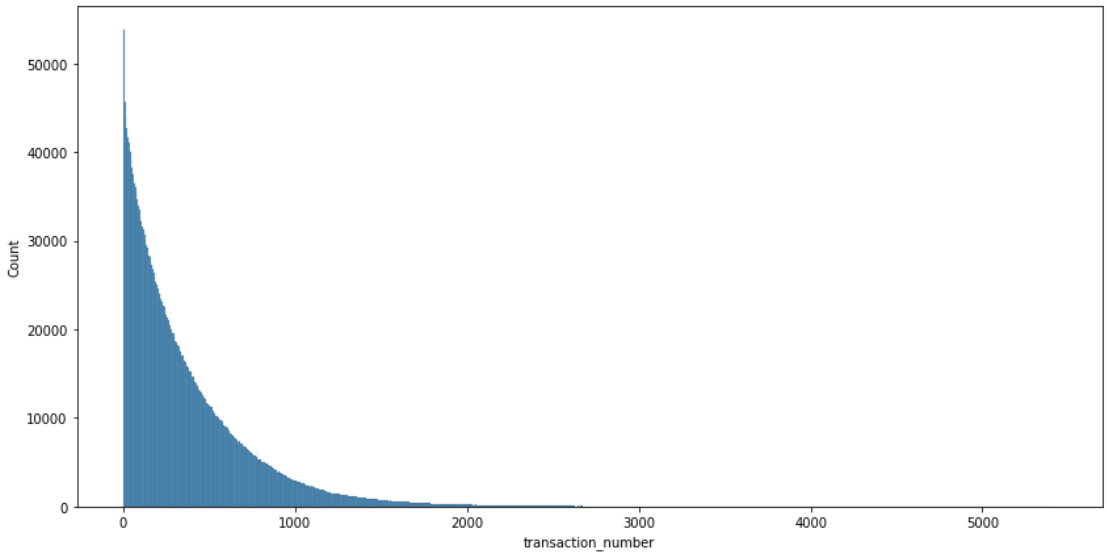


График 19 «Распределение transaction_number»

По графику номеров транзакций хорошо видно, что большая их часть укладываются в первые две тысячи, дальше которых заёмщик обычно не

заходит, т.к. на достижение этого порога уходит уже достаточно большое количество времени, а клиенты банка всё же не ждут годами, чтобы попытаться получить кредитный продукт.

product - Продукт по которому нужно принять решение, уйдет ли заявитель в дефолт или нет, категориальная переменная

flag - Целевая переменная, 1 - факт ухода в дефолт, бинарная переменная

Предполагаемые наиболее значимые предикторы и целевые переменные

Предполагается, что наиболее важными переменными в моделях окажутся:

- порядковый номер транзакции, т.к. чем дольше один и тот же человек пользуется кредитными продуктами компании, тем больше к нему доверия;
- тип кредитного продукта, т.к. по разным кредитным продуктам мы можем наблюдать различный уровень риска. Например, риск по кредитке значительно меньше, чем по ипотеке из-за несопоставимых размеров кредита;
- число часов с момента последней транзакции, т.к. клиенты, которые часто пользуются услугами нашего банка, скорее всего более склонны не уйти в дефолт при выплате кредита;
- категория магазина покупки, т.к. в магазинах низкой категории с продукцией плохого качества платёжеспособные клиенты, как правило, закупаются достаточно редко.

Остальные переменные скорее всего так же окажут определённое влияние, однако на данный момент трудно сказать, насколько сильным оно окажется.

Качество данных

Целевой переменной является то, уйдёт ли данный кредитный продукт в дефолт, которая привязывается к конкретному идентификатору заявки и кредитному продукту.

Всего в представленном датасете 1 миллион строк. 900 тысяч выделены под обучающую выборку. Оставшиеся 100 тысяч строк являются тестовой выборкой, которая и будет использоваться для оценки качества нашей модели. Тестовая выборка хронологически идёт после обучающей и валидационной.

Данные не имеют пропусков и выбросов, так как были отфильтрованы и нормированы заранее, благодаря чему команде разработки не придётся заниматься улучшением их качества.

Ожидания от модели

Основными метриками нашей модели были выбраны AUC-ROC, т.к. он позволяет качественно оценить точность предсказания вероятности ухода кредитного продукта в дефолт, учитывая внутри себя ряд других более простых метрик, F1 Score, т.к. это одна из наиболее широко используемых метрик для классификации. На основе выбранных метрик ожидается, что ансамбль из нескольких моделей разных типов (градиентный бустинг через три библиотеки (LightGBM, XGBoost, CatBoost) и нейросеть) сможет получить значение AUC-ROC и F1 Score выше 0,75.

Второй раздел – Формализация требований

Требования к архитектуре приложения BI

Функциональные требования:

1. Качественный UI/UX интерфейс. Понятные пользователю и простые во взаимодействии схемы. Дизайн информационного монитора минималистичен и соответствует современным канонам.
2. Понятность для конечного пользователя. При использовании BI-интерфейса у клиента не должно возникать никаких функциональных вопросов.
3. Простота установки и запуска. Формат web-страницы был выбран в связи с тем, что в данном случае пользователю не требуется устанавливать лишнего программного обеспечения и проводить никаких дополнительных настроек. Все, что нужно для начала использования BI-интерфейса – просто зайти на страницу в интернете.

Технологические требования:

1. Архитектурная независимость. BI-приложение должно быть реализовано как отдельное web-приложение, которое самостоятельно взаимодействует с данными из хранилища. Таким образом, при внесении изменений в модель или же в хранилище это никак не отразится на работе BI-интерфейса.
2. Непрерывность взаимодействия. BI-приложение должно быть доступно для взаимодействия и обработки данных в любое время суток, чтобы аналитики всегда имели доступ к данным и инфографике.
3. Модульность. BI-интерфейс должен быть легко перестраиваемым под текущие нужды аналитиков. Необходимо иметь возможность с легкостью добавлять новые и изменять текущие информационные компоненты без внесения крупных изменений в код.
4. Работа с новыми данными. BI-приложение должно быть способно к обработке и трансформации новых поступивших в хранилище данных.

Требования к модели

Функциональные требования:

1. Уровень качества модели - Качество модели на тестовой выборке данных по AUC ROC и F1 Score не должно быть ниже 0.75.
2. Обрабатываемые данные – Модель должна быть способна повторно обучаться на новых данных, не теряя качества прогнозирования.
3. Расходы на работу модели - Разработка, обновление и функционирование модели не должны суммарно требовать большее финансовое обеспечение, чем это прописано в управляющих документах.

Технологические требования:

1. Время изменения параметров модели - Все элементы модели должны быть построены в течение 3 дней после изменения ее параметров.
2. Конечное время построения прогноза - Время построения прогноза не должно превышать 60 минут.
3. Повторная настройка модели из приложения - Параметры модели должны быть доступны для изменения людьми, не обладающими навыками программирования.
4. Вычислительные мощности - Модель по вычислительным мощностям должна умещаться в располагаемые аналитическим отделом банка.
5. Формат данных – Модель должна быть способной к обработке и трансформации любых видов данных, поступающих из других банковских систем.
6. Непрерывность взаимодействия – Модель должна быть доступной для обработки данных в любое время суток, чтобы не создавать очереди и не затягивать процесс принятия решения о предоставлении кредитного продукта.

Требования к хранилищу данных

Функциональные требования:

1. Обеспечение безопасности данных
2. Контроль качества данных
3. Создание и сохранение резервных копий
4. Поддержка согласованности данных
5. Разделение прав доступа к данным для групп пользователей

Технологические требования:

1. Поддержка параллельной работы ядер СУБД (системы управления базами данных) – Необходима для обеспечения быстродействия работы хранилища
2. Поддержка кириллицы – Часть данных имеет текстовую форму представления на русском языке. Существует риск выбора СУБД, не имеющей поддержки русского языка
3. Использование облачных технологий – Слабая инфраструктурная база проекта требует физического расположения хранилища на удаленном сервере
4. Поддержка популярных протоколов для подключения – СУБД должна позволять использовать ODBC-подключения, одного из самых широко использующихся протоколов

Третий раздел – Выбор ИТ-решения

Основными критериями при выборе ИТ-решения для всех задач была простота разработки, бесплатность, возможность работы с большими объёмами данных и совместимость с современными операционными системами.

Модель прогнозирования

Основными языками программирования, на которых возможна разработка решения для подобных задач являются Python 3.x и R. Речь идёт именно о языках программирования, а не об уже готовых статистических пакетах, т.к. они в абсолютном большинстве своём платные, ограниченные по функционалу и не очень популярны на рынке. С точки зрения обеспечения библиотеками Python 3.x и R практически идентичны, т.к. большинство необходимых библиотек подготовки данных и машинного обучения разрабатываются сразу под оба языка.

