МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)»

Институт№8 «Информационные технологии и прикладная математика» **Кафедра** 810Б
Направление подготовки 02.04.02 ФИИТ **Группа** М8О-203М-19
Квалификация (степень) магистр

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРА (МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)

На тему: «Применение методов машинного обучения при моделировании вероятности дефолта в задаче оценки кредитного риска»

Автор диссертации Лопатенко Валентин Васильевич
(Фамилия, имя, отчество) подпись

Научный руководитель Абгарян Каринэ Карленовна
(Фамилия, имя, отчество) подпись

Рецензент Думин Павел Николаевич
(Фамилия, имя, отчество) подпись

К з а щ и т е д о п у с т и т ь

Зав. кафедрой Абгарян Каринэ Карленовна
(Фамилия, инициалы) подпись

« 24 » мая 2021 г.

Москва 2021 г.

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)»

Факультет №8 *«Информационные технологии и прикладная математика»* **Кафедра** 810Б

Направление подготовки 02.04.02 ФИИТ **Группа** М8О-103М-19

Квалификация (степень) *магистр*

УТВЕРЖДАЮ

Зав. кафедрой 810Б _____ Абгарян К.К.
(подпись) (фамилия, инициалы)

«02» сентября 2019 г.

ПЛАН РАБОТЫ
над выпускной квалификационной работой магистра (магистерской диссертацией)

Студент Лопатенко Валентин Васильевич

Научный руководитель Абгарян Каринэ Карленовна
(фамилия, имя, отчество полностью)

д.ф.-м.н., доцент., зав кафедры 810Б МАИ
(ученая степень, ученое звание, должность и место работы)

1. Наименование темы: «Применение методов машинного обучения при моделировании вероятности дефолта в задаче оценки кредитного риска»

2. Срок сдачи студентом законченной работы 24 мая 2021 г.

3. Техническое задание и исходные данные к работе
Требуется реализовать модель количественной оценки вероятности дефолта корпоративных заёмщиков банка– юридических лиц. Провести моделирование на основе информации о судебных исках заёмщика и его финансовой отчетности. Провести сравнение между несколькими алгоритмами машинного обучения и выбрать лучший.

4. Перечень подлежащих разработке разделов и этапы выполнения работы

№ п/п	Наименование раздела или этапа	Трудоёмкость в % от полной трудоёмкости работы	Срок выполнения
1	2	3	4
1	Обзор литературы (кредитный риск и его компоненты)	9	02.09.2019 – 15.10.2019

2	Обзор литературы (задача кредитного скоринга)	8	16.10.2019 – 18.03.2020
3	Сбор данных на Hadoop кластере	20	19.03.2020 – 22.09.2020
4	Конструирование признаков и приоритезация даных	10	23.09.2020 – 20.10.2020
5	Написание фреймворка для однофакторного анализа на языке python	15	21.10.2020 – 21.12.2020
6	Проведение однофакторного анализа, формирование короткого списка факторов	15	22.12.2020– 18.02.2021
7	Многофакторный анализ, построение модели	10	19.02.2021 – 22.03.2021
8	Валидация построенной модели	5	23.03.2021 – 29.03.2021
9	Проведение альтернативного моделирования с использованием альтернативных алгоритмов машинного обучения	3	30.03.2021 – 05.04.2021
10	Написание отчета о разработке модели	3	06.04.2021 – 12.04.2021
11	Подготовка и оформление работы к защите диплома	4	13.04.2021 – 23.05.2021

5. Перечень иллюстративно-графических материалов:

№ п/п	Наименование	Количество листов
1	Раздаточный материал	14

6. Исходные материалы и пособия

1. Деан Фантаццини. Управление кредитным риском // Прикладная эконометрика 4(12) 2008: 84-137.
2. Сорокин А.С. Построение скоринговых карт с использованием модели логистической регрессии // Интернет-журнал «Науковедение» 2(3) 2014: 1-29.
3. Ковалев М., Корженевская В. Методика построения банковской скоринговой модели для оценки кредитоспособности физических лиц // Банки Казахстана 1 2008: 43-48
4. Карминский А.М. Кредитные рейтинги и их моделирование. М.: НИУ ВШЭ, 2015.

7. Дата выдачи задания 02 сентября 2019 г.