Однако для разработки был выбран именно Python 3.x на дистрибутиве Anaconda, т.к. вся команда разработки с ним знакома и реализовывала на нём свои проекты ранее. В решении будут использованы библиотеки:

- Pandas
- NumPy
- scikit-learn
- Tqdm
- Pickle
- LightGBM
- XGBoost
- CatBoost
- Keras

Первые 5 для подготовки, обработки и визуализации данных, последние 4 для построения моделей.

BI-приложение

В архитектурном плане ИТ-решение будет построено при помощи

фреймворка Flask. Благодаря этому, при необходимости будет возможно удобно и быстро развернуть его на различных платформах, что добавит гибкости при запуске и позволит подстроиться под требования потенциального заказчика. Модульный тип архитектуры проекта, а также интеграция данных напрямую из облачного хранилища позволит развивать проект без вмешательства в блок моделирования. Также возможно легкое и быстрое подключение к новому проекту, требующему визуализацию своих данных. Формат web-приложения позволяет достичь максимальной гибкости в работе. Таким образом, выбор построения архитектуры ИТ-решения сразу пал на Flask, так как данный фреймворк позволяет в кратчайшие сроки производить переработку кода и адаптацию под нужды клиента.

Для визуализации данных был выбран данный набор библиотек:

- Pandas
- Numpy
- Plotly
- PlotlyExpress

Также интеграция с хранилищем данных происходит на базе библиотеки SQLAlchemy, позволяющей в несколько строк реализовать подключение к удаленной базе данных.

На данных скриншотах можно увидеть примеры страниц из интерфейса BI:

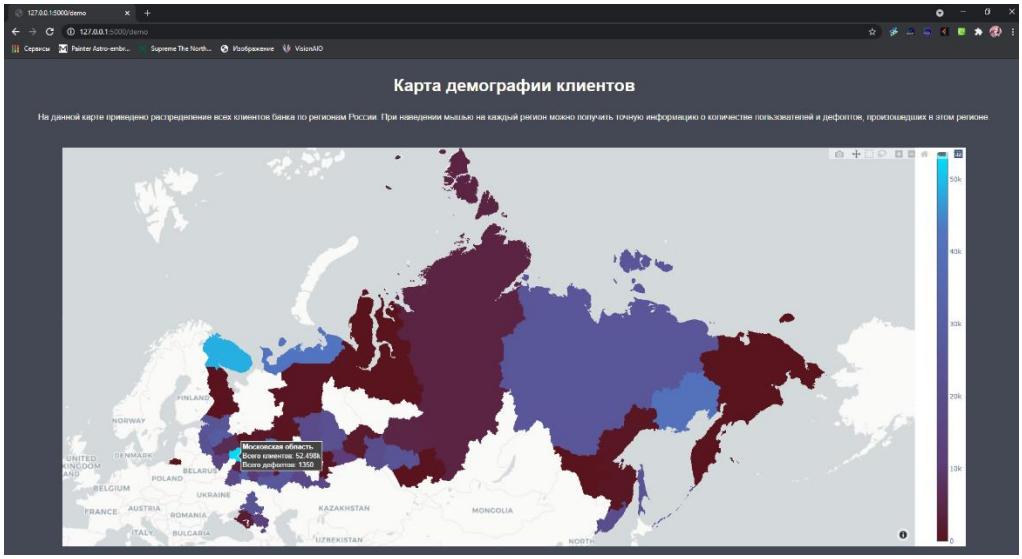


Рисунок 1. Карта демографии клиентов.

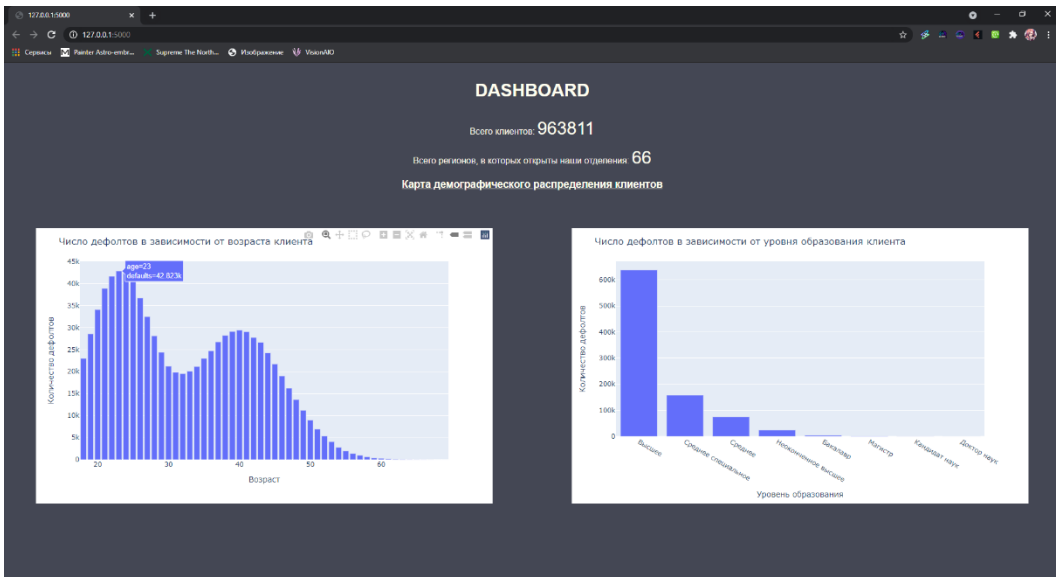


Рисунок 2. Главная страница BI-интерфейса.

Хранилище данных

В рамках проекта предполагается использование облачного хранилища данных для создания прототипа системы. Использование облачных технологий обусловлено необходимостью наличия доступа к хранилищу со стороны всех участников проекта и компонент, за которые они ответственны: моделей и BI-интерфейса.

На начальном этапе проекта предполагается использовать учебный сервер Высшей Школы Экономики, затем, после этой, так называемой, «песочницы», планируется использовать облачное хранилище, предлагаемое компанией Google. Ближайшие аналоги – аналогичные сервисы от Amazon и Яндекса – не предоставляют пользователям условно бесплатный доступ: в Google вновь созданный аккаунт получает некоторую сумму на счет, которую владелец аккаунта может тратить в течение ограниченного времени после регистрации.

В рамках курсовой работы планируется перенос разработанного в облаке хранилище на локальный сервер одного из участников с последующим предоставлением удаленного доступа остальным членам команды разработки в виду того, что сроки выполнения работы превышают длительность пробного периода облачного сервиса.

В качестве СУБД (системы управления базами данных) был выбран Postgres, так как он является одним из самых популярных представителей реляционных СУБД, а потому предоставляет широкий спектр возможностей для установки, подключения и настройки. Бесплатная версия Postgres активно развивается и незначительно уступает платным аналогам. В дальнейшем рассматривается возможность использовать СУБД, поддерживающую MPP (massive parallel processing) – массивные параллельные вычисления. В рамках проекта планируется исследовать возможность использования Google BiqQuery – облачное решение, аналогичное с поддержкой MPP.

В результате предпринятых попыток включения Google BiqQuery в технологический стек нашего проекта команда пришла к выводу, что, несмотря на высокую производительность и довольно дружелюбный к пользователю интерфейс, данная технология не может использоваться в работе ввиду отсутствия ряда критически важных функций таких, как простой способ интеграции хранилища с Python.

В качестве ETL (Extract, Transform, Load) -инструмента был выбран Python, так как, благодаря наличию широкого спектра библиотек, он является гибким и простым в использовании средством, на котором одинаково удобно использовать как для проверки гипотез и разработки, так и для создания промышленных решений. Python позволяет реализовать собственные алгоритмы обработки и загрузки данных, а также кастомизированные отчеты о выполнении работ, сообщения об ошибках. Кроме того, Python является бесплатным для использования продуктом, не требующим лицензии, что упростит его дальнейшее использование в промышленной эксплуатации.