Научный руководитель _____ Абгарян К.К.
(подпись) (ФИО)

Задание принял к исполнению _____ Лопатенко В.В.
(подпись) (ФИО)

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)»

О Т З Ы В
НАУЧНОГО РУКОВОДИТЕЛЯ

Студент Лопатенко Валентин Васильевич
(фамилия, имя, отчество полностью)

Институт *№ 8 «Информационные технологии и прикладная математика»* **Кафедра** 810Б

Направление подготовки 02.04.02 ФИИТ **Группа** М8О-203М-19

Квалификация (степень) *магистр*

Тема диссертации: «Применение методов машинного обучения при моделировании вероятности дефолта в задаче оценки кредитного риска»

Научный руководитель Абгарян Каринэ Карленовна, д.ф.-м.н., доцент,зав. каф. 810Б МАИ
(фамилия, имя, отчество полностью, ученая степень, ученое звание, должность и место работы)

Отмеченные достоинства: Соискателю была поставлена задача разработать модель оценки кредитного риска корпоративных заёмщиков, а именно одной из его основных компонент – вероятности дефолта. Практическая предметная область, данные, а также постановка задачи исходили от заказчика – ПАО «Сбербанк», который представляет собой крупнейший банк в России. Особенностью данной работы является не только разработка модели, но и подготовка ее к внедрению в промышленную эксплуатацию для обработки в кластере на больших объемах данных. Более того, в ходе процесса моделирования студент самостоятельно инициировал и написал библиотеку для внутреннего использования, позволяющую рассчитывать и визуализировать множество статистических метрик для отбора факторов в модель. Данная работа выполнялась практически в течение всего срока обучения в магистратуре, в ходе которого Валентин показал высокий уровень профессионализма, а также самостоятельности в принятии решений и реализации задач.

Отмеченные недостатки: Существенных недостатков в работе не отмечается.

Заключение: Магистерская работа заслуживает оценки «отлично», а ее автор – Лопатенко В.В. присвоения степени «магистр» по направлению «Фундаментальная информатика и информационные технологии»

По теме диссертации были опубликованы одноименные научные тезисы.

Работа проверена на объем заимствования. % заимствования – 6%

«24» мая 2021 г. Научный руководитель _____ Абгарян К.К.
(подпись)

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)»

ЗАКЛЮЧЕНИЕ
РЕЦЕНЗЕНТА
О ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ МАГИСТРА
(МАГИСТЕРСКОЙ ДИССЕРТАЦИИ)

студента Лопатенко Валентина Васильевича

(Фамилия, Имя, Отчество)

Институт № 8 «Информационные технологии и прикладная математика» **Кафедра** 810Б

Направление подготовки 02.04.02 ФИИТ **Группа** М8О-203М-19

Квалификация (степень) магистр

Наименование темы: «Применение методов машинного обучения при моделировании вероятности дефолта в задаче оценки кредитного риска»

Рецензент Думин Павел Николаевич

(Фамилия И.О., ученая степень, ученое звание, должность и место работы)

к.ф.-м.н., доцент, зав. лабораторией МГППУ

Отмеченные достоинства: Автором, Лопатенко В.В., разработана и реализована модель оценки кредитного риска заемщика. В работе особенное внимание уделено анализу важности признаков: проведен как статистический анализ влияния факторов на целевую, так и эскпертные заключения. Модель валидирована и может быть применена в соответствующих предметных областях.

Отмеченные недостатки: В работе автору стоило более подробно визуализировать исходные данные.

Заключение: Магистерская диссертация заслуживает оценки «отлично», а ее автор – Лопатенко В.В.присвоения степени «магистр» по направлению «Фундаментальная информатика и информационные технологии»

“ 24 ” мая 2021 г. Рецензент Думин П.Н.

(подпись)

Подпись / печать верны.
Ведущий специалист по кадрам
отдела по работе с персоналом
Васильев А.М.

МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ОТДЕЛ ПО РАБОТЕ С ПЕРСОНАЛОМ

РУКОНТЕКСТ
text.rucont.ru

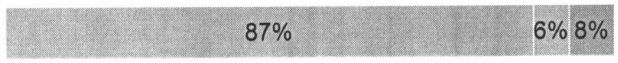
Справка о проверке на наличие заимствований

Имя файла: Магистерская диссертация (Лопатенко В.В.) (1).docx
Автор: Лопатенко Валентин Васильевич
Заглавие: Применение методов машинного обучения при моделировании вероятности дефолта в задаче оценки кредитного риска
Год публикации: 2021
Комментарий: *Не указан*
Коллекции: Интернет 2.0, Русскоязычная Википедия, Англоязычная Википедия, Коллекция Энциклопедий, Библиотека Либрусек, Университетская библиотека, Коллекция КФУ, ВКР Российского университета кооперации, Коллекция АПУ ФСИН, Коллекция ПГУТИ, Научная электронная библиотека "КиберЛенинка", ЦНМБ Сеченова, Авторефераты ВАК, Диссертации ВАК, Диссертации РГБ, Авторефераты РГБ, Готовые рефераты, ФИПС. Изобретения, ФИПС. Полезные модели, ФИПС. Промышленные образцы, Коллекция Руконт, Библиотека им. Ушинского, Готовые рефераты (часть 2), Открытые научные источники, eLIBRARY.RU, БиблиоРоссика, Правовые документы I, Правовые документы II, Правовые документы III, Собрание законодательства Российской Федерации

Результат проверки

Оценка оригинальности документа: 87%

Оригинальные фрагменты: 86,63%
Обнаруженные заимствования: 5,81%
Цитирование: 7,56%



Работу проверил: Абгарян Каринэ Карленовна

Дата: 24.05.2021

Подпись:

СПИСОК
научных трудов с 2019 по 2021 год
Лопатенко В.В.

№ пп	Наименование	Печ. или рук.	Название изд-ва	Год издания	Кол. стр.	Примечание соавторы
1	Использование дерева решений для нахождения оптимальных границ WOE-преобразования.	Печ.	XLVII Международная молодёжная научная конференция «Гагаринские чтения».	2021	2	

Соискатель _____ Лопатенко В.В.

Зав. каф. 810Б _____ Абгарян К.К.

РЕФЕРАТ

Магистерская диссертация содержит 52 страницы, 5 таблиц, 7 рисунков. Список использованных источников содержит 6 позиций.

КРЕДИТНЫЙ РИСК, ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ, КРЕДИТНЫЙ СКОРИНГ, WOE-ПРЕОБРАЗОВАНИЕ, ОЦЕНКА ВЕРОЯТНОСТИ ДЕФОЛТА

Магистерская диссертация посвящена построению модели оценки кредитного риска корпоративных заёмщиков, с использованием метода логистической регрессии в связке с алгоритмом WOE-преобразования в качестве ключевого метода. В качестве входных данных в модель поступают значения факторов заёмщика, на выходе модель предсказывает число, лежащее в интервале [0; 1]- вероятность дефолта в течение одного года.

Для решения поставленной задачи были выгружены данные о судебных актах в отношении заёмщика и финансовая отчетность компаний, с помощью которых были рассчитаны риск-факторы и построена модель. В качестве формы связи была выбрана модель логистической регрессии. При конструировании риск-факторов модели была учтена специфика предметной области, каждый фактор был подвержен ряду статистических и экономических проверок, в результате которых были отобраны лучшие с точки зрения метрики качества. Построенная модель показывает высокое значение коэффициента Джини.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Банковская система является звеном финансовой системы государства, обеспечивая жизнеспособность реальной экономики. Выступая в роли посредников, банки выполняют важную роль, участвуя в процессе эффективного перераспределения накоплений и инвестиций. Поскольку банки принимают на себя риски, они могут оказаться неплатежеспособными и потерпеть банкротство. В этом случае их вкладчики теряют сбережения, что может иметь разрушительные последствия и повлечь за собой потерю доверия ко всей банковской системе со стороны клиентов. Только устойчивая банковская система может выполнять возложенные на нее задачи и служить определенной гарантией общей стабильности экономики.

Вопрос управления рисками в банках вышел на первый план, особенно во время мирового финансового кризиса, который во многом стал результатом того, что финансовые организации неправильно оценивали последствия тех или иных своих действий и принятых на себя рисков. Недостатки систем управления рисками стали одной из основных причин банкротства банков. В России в наибольшей степени от кризиса пострадали кредитные организации со слабо развитой культурой управления рисками.