Архитектура системы

С учетом выбранных нами технологий архитектура нашей системы выглядит следующим образом (стрелки показывают потоки данных):

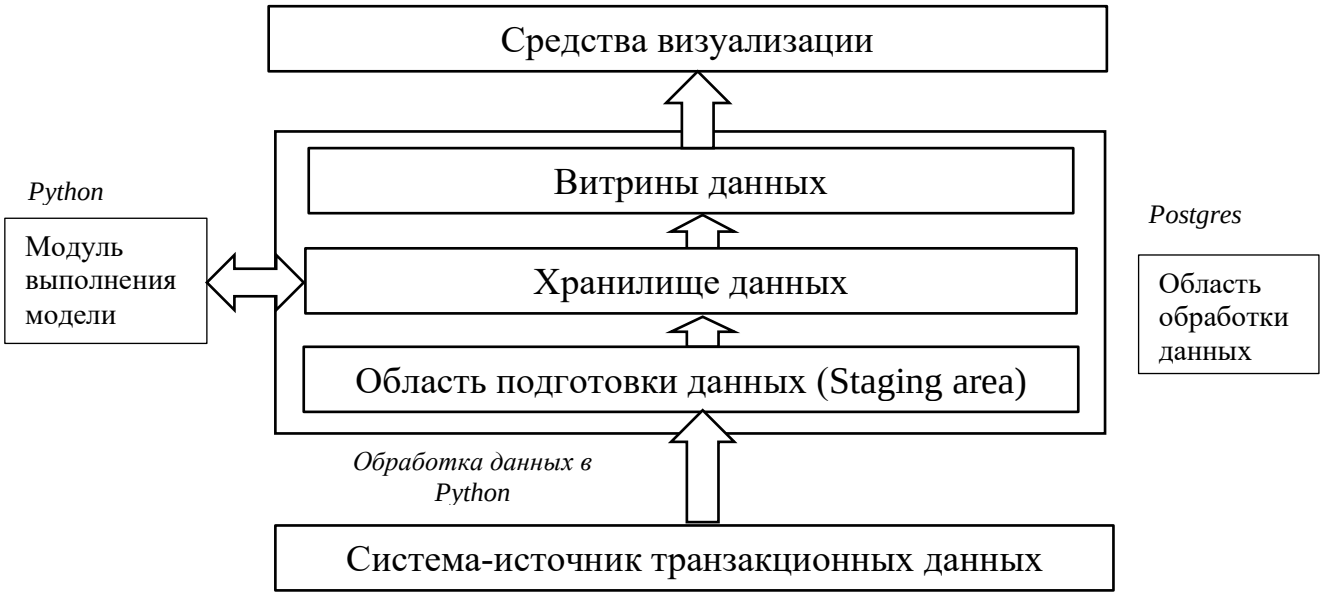


Схема 3 «Архитектура системы в период разработки»

До вывода в промышленную эксплуатацию наша система имеет следующую архитектуру: подготовленный для нашего использования отчет с транзакционными данными, структура которых описана в разделе «Исходный набор переменных», при помощи языка программирования Python загружается в область подготовки данных (staging area), которая находится в СУБД Postgres. В этой же области находится массив персональных данных клиентов банка – заемщиков, в нем находятся следующие атрибуты: ФИО, возраст, регион и город проживания, уровень образования.

Транзакционные данные из staging area при помощи выбранного нами ETL-инструмента - Python - преобразуются в структуры, необходимые для работы ансамбля моделей машинного обучения: для нейросети отдельные транзакции, выполненные одним клиентом, преобразуются во временные ряды произвольной длины, в полях полученной таблицы хранятся массивы с характеристиками каждой транзакции; для моделей градиентного бустинга различные метрики транзакций, относящихся к одной заявке на кредитный продукт, преобразуются в агрегированные показатели: среднее, медиану и другие квантили, минимум, максимум для числовых признаков и различные счетчики для категориальных переменных. Преобразованные данные загружаются в область хранилища данных, к таблицам в которой обращается модуль исполнения моделей для получения входных данных перед построением новых прогнозов.

Модуль выполнения модели с заданной периодичностью (3 дня) обращается к хранилищу для получения новых данных, использующихся для построения новых прогнозов дефолта и обновления старых. Результаты работы ансамбля моделей загружаются обратно в хранилище в новые объекты в том же пространстве хранилища.

Последний слой хранилища данных содержит витрины данных, которые используются для построения дэшбордов и других способов визуализации данных. В рамках работы были созданы следующие витрины:

- основная бизнес-витрина, в которой каждой заявке приписан актуальный прогноз ансамбля моделей - флаг вхождения заявителя в дефолт при одобрении его заявки на кредитный продукт;
- детальная бизнес-витрина, в которой к каждой паре «заявка – прогноз ансамбля моделей» добавлены персональные данные клиента с целью построения отчетов в различных разрезах – по возрасту заявителей, их местоположению, уровню образования;
- витрина для сотрудников кредитного департамента, в которой к каждой паре «заявка – прогноз ансамбля моделей» добавлены персональные данные клиента и все его транзакции для просмотра детальной информации о любом человеке и возможного принятия решения, отличного от предложенного предиктивной системой;
- витрина контроля качества данных, в которой находится информация о загруженных в хранилище данных и сопоставление этих метрик с метриками данных, заявленными поставщиком данных.

BI-интерфейс обращается к витринам данных для обновления информации на дэшбордах.

В качестве целевой архитектуры системы нами предполагается следующая схема:

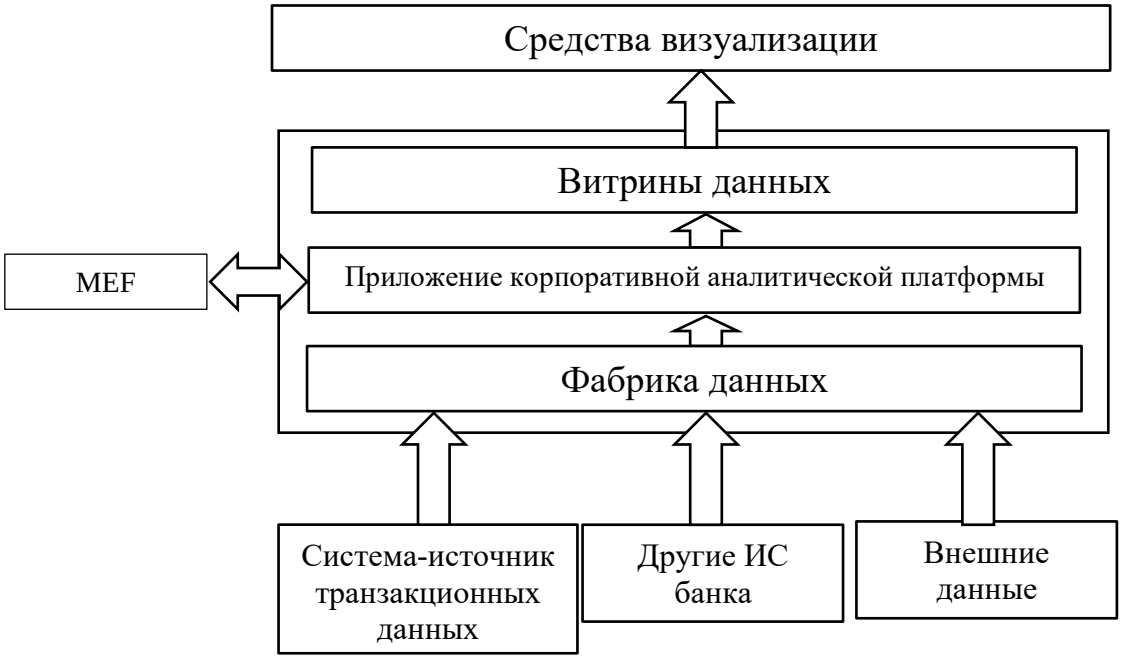


Схема 4 «Целевая архитектура системы»

В виде целевого решения планируется интеграция с существующими в банке системами: Фабрикой данных, ИС-источником транзакционных данных и, потенциально, другими системами. Фабрика данных - единое аналитическое хранилище данных (АХД) банка, в котором находятся данные всех ИС и структурных подразделений банка. Для каждой команды выделяется свое пространство в Фабрике - Приложение корпоративной аналитической платформы (ПКАП), с помощью которого команда получает доступ к Фабрике, а также к MEF (model execution framework) - инструменту для регулярного запуска моделей. Благодаря доступа ПКАП к Фабрике данных, куда попадают данные из всего банка, вопрос интеграции нашей системы с другими сводится или к настройке соединений между различными областями Фабрики - нашим ПКАП и ПКАП других команд, или получению доступа в области других команд в АХД.

В Фабрику также можно самостоятельно загружать данные из внешних источников (при помощи ETL инструментов или средств для crawling`a данных), что может позволить расширить функционал нашей системы в части визуализируемых данных.

В рамках ПКАП помимо витрин, определенных в схеме текущего состояния системы, могут появиться и иные витрины: например, витрина для

performance dialogue - демонстрации потенциальной успешности тех или иных кредитных продуктов в разных регионах с целью построения лучшей маркетинговой кампании там.

Собранные витрины извлекаются из ПКАП и попадают в модуль визуализации.

Четвёртый раздел – Построение модели

Модель прогнозирования

Перед созданием модели важнейшим решением является выбор метода её построения и соответствующая подготовка данных, т.к. разные модели могут требовать кардинально отличающиеся методы. Для первичного отбора были взяты три модели, использующие градиентный бустинг на решающих деревьях (далее GBDT), и одна рекуррентная нейросетевая модель (далее RNN). Исходный выбор двух разных типов моделей обоснован часто используемой в наше время практикой создания ансамблей на основе нескольких базовых моделей[5]. При помощи комбинирования (в нашем случае взвешенного усреднения методом перебора) результатов работы базовых моделей удаётся сократить число выбросов в прогнозах и улучшить общий результат в сравнении с независимыми моделями. Для комбинации лучше всего подходят модели разных типов (GBDT, нейросети, регрессионные модели и др.), т.к. между ними будут наиболее значимые различия, исправление которых сильно улучшит качество прогноза.

Т.к. наиболее продвинутыми на сегодняшний день моделями, целью которых может быть предсказание вероятности дефолта, являются GBDT (благодаря внутреннему ансамблю из множества слабых моделей) и нейросети (для большего разнообразия между моделями была выбрана рекуррентная сеть, принимающая на вход временной ряд любой длины). Главное преимущество возможности получать на вход ряд различной длины заключается в том, что мы можем обучать нейросеть на полном наборе исторических транзакционных данных, связанных с конкретным человеком и его заявкой на кредитный продукт, без необходимости серьёзно менять их формат, однако это потребует много времени на подбор коэффициентов.

Для реализации моделей с использованием GBDT было выбрано три продукта от разных создателей, которые часто встречаются вместе в одном ансамбле: XGBoost (независимая разработка, медленная, но способна давать хорошие прогнозы без длительного подбора параметров), LightGBM (разработка Microsoft, самая быстрая и при этом дающая точные прогнозы) и CatBoost (разработка Yandex, быстрая, но требует точного подбора

[5] toptal.com/machine-learning/ensemble-methods-machine-learning – В приведённом обзоре можно коротко ознакомиться с методикой

параметров). При этом данные для этих моделей потребуют значительно большей предварительной обработки, т.к. они не способны принимать ряды изменяемой длины, которыми по сути и являются получаемые на входе данные. Для работы модели нам потребуется пересчитать ряд агрегированных метрик из набора исходных (среди которых есть и категориальные, и количественные). Категориальные метрики будут трансформированы в число упоминаний конкретного параметра в ряду и в среднее число появлений конкретного параметра в ряду. Количественные метрики преобразуются в среднее, медианное, минимальное и максимальное значение и другие параметры. Всего получено 127 параметров, участвующих в построении моделей.

Обработанные данные поступают в специально подготовленные для этого облачные витрины, откуда они уже попадают в выбранные нами модели машинного обучения. В ходе обучения моделей нам удалось вывести качество их предсказаний на приблизительно схожий уровень, что позволило оставить для финальной системы все выбранные изначально методы построения моделей.

Схема 5 «Схема работы ансамбля моделей[6]»

[6] В модели все усреднённые предсказания подразумевают взвешенное усреднение

После построения прогнозов результаты работы GBDT-моделей формируют свой ансамбль методом перебора лучшей комбинации по выбранным метрикам, после чего создаётся второй ансамбль с учётом RNN-модели. Создание ансамбля разделено на два этапа, т.к. важной частью построения модели является оценка вклада отдельных базовых GBDT-моделей в общее качество их ансамбля. При этом по приведённым ниже графикам можно заметить, что для разных моделей сформированные нами переменные имеют разные веса, лишний раз подчёркивая обоснованность использование трёх схожих моделей в нашей работе. При этом несмотря на то, что в ряде моделей отдельные переменные оказывают минимальное влияние на итог, мы не собираемся их исключать, т.к. при работе модели оценки кредитного риска значительно важнее точность модели, чем сокращение времени её работы. К тому же, в нашей системе будет предусмотрен расчет рисков выдачи каждого кредитного продукта для действующих клиентов банка заранее, до их фактического обращения за тем или иным продуктом, что позволит сократить длительность процесса обработки заявок клиентов.

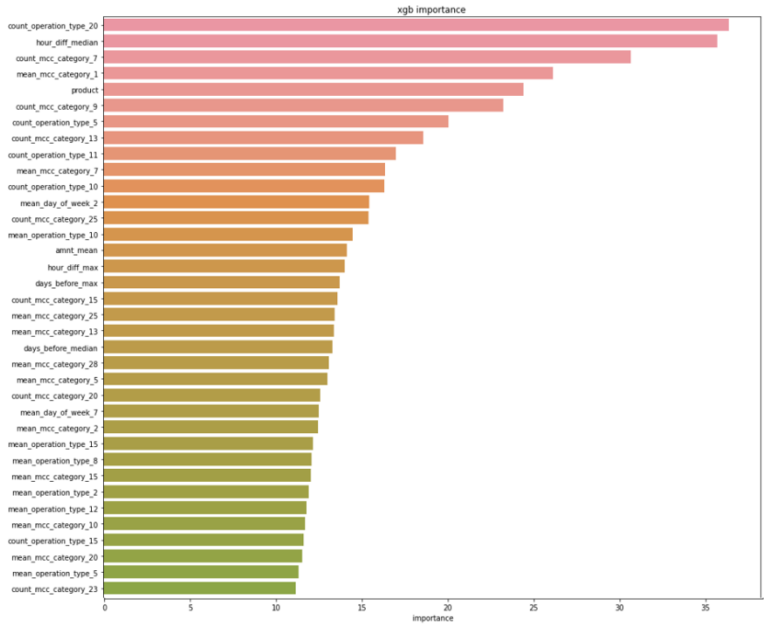


График 20 «Наиболее важные параметры в XGBoost-модели»

На общем фоне XGBoost-модель показывает наиболее сбалансированное использование переменных при разделении вершин, что повышает эффективность работы модели и лишний раз обосновывает использование столь большого набора параметров.

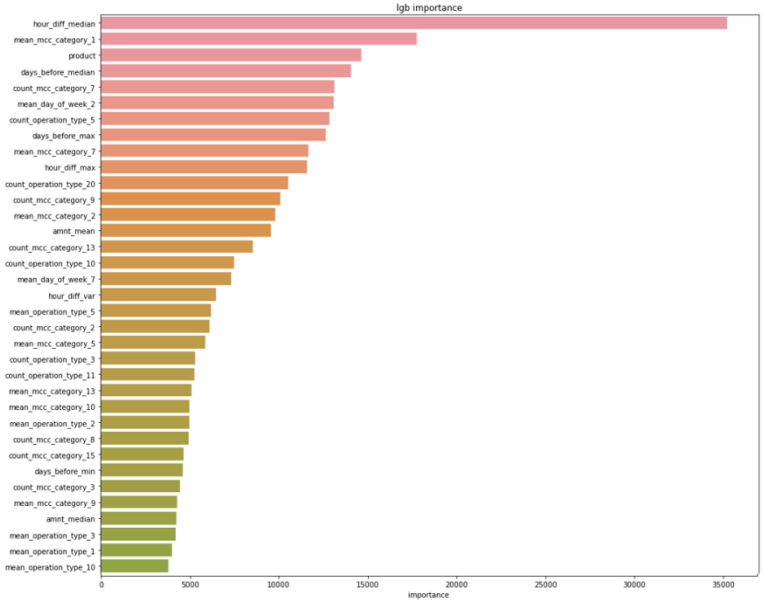


График 21 «Наиболее важные параметры в LightGBM-модели»

В LightGBM-модели можно наблюдать значительно больший разрыв между значимостью отдельных переменных в пользу переменной, описывающей разность в часах между транзакциями, важность которой почти в два раза превышает значимость следующей переменной.