Кредитный риск является наиболее значимым банковским риском, что обусловлено важнейшей ролью кредитования в банковской деятельности. Именно кредит стимулирует развитие производительных сил, ускоряет формирование источников капитала для расширения воспроизводства, без кредитной поддержки невозможно обеспечить развитие хозяйств, предприятий, внедрение новых продуктов и технологий. В то же время операции по кредитованию - это самая доходная статья банковского бизнеса, за счет этого источника формируется основная часть чистой прибыли,

отчисляемой в резервные фонды и идущей на выплату дивидендов акционерам банка. Однако невозврат кредитов, особенно крупных, может привести банк к банкротству, а в силу его положения в экономике, к целому ряду банкротств связанных с ним предприятий, банков и частных лиц, поэтому для обеспечения стабильности экономической системы государства так важно, чтобы банки имели развитую систему управления кредитными рисками.

В российских банках несовершенство управления кредитным риском во многом вызывает неразвитость рынка кредитования, высокие процентные ставки по кредитам и высокую долю просроченной задолженности в кредитном портфеле. За период с февраля 2020 года по март 2021 года суммарная просроченная задолженность в структуре кредитного портфеля российских банков выросла на 577 млрд руб. (или на 17,5%) и на 1 марта 2021 года составила 3,848 трлн руб. Ее доля в кредитном портфеле российских банков составляет 38% на 1 января 2021 года и на следующий год прогнозируется, что она сохранится на уровне 44,5%. Системы оценки и управления кредитным риском существуют в том или ином виде в каждой кредитной организации, однако часто они носят исключительно формальный характер или реализуются за счет применения западных методик без адаптации их к отечественным условиям, что ведет к ослаблению позиции банка в повседневной конкурентной борьбе и при наступлении кризисных ситуаций. В связи с этим актуальной задачей является разработка подхода или модели, способных в условиях нынешних реалий как можно более точно оценить величину кредитного риска, а, в частности, одну из главных его компонент – вероятность выхода в дефолт или вероятность неисполнения заёмщиком своих долговых обязательств.

ОСНОВНАЯ ЧАСТЬ

1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Понятие кредитного риска

Кредитный риск представляет собой риск нарушения обязательств по платежам – неопределенность, связанную со способностью компании по обслуживанию своих долгов и возможностью отвечать по взятым на себя финансовым обязательствам. Очевидно, что заранее невозможно идентифицировать те компании, которые смогут выполнить свои обязательства и те, которые этого сделать не смогут. Так, в лучшем случае можно лишь дать вероятностную оценку кредитного риска. В результате компании, которые подвержены кредитному риску, обычно платят за пользование заемными средствами по процентной ставке, равной безрисковой процентной ставке, умноженной на коэффициент, который пропорционален вероятности нарушения обязательств по платежам. С помощью этой величины регулируется размер возмещения кредиторам за неопределенность, связанную с такими кредитными обязательствами.

В случае, когда заёмщик неспособен выполнить собственные обязательства перед кредитором или один из контрагентов не имеет возможности выполнить условия финансового соглашения, говорят, что контрагент находится в состоянии дефолта. Кредитный риск включает риск, связанный с событиями, отличными от дефолта, например, с изменением кредитного рейтинга заёмщика.

Риск возникает в ситуации неопределенности, связанной со способностью заёмщика выполнять свои контрактные обязательства. Этот риск характерен для банковской коммерции. Отсутствия диверсификации кредитного риска в банках в ряде случаев приводило к банкротству. Кроме того, при введении свопов и фьючерсных контрактов, а также в результате

роста объемов внебиржевых рынков кредитный риск стал иметь ключевое значения при управлении инвестициями.

Компоненты кредитного риска определены следующим образом:

1. Вероятность дефолта (ВД), которую можно рассматривать с позиции двух элементарных исходов:
    a. Платежеспособность
    b. Неплатежеспособность заёмщика
    c. А также с позиции ухудшения кредитного рейтинга, которое указывает на увеличение вероятности дефолта.
2. Доля невозвращенных средств при дефолте: в случае реализации события дефолта при наличии залогового обеспечения теряется не вся сумма кредита. Так, мы приходим к понятию нормы восстановления заёмщика, которая определяется как доля всех кредитных обязательств заёмщика и может быть покрыта в случае дефолта.
3. Величина номинальных потерь при дефолте – сумма кредитных обязательств заёмщика в момент реализации некоторого кредитного события, например, дефолта или банкротства.

Задача банка состоит в том, чтобы управлять риском своего кредитного портфеля. Управление кредитным риском включает в себя ряд мер, в том числе удержание определенной величины резервного капитала на случай кризисных ситуаций, которая определяется надзорными за банковской деятельностью органами. Существуют следующие виды деятельности, направленные на управление кредитным риском:

1. Оценка риска кредитного портфеля
2. Расчет величины регулирующего капитала и ее сохранение
3. Определение экономического капитала для внутренних целей

4. Обеспечение диверсификации портфеля, а также выявление очагов риска
5. Уменьшение концентраций риска путем хеджирования кредитными деривативами
6. Приобретение новых кредитных рисков или «разгрузка» старых
7. Сопоставление скорректированных риском эффективностей различных секторов портфеля

1.2. Виды убытков в кредитном риске

Ожидаемые убытки – математическое ожидание $E(L)$ фактических убытков L. Убытки такого рода не представляют опасности для банка, поскольку если фактические убытки в точности равнялись бы ожидаемым убыткам, для банка не было бы никаких негативных последствий: в случае, если банк отложил резервный капитал, размер которого равен ожидаемым убыткам, то никаких неожиданных изменений величины прибыли в этом случае не произошло бы. Кроме того, банк покрывает эти убытки, взымая их со своих заёмщиков в форме премии за риск. Это явным образом происходит, например, при ценообразовании кредитом, когда заёмщик платит премии, напрямую зависящие от его кредитоспособности. Так, например, в случае облигаций, купонные платежи по ним являются неявными рисковыми премиями.

Так, в банке больше всего беспокоятся об убытках, превышающих величину ожидаемых убытков – о неожидаемых убытках – разницей между величиной фактических убытков и ожидаемых убытков. Последняя измеряется с помощью математического ожидания распределения, описывающего фактические убытки.

Экономический капитал – это капитал, требуемый банку для того, чтобы ограничить вероятность банкротства до данного доверительного уровня на заданный промежуток времени. Это своего образа попытка оценить риск с точки зрения нынешних экономических реалий. Модели оценки экономического капитала более реалистичны, чем модели регулирующего капитала. Во многих крупных банках имеются проекты разработки таких моделей. Банк, имеющий хорошую модель экономического капитала, может эффективнее использовать свой капитал.

Подводя промежуточный итог, следует сказать, что в процессе управления кредитным риском помимо задач, аналогичных тем, которые возникают при управлении рыночным риском, существует и ряд специфических задач:

- Данные: недостаток публичной информации, касающейся кредитного качества заёмщиков. Важную роль в этом пункте играют кредитные агентства.
- Более длинный временной горизонт (как правило, не менее года)
- Функции распределения убытков обычно сильно скошены, имеют длинный правый хвост, что указывает на частые малые и редкие большие убытки
- Моделирование зависимости между убытками в портфеле – более приоритетная задача, чем при управлении рыночным риском, поскольку на хвост функции распределения убытков сильно влияет спецификация зависимости между дефолтами.

1.3. Регулирование кредитных рисков

Для регулирования в области банковского надзора в отношении требований к собственному капиталу банка в 1988 году было разработано

первое Базельское соглашение как реакция со стороны банковского сообщества и надзорных органов на случае крупных потерь и банкротств финансовых посредников, наблюдавшееся в 1970-1980 годы. Ключевая его идея – ограничение кредитного риска и возможных потерь банков путем построения системы контроля регулятора за достаточностью капитала банков.

В первом Базельском соглашении предприняты первые шаги к созданию международных стандартов расчета минимальной величины регулирующего капитала. Тем не менее, основной подход, предложенный в соглашении, являлся довольно грубым и недостаточно дифференцированным, поскольку основное внимание в нем уделялось различию кредитного риска, связанного с государственными, банковскими и ипотечными облигациями небанковского частного сектора, очевидно, имеющих низкий риск и коммерческих кредитных обязательств, подверженных высокому риску. При этом практических ничего не говорилось относительно дифференцирования кредитного риска в рамках классификации коммерческих кредитов. Для всех таких кредитов должна была быть обеспечена 8%-совокупный резервный капитал вне зависимости от кредитоспособности заёмщиков, их кредитного рейтинга, предложенного обеспечения, условий сделки. Вышло так, что для кредитов с высоким уровнем риска был установлен слишком низкий уровень достаточности капитала, а для сделок с низким риском – слишком высокий.