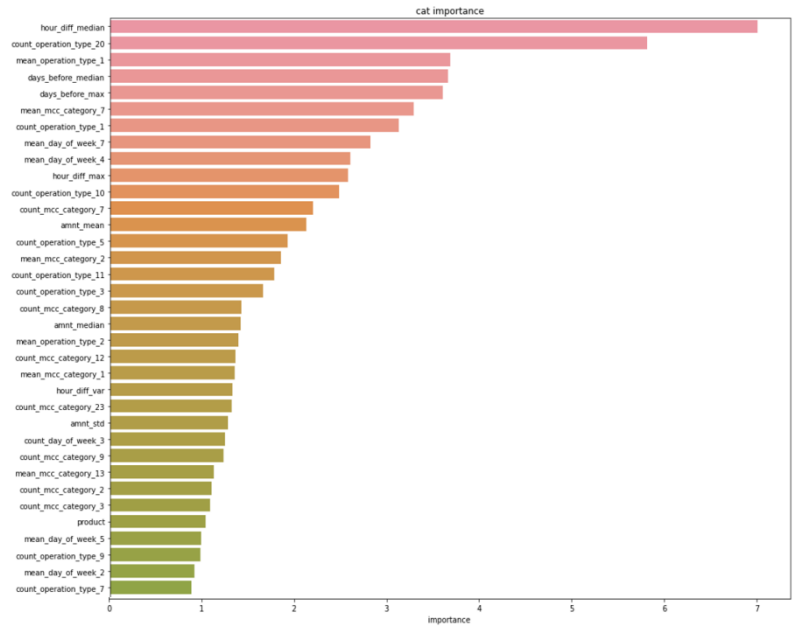


График 22 «Наиболее важные параметры в CatBoost-модели»

В CatBoost-модели мы можем наблюдать значительно меньшее разброс значений важности признаков, однако к наиболее значимым переменным мы все равно можем отнести достаточно малое их число.

Оценка качества работы модели

По результатам построения ансамбля из 4 базовых моделей были получены следующие результаты:

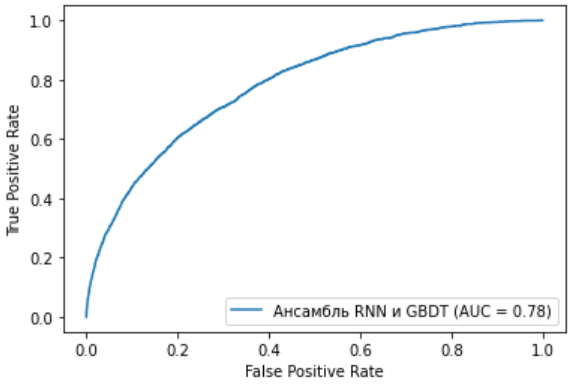


График 20 «AUC ROC кривая получена на базе тестовых данных в ансамбле из 4 моделей»

Получившийся уровень AUC смог превысить исходные требования к модели, но также стоит отметить достаточно ровное искривление AUC ROC кривой, что говорит о хорошей сбалансированности построенной на тренировочных данных модели.

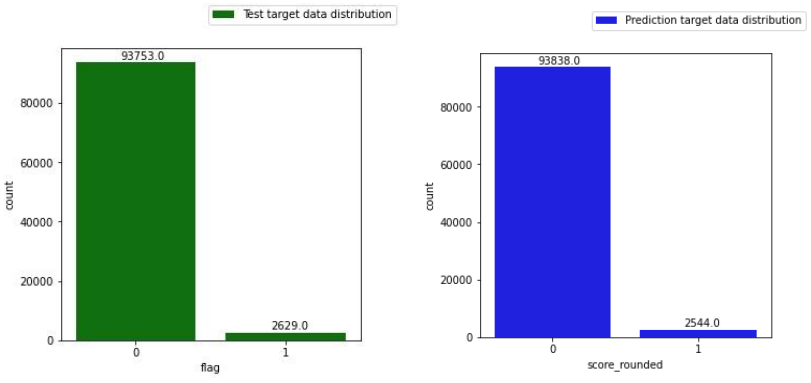


График 21 «Графики соотношения числа успешно выплаченных кредитных продуктов и уходов в дефолт в сравнении между тестовыми данными и предсказанными в ансамбле из 4 моделей»

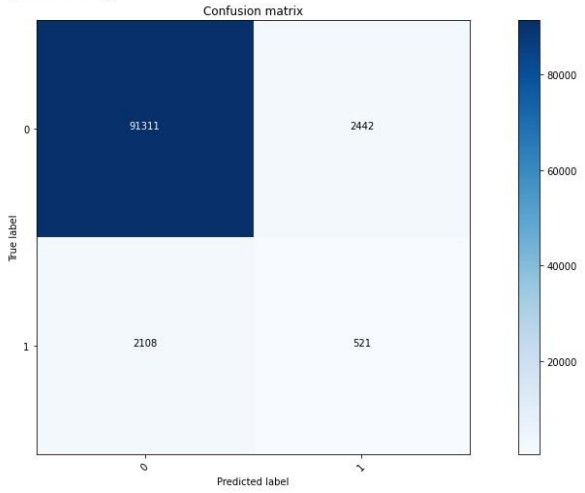
Для точной классификации было необходимо конвертировать вероятности дефолта в предсказания конкретного класса. При помощи перебора threshold-значений удалось найти границу 0,12, на которой F1-Score для тестовой выборки оказался бы наибольшим. Для сравнения можно

привести результаты проверки работы отдельной RNN-модели и её ансамбля с 3 GBDT-моделями.

Таблица 7 «Матрица ошибок для тестовых данных в RNN-модели и ансамбле из 4 моделей»

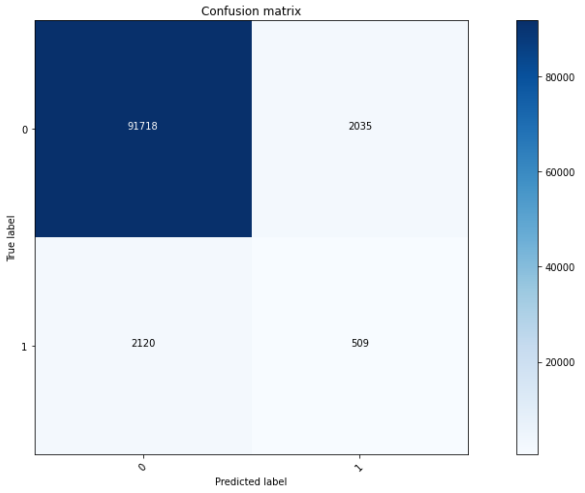
	precision	recall	f1-score	support
0	0.98	0.97	0.98	93753
1	0.18	0.20	0.19	2629
accuracy			0.95	96382
macro avg	0.58	0.59	0.58	96382
weighted avg	0.96	0.95	0.95	96382

```
[[91311  2442]
 [ 2108   521]]
```



	precision	recall	f1-score	support
0	0.98	0.98	0.98	93753
1	0.20	0.19	0.20	2629
accuracy			0.96	96382
macro avg	0.59	0.59	0.59	96382
weighted avg	0.96	0.96	0.96	96382

```
[[91718  2035]
 [ 2120   509]]
```



Хорошо видно, что при помощи ансамбля мы смогли немного улучшить F1-score с 0,95 до 0,96, что достаточно много, учитывая несбалансированность исходных данных. При этом ансамбль позволил серьёзно сократить число FN-прогнозов (на 400 из исходных 2400), что означает возможность для банка выдать дополнительные 400 кредитов (из суммарных 96 тысяч заявок), которые будут нам успешно возвращены. При этом стоит отметить, что наша система всё ещё может быть улучшена в направлении поиска потенциальных дефолтов. Возможно, комбинирование нашей системы с уже существующей в банке моделью, оценивающей риски выдачи кредитных продуктов по персональным данным, поможет значительно улучшить качество прогнозирования.

Заключение

В результате работы над курсовым проектом нами были выполнены все запланированные задачи:

- Проведен экспресс-анализ рынка банковского кредитования
- Изучены и описаны основные подходы к оценке кредитных рисков
- Выявлены и проанализированы требования банка к прогнозной модели
- Данные о транзакциях клиентов исследованы и подготовлены для дальнейшей работы
- Выделены значимые для модели факторы (предикторы)
- Выбраны метрики для оценки качества моделей, а также методы прогнозирования
- Разработаны требования к создаваемой нами системе
- Описаны ожидаемые бизнес-эффекты от внедрения новой системы предиктивной аналитики
- Разработан ансамбль моделей для предсказания кредитного риска, проведена оценка качества ансамбля
- Разработана система предиктивной аналитики для управления рисками коммерческого банка

По результатам завершения указанных выше этапов проекта заказчик смог ознакомиться с анализом рынка кредитных продуктов и решениями, которые существуют для оценки кредитных рисков банка; у него появилось понимание, что использующиеся в банке методы оценки заемщиков не оптимальны и требуется внедрение новых решений – системы предиктивной аналитики для управления рисками коммерческого банка.

В результате внедрения этой системы ожидается повышение эффективности кредитных продуктов (правильная оценка не даст отказать «хорошим» клиентам и выдать кредит потенциально «плохим» заемщикам); уменьшится время обработки заявок на кредитные продукты, что может повысить привлекательность банка для потенциальных заемщиков.

С инфраструктурной точки зрения система позволит снизить нагрузку на аппаратное обеспечение банка, так как модели будут использовать уже имеющиеся данные – историю транзакций клиентов, что не потребует сбора дополнительной информации о заемщиках посредством анкетирования и снизит скорость роста информационного актива банка.

Также механизм оценки заемщиков станет более прозрачным для руководства банка, у них появится доступ к результатам работы модуля визуализации нашей системы, с помощью которого они смогут наблюдать за ситуацией в режиме «near real time».

В результате внедрения системы decision intelligence процесс оценки заемщиков очистится от ряда субъективных факторов, которые возможны при «ручной» обработке анкет, что сделает весь механизм выдачи кредитов более прозрачным для заинтересованных лиц как внутри банка, так и вне его. Так, например, могут быть привлечены новые клиенты или повышена лояльность текущих (что, при выборе ими банка для получения кредита, может принести дополнительную прибыль).

Спрогнозированные нами бизнес-эффекты, которые должны быть достигнуты после успешного внедрения разработанной системы предиктивной аналитики представлены в таблице ниже.

Таблица 8 «Прогнозируемые бизнес-эффекты от внедрения разработанной системы предиктивной аналитики»

Количественные бизнес-эффекты	Качественные бизнес-эффекты
Сокращение срока рассмотрения кредитных заявок	Оптимизация документооборота (работа с меньшим количеством бумажной документации, исключение лишней документации и пр.)
Снижение доли невозвратных кредитов в кредитном портфеле банка	Повышение рейтинга надежности банка
Увеличение кредитного портфеля банка из-за снижения числа необоснованных отказов	Повышение привлекательности банка у заёмщиков
Сокращение времени обучения кредитных специалистов	Повышение прозрачности механизма оценки заёмщиков для руководства банка
Повышение точности вероятностной оценки кредитоспособности заёмщика	Уменьшение роли субъективного фактора в принятии решения по кредитной заявке

В качестве путей развития данной работы нашей командой предлагаются следующие варианты:

- дополнение ансамбля моделей для включения новых признаков транзакций клиентов в модель скоринга
- отслеживание изменения результатов работы предиктивной системы для предсказания оттока клиентов банка
- использование транзакционных данных пользователей продуктов банка в иных процессах, требующих скоринга клиентов

Библиографический список

1. Машинное обучение («Machine Learning») [Электронный ресурс]. - URL: machinelearning.ru (дата обращения: 24.05.2021).

2. Обзор методов и моделей кредитного и поведенческого скоринга [Электронный ресурс]. - URL: craftappmobile.com/obzor-metodov-kreditnogo-skoringa/#i-3 (дата обращения: 24.05.2021).

3. Официальный сайт Министерства экономического развития Российской Федерации [Электронный ресурс]. - URL: economy.gov.ru (дата обращения: 24.05.2021).

4. Официальный сайт Центрального банка Российской Федерации [Электронный ресурс]. - URL: cbr.ru/statistics (дата обращения: 24.05.2021).

5.Охлаждение пройденного: Банк России оценил уровень долговой нагрузки граждан [Электронный ресурс]. - https://rg.ru/2021/05/17/bank-rossii-ocenil-uroven-dolgovoj-nagruzki-grazhdan.html (дата обращения: 24.05.2021).

6. Полищук, Ф.С.. Романов, А.Ю. Кредитный скоринг: разработка рейтинговой системы оценки риска кредитования физических лиц // Новые информационные технологии в автоматизированных системах. - 2016. - №. 19. - С. 280–282.

7. Российские банки: финансовые итоги 1-го квартала 2021 года [Электронный ресурс]. - finversia.ru/publication/rossiiskie-banki-finansovye-itogi-1-kvartala-2021-goda-94834 (дата обращения: 24.05.2021).

8. Рыбальченко, Ю.С. Скоринг как инструмент оценки и минимизации кредитного риска // Молодой ученый. - 2017. - №35. - С. 37–40 [Электронный ресурс] - URL: moluch.ru/archive/169/45538 (дата обращения: 24.05.2021).

9. Скачкова, Е.К. Скоринг как метод оценки кредитного риска // Молодой ученый. - 2016. - №8. - С. 667–671 [Электронный ресурс] - URL:Вирз/ивошев.пуагевуе/1 12/28529/ (дата обращения: 24.05.2021).

10. Скоринг (Scoring) [Электронный ресурс]. - URL: banki.ru (дата обращения: 19.03.2021)

11. Скоринговая карта [Электронный ресурс]. - URL: basegroup.ru (дата обращения: 24.05.2021).

12. Современный скоринг: использование big data и machine learning [Электронный ресурс]. - nbj.ru (дата обращения: 24.05.2021).

13. ЦБ РФ, АНАЛИЗ ДИНАМИКИ ДОЛГОВОЙ НАГРУЗКИ НАСЕЛЕНИЯ РОССИИ В II–III КВАРТАЛАХ 2020 ГОДА НА ОСНОВЕ ДАННЫХ БЮРО КРЕДИТНЫХ ИСТОРИЙ [Электронный ресурс]. - URL: cbr.ru/collection/collection/file/31945/review_03022021.pdf (дата обращения: 24.05.2021).

14.Долговая нагрузка россиян достигла нового рекорда во время пандемии [Электронный ресурс]. - rbc.ru/finances/28/05/2020/ 5ecec4c89a7947f447db3619 (дата обращения: 24.05.2021).

15. Alternative Credit Scoring — Financial Salvation For Those With Low or No Credit [Электронный ресурс]. - URL: lending-times.com/2018/04/04/alternative-credit-scoring-financial-salvain-for-those-with-low-or-no-credit-score (дата обращения: 19.03.2021).

16.ARIS Community [Электронный ресурс]. - ariscommunity .com/help/aris-express (дата обращения: 24.05.2021).

17. Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Results [Электронный ресурс]. - toptal.com/machine-learning/ensemble-methods-machine-learning (дата обращения: 24.05.2021).

18. How market leaders are managing change with cloud-driven innovation - [Электронный ресурс]. - URL: oracle.com/ru/a/ocom/docs/esg-research-oracle-emerging-technologies.pdf (дата обращения: 24.05.2021).

19. Object Management Group, Business Process Model and Notation [Электронный ресурс]. - bpmn.org (дата обращения: 24.05.2021).

Приложения

Таблица 4 «План реализации проекта»

Название задачи	Длительнс	Начало	Окончани	Предшествен	Названия ресурсов	авить новый стол
Определение цели	2 дней	Чт 28.01.21	Пт 29.01.21			
Определение бизнес-требований к прогнозной модели	**8,63 дней**	**Пн 01.02.21**	**Чт 11.02.21**	**1**		
Описание объект	1 день	Пн 01.02.21	Пн 01.02.21			
Экспресс-анализ	2 дней	Вт 02.02.21	Ср 03.02.21			
Изучение подход	2 дней	Ср 03.02.21	Чт 04.02.21			
Идентификация с	3 дней	Пт 05.02.21	Вт 09.02.21	5		
Анализ требовани	2 дней	Пн 08.02.21	Чт 11.02.21			
Подготовка входных параметров для модели	**8 дней**	**Чт 11.02.21**	**Вт 23.02.21**	**2**		
Исследование и г	5 дней	Чт 11.02.21	Чт 18.02.21			
Выделение значи	3 дней	Чт 18.02.21	Вт 23.02.21	9		
Формализация требований	**1,38 дней**	**Вт 23.02.21**	**Ср 24.02.21**	**8**		
Формализация тр	1 день	Вт 23.02.21	Ср 24.02.21			
Формализация тр	1 день	Вт 23.02.21	Ср 24.02.21			
Формализация тр	1 день	Ср 24.02.21	Ср 24.02.21			
Выбор ИТ-решения	**6 дней**	**Чт 25.02.21**	**Ср 03.03.21**	**11**		
Выбор ИТ-решен	2 дней	Чт 25.02.21	Пт 26.02.21			
Выбор ИТ-решен	2 дней	Сб 27.02.21	Пн 01.03.21	16		
Выбор ИТ-решен	2 дней	Вт 02.03.21	Ср 03.03.21	17		
Построение демонстрационной модели	**7 дней**	**Чт 04.03.21**	**Пт 12.03.21**	**15**		
Выбор метрики	2 дней	Чт 04.03.21	Пт 05.03.21			
Выбор метода пр	2 дней	Пн 08.03.21	Вт 09.03.21	20		
Построение моде	3 дней	Ср 10.03.21	Пт 12.03.21	21		
Оценка эффективности построенной модели и выводы	**5 дней**	**Пн 15.03.21**	**Пт 19.03.21**	**19**		
Предварительная	1 день	Пн 15.03.21	Пн 15.03.21			
Написание Устава	2 дней	Вт 16.03.21	Ср 17.03.21	24		
Подготовка отчёт	2 дней	Чт 18.03.21	Пт 19.03.21	25		
Построение итоговой модели	**20 дней**	**Пн 22.03.21**	**Пт 16.04.21**	**23**		
Доработка исход	3 дней	Пн 22.03.21	Ср 24.03.21			
Выбор метода пр	4 дней	Чт 25.03.21	Вт 30.03.21	28		
Построение ново	7 дней	Ср 31.03.21	Чт 08.04.21	29		
Предварительная	1 день	Пт 09.04.21	Пт 09.04.21	30		
Создание ансамб	4 дней	Пн 12.04.21	Чт 15.04.21	31		
Предварительная	1 день	Пт 16.04.21	Пт 16.04.21	32		
Построение хранилища данных	**12 дней**	**Пн 19.04.21**	**Вт 04.05.21**	**33**		
Проектирование	5 дней	Пн 19.04.21	Пт 23.04.21			
Разработка храни	7 дней	Пн 26.04.21	Вт 04.05.21	35		
Построение BI-интерфейса	**11 дней**	**Пн 26.04.21**	**Пн 10.05.21**			
Проектирование	4 дней	Пн 26.04.21	Чт 29.04.21			
Разработка BI-инт	7 дней	Пт 30.04.21	Пн 10.05.21	38		
Построение системы предиктивной аналитики	**5 дней**	**Вт 11.05.21**	**Пн 17.05.21**	**39**		
Интеграция разра	5 дней	Вт 11.05.21	Пн 17.05.21			
Оценка эффективности разработанной системы предиктивной аналитики	**6 дней**	**Вт 18.05.21**	**Вт 25.05.21**	**41**		
Предварительное	3 дней	Вт 18.05.21	Чт 20.05.21			
Прогнозирование	2 дней	Чт 20.05.21	Пт 21.05.21			
Подготовка отчёт	2 дней	Пн 24.05.21	Вт 25.05.21	44		

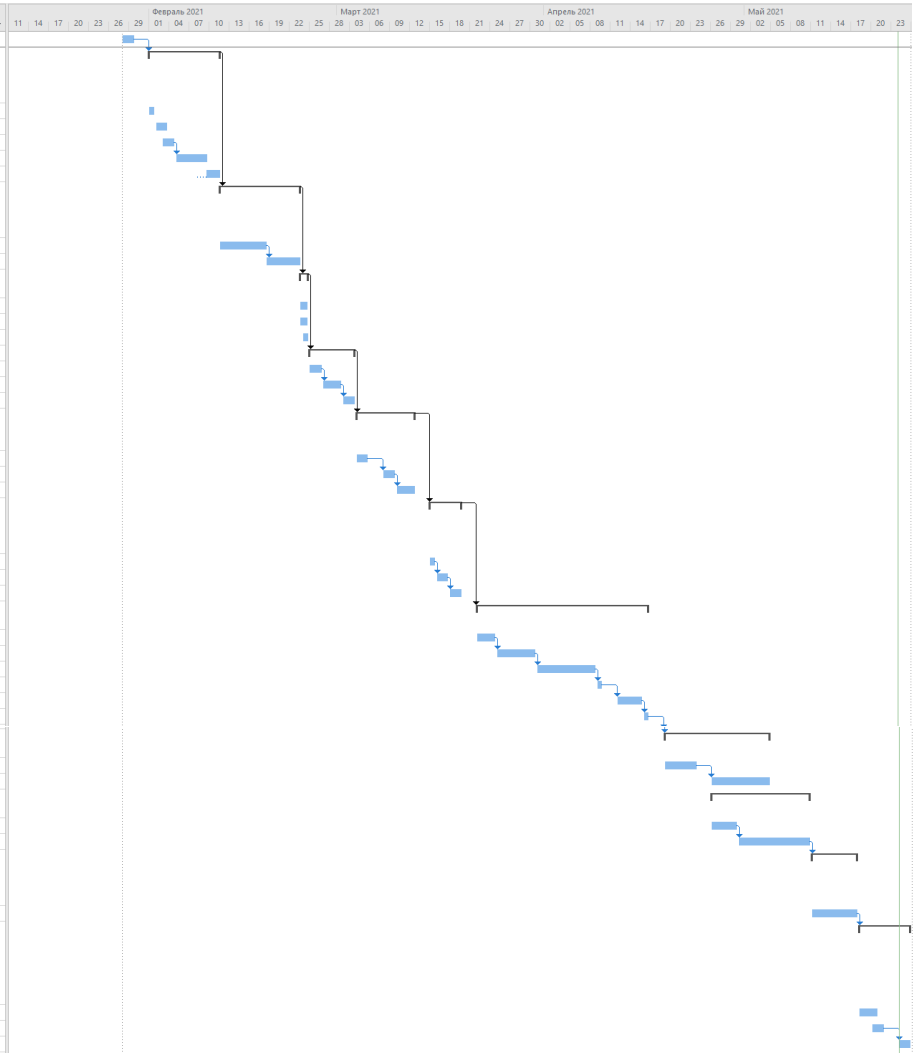


Таблица 5 «Пример исходных данных»

	app_id	amnt	currency	operation_kind	card_type	operation_type	operation_type_group	ecommerce_flag	payment_system	income_flag	mcc	country	city	mcc_category	day_of_week	hour	days_before	weekofyear	hour_diff	transaction_number
0	0	0.465425	1	4	98	4	2	3	7	3	2	1	37	2	4	19	351	34	-1	1
1	0	0.000000	1	2	98	7	1	3	7	3	2	1	49	2	4	20	351	34	0	2
2	0	0.521152	1	2	98	3	1	3	7	3	2	1	37	2	4	20	351	34	0	3
3	0	0.356078	1	1	5	2	1	3	7	3	10	1	49	7	2	0	348	34	52	4
4	0	0.000000	1	2	98	7	1	3	7	3	2	1	49	2	4	16	337	53	280	5

Устав проекта

Разработка и внедрение системы предиктивной аналитики

Версия 3.0

Журнал изменений

Дата	Версия	Описание	Автор
10.03.2021	1.0	Первая версия устава. В ней отражены все основные аспекты проведения проекта на момент написания устава.	Губин Д. М.
18.03.2021	2.0	Обновленная версия устава. Пересмотрены критерии успешности и риски проекта.	Губин Д. М.
22.03.2021	3.0	Обновленная версия устава. Разработаны и оценены риски проекта.	Губин Д. М.

Содержание

Название проекта **52**
Цели проекта **52**
Задачи проекта **52**
Критерии успешности проекта **52**
Ограничения проекта **53**
Команда проекта **53**
Этапы проекта **54**
Бюджет проекта **54**
Риски проекта **54**
Взаимосвязь с другими проектами **57**

Устав проекта “Разработка и внедрение системы предиктивной аналитики”

Название проекта

Код: 001

Символьное наименование: 001

Полное определение: Разработка и внедрение системы предиктивной аналитики для управления рисками коммерческого банка

Цели проекта

Цель проекта: Разработка и внедрение системы предиктивной аналитики для управления рисками коммерческого банка

Задачи проекта

1.Провести экспресс-анализ рынка банковского кредитования
2.Изучить и описать основные подходы к оценке кредитных рисков на рынке
3.Построить и описать схемы кредитного конвейера as-is и to-be
4.Выявить и проанализировать требования банка к прогнозной модели
5.Разработать требования к системе предиктивной аналитики для управления рисками коммерческого банка
6.Провести исследование и подготовку данных о транзакциях заёмщиков
7.Выделить значимые факторы для скоринговой модели
8.Выбрать метрики качества модели
9.Выбрать и обосновать методы прогнозирования
10.Построить модели машинного обучения для прогнозирования кредитного риска, а затем создать итоговую модель как ансамбль исходных
11.Провести предварительную оценку качества итоговой модели
12.Спроектировать и разработать хранилище данных для системы предиктивной аналитики
13.Спроектировать и разработать BI-интерфейс для системы предиктивной аналитики
14.Построить систему предиктивной аналитики путем интеграции разработанного хранилища данных, модели машинного обучения и BI-интерфейса
15.Провести предварительное тестирование системы предиктивной аналитики
16.Описать ожидаемые бизнес-эффекты от внедрения системы предиктивной аналитики

Критерии успешности проекта

1. Аналитический отдел успешно справился с задачей подготовки данных в предусмотренный срок.
2. Модель обучена и отлажена на локальном сервере в предусмотренный срок.

3. На этапе разработки BI-и не возникло проблем со связью с хранилищем. Приложение разработано в сроки и успешно работает.
4. Произведено успешное развертывание проекта на облачном сервере.
5. Облачный сервер выдержал нагрузку при первичном обучении и тесте модели.
6. Не превышен лимит трафика по тарифному плану сервера.
7. В течение первой недели запуска проекта команда технической поддержки успешно справляется с возникающими неполадками. Приложение не должно приостанавливать свою работу по техническим причинам более чем на 3 часа суммарно.
8. В течение первого месяца работы приложения произойдет менее 5 технических сбоев.
9. Не пришлось привлекать внешних экспертов для решения возникающих сбоев.

Ограничения проекта

Основное ограничение проекта – неполная адаптированность системы к поступлению новых данных, что влечет за собой потенциальное снижение точности работы модели в будущем. На текущий момент принято решение переобучать модель в фиксированных временных промежутках, однако, это менее оперативное решение, в сравнении с дообучением при получении новых данных. Однако, это ограничение необходимо для обеспечения быстродействия модели.

Команда проекта

Команда и роли в проекте		Описание
Лебедев Андрей	Team leader и разработчик модели, бизнес-аналитик	Координация проекта; выявление и анализ требований к модели; выделение значимых факторов, выбор метрики качества и методов прогнозирования для моделей; построение моделей машинного обучения; интеграция итоговой модели в систему предиктивной аналитики
Губин Даниил	Разработчик BI-интерфейса, бизнес-аналитик	Разработка требований к интерфейсу BI; анализ рисков проекта; создание устава проекта; проектирование и разработка BI-интерфейса; Интеграция BI-интерфейса в систему предиктивной аналитики; связь BI-интерфейса с хранилищем данных
Сухоруков Георгий	Рыночный аналитик, бизнес-аналитик	Проведение экспресс-анализа рынка; изучение и описание подходов к оценке кредитных рисков; создание схем кредитного конвейера as-is и to-be; составление плана проекта; оценка ожидаемых бизнес-эффектов от внедрения системы

Мамедов Артём	Data-engineer, data-scientist, бизнес-аналитик	Разработка требований к хранилищу данных; проведение исследования и подготовки данных о транзакциях заёмщиков; проектирование и разработка хранилища данных; интеграция хранилища данных в систему предиктивной аналитики

Этапы проекта

1. Этап подготовки данных и выбора ключевых атрибутов для предсказания – 1 неделя
2. Этап обучения модели – 2 недели
3. Этап тестирования и доработки модели – 1 неделя
4. Этап разработки BI приложения – 2 недели
5. Этап связывания BI приложения и хранилища – 5 дней
6. Этап развертывания проекта на облачном сервере – 3 дня
7. Этап тестирования проекта – 1 неделя
8. Дополнительное резервное время – 1 неделя до запуска проекта. Оно может быть использовано в случае сбоя в одном из этапов, чтобы не сдвигать финальный дедлайн запуска.
9. Запуск проекта – 1 день

Бюджет проекта

1. Облачный сервер, на котором будет держаться модель: 20 000 тыс. руб./мес.
2. Техническая поддержка BI приложения: 30 000 тыс. руб./мес.
3. Резерв на случай необходимости проведения дополнительного технического обслуживания – 15 000 тыс. руб./мес.

Риски проекта

1. **Риск недооценки масштаба работ**. [Риск возникновения: 15% | Степень ущерба: Высокая]

   С учетом грамотного планирования на этапе подготовки проекта, данный риск имеет сравнительно небольшой шанс возникновения. Однако, так как на текущий момент все этапы проекта обладают жесткими дедлайнами, при недооценке масштаба одного из этапов и, соответственно, временном сдвиге в нем, по цепочке пойдет сдвиг во всех последующих и отложит дату сдачи проекта. Этого возможно избежать 3 решениями:

   a. Сделать плавающие дедлайны, однако, это может негативно сказаться на мотивации работников.

b. Добавить резервное время перед этапом финального запуска проекта. При сбое в одном из этапов это все-еще повлечет временной сдвиг всех последующих, однако, дата сдачи проекта останется неизменной.
c. Расчет критического пути и раннее начало наиболее трудоемких и рискованных работ

2. **Риск перерасхода средств**. [Риск возникновения: 50% | Степень ущерба: Низкая]
Несмотря на то, что на этапе планирования проекта проведен анализ рынка, в рамках работы экономическая ситуация может измениться из-за внешних факторов (которые в период пандемии и так не показывают стабильного поведения), поэтому степень возможного возникновения данного риска довольно высока. Однако, степень ущерба мала, так как данный риск легко решается финансовыми подушками и грамотными контрактами с вендорами оборудования и ПО. Может произойти увеличение тарифов поддержки серверов в следствие различных экономических факторов, как, например, рост курсов валют или рост цен на серверное оборудования. Данный риск возможно минимизировать несколькими способами:
   a. Заключение контракта таким образом, что оплата производится перед рабочим периодом и в стоимость уже заложены риски увеличения цены. Таким образом сумма оплаты фиксируется и не может быть скорректирована в процессе работы.
   b. Увеличение финансовых резервов с учетом возможных неблагоприятных экономических условий в периоде работы оборудования.
3. **Технологический риск**. [Риск возникновения: 65% | Степень ущерба: Средняя]
Данный риск имеет высокую долю возникновения, так как зачатую невозможно учесть все особенности сервера до момента настройки системы непосредственно на нем. В рамках развертывания модели на сервере могут случиться технические сбои, которые повлекут за собой излишние траты трафика сервера, что может привести к превышению выделенного на проект бюджета. Этого можно избежать несколькими способами:
   a. Приобретение резервного локального сервера, включающегося в работу автоматически при сбое на облачном сервере.
   b. Заключение резервного контракта с другим вендором серверов, который предусматривает выставление счета по суммарному времени работы оборудования, сложенную с минимальной резервной оплатой за простой оборудования.
4. **Риск выставления новых требований заказчиком**. [Риск возникновения: 15% | Степень ущерба: Очень высокая]

Возможность возникновения данного риска очень мала в случае хорошо продуманных контрактов. Однако, в случае возникновения это может повлечь большие финансовые и временные потери. Данный риск повлечет перерасход как временных, так и трудовых ресурсов. Кадровый состав может оказаться не готовым к новым требованиям. Путей минимизации данного риска 2:

a. Составление договора о выполнении только конкретных задач в рамках проекта, а также заверении критериев успешности каждой из задач.
b. Резервация дополнительных кадров и увеличение временного резерва за счет заказчика.

5. **Риск недостаточной компетентности кадров к выполнению проекта на необходимом заказчику уровне**. [Риск возникновения: 25% | Степень ущерба: Очень высокая]

   С учетом тщательного тестирования персонала шанс возникновения данного риска довольно мал, однако, в случае возникновения это повлечет огромные финансовые и временные издержки на поиск решения тех проблем, которые возникли вследствие некомпетентности работников. Данный риск возможно минимизировать некоторыми путями:

   a. Ужесточение критериев отбора кадров.
   b. Проведение дополнительных тренингов.
   c. Заключение договоров со внешними специалистами, привлекаемыми в случае неудачи основного состава кадров.

Матрица рисков

Степень ущерба	критическая					
	очень высокая	Риск выставления новых требований заказчиком	Риск недостаточной компетентности кадров			
	высокая	Риск недооценки масштаба работ				
	средняя				Технологический риск	
	низкая			Риск перерасхода средств		
		0-20%	21-40%	41-60%	61-80%	81-100%
		Вероятность возникновения				

Таблица 1. Команда проекта.

Взаимосвязь с другими проектами

Модульная архитектура проекта позволит, при необходимости, встраивать его компоненты в другие проекты. Таким образом, разрабатываемый продукт в первую очередь будет взаимодействовать с системой, являющейся источником транзакционных данных, поступающих от клиента. Также проект предположительно будет встроен в ИС-инфраструктуру банка, что означает его связь с различными уже готовыми ИС. Также должна быть предусмотрена возможность добавления новых внешних данных.

Однако, на момент разработки проекта дугие проекты, не влияющие на результат нашего проекта, не рассматриваются.