Новое Базельское соглашение («Базель – II») призвано скорректировать процедуру оценки уровня резервного капитала, принятую в первом Базельском соглашении, и сделать ее более гибкой и чувствительной к риску. Так, для оценки кредитного риска банки теперь могут использовать стандартизированный подход, введенный в соглашении «Базель – I», однако при этом крупные банки имеют возможность выбора в пользу подхода, основанного на внутренних рейтингах (ПВР).

1.3.1. Стандартизированный подход

При использовании стандартизированного подхода риск определенного актива вычисляется путем умножения величины номинальных потерь при дефолте на соответствующий вес риска. Веса риска при этом определяются посредством внешних рейтингов заёмщика.

После того, как риск каждого входящего в портфель актива вычислен, можно вычислить риск всего кредитного портфеля, который представляет собой сумму рисков активов. В этом случае резервный капитал определяется путем умножения риска портфеля на величину, известную как норма покрытия Куки, которая приблизительно равна 0,08.

1.3.2. Подход на основе внутренних рейтингов

На данный момент можно выделить две разновидности подхода, основанного на внутренних рейтингах:

- Базовый ПВР
- Усовершенствованный ПВР

При ПВР-подходе банки вправе оценивать вероятность дефолта заёмщиков самостоятельно. Более того, при использовании усовершенствованного ПВР-подхода банки могут оценивать долю невозвращенных средств при дефолте. Эти оценки должны быть получены на основании специфических моделей, которые являются приемлемыми с точки зрения внешнего регулятора. Во втором Базельском соглашении величина капитала на покрытие кредитного риска определяется с помощью ранее оцененных величин вероятности дефолта, величины номинальных потерь при дефолте и доли невозвращенных средств при дефолте.

Так, в любом подходе, основанном на внутренних рейтингах, можно выделить пять основных компонент:

- Внутренняя рейтинговая модель
- Компоненты риска
- Весовая функция рисков
- Перечень минимальных требования для применения ПВР
- Обзор случаев соблюдения минимальных требования, предоставленный надзорными органами

Для компонент риска при применении ПВР-подхода справедливо следующее:

- Величина вероятности дефолта может быть определена на основании ретроспективной информации или на основании скоринговых моделей
- Величина номинальных потерь определяется как номинальная сумму невозвращенной задолженности, при этом в качестве факторов, влияющей на корректировку данной величины, могут выступать, например, наличие обеспечения по кредиту.
- Доля невозвращенных средств при дефолте определяется на уровне 45% в случае использования базового ПВР. В случае использования усовершенствованного ПВР-подхода она определяется как оценка фактической величины доли невозвращенных средств, полученной банком.

1.4. Обзор моделей определения вероятности дефолта

1.4.1. Экспертный подход

При таком подходе решение о выдаче или невыдаче кредита принимается экспертом – кредитным инспектором, который может учитывать разные факторы, влияющие на кредитный риск. Как правило, это решение зависит от следующих факторов:

- Кредитная история заёмщика
- Деловая репутация
- Финансовые показатели на основе финансовой отчетности заёмщика
- Наличие обеспечения по кредиту
- Текущее состояние экономического цикла

1.4.2. Подход, основанный на кредитном скоринге

Кредитный скоринг представляет собой математическую модель, с помощью которой на основании такой информации как:

- Кредитная история
- Финансовая отчетность
- Данные об арбитражных делах
- Транзакционные данные

о заёмщике предсказывает вероятность невозврата кредита в определенный срок.

В самом простом виде такая модель представляет собой линейную комбинацию определенных характеристик (факторов) заёмщика и соответствующих им весовым коэффициентам. Величина на выходе модели –

скорбалл – интегральный показатель, прямым образом связанный со степенью надежности клиента, позволяющий отранжировать заёмщиков по степени величины кредитного риска.

1.5. Постановка задачи определения вероятности дефолта

Итак, после того, как мы определились с основными понятиями и описали процесс моделирования, можно сформулировать задачу на математическом языке.

Пусть $x_1, x_2, \dots, x_n$ – риск-факторы модели. То есть факторы, которые определенным образом описывают уровень кредитного риска заёмщика. Обозначим через PD вероятность дефолта заёмщика в течение года. Под дефолтом будем понимать дальнейшее невозможность заёмщика исполнить свои долговые обязательства перед кредитной организацией.

Тогда по имеющейся исторической информации о значениях риск-факторов заёмщика и соответствующим им меткам из множества $\{0; 1\}$, означающих исполнение своих обязательств и выход в дефолт соответственно, необходимо восстановить зависимость вида:

$$PD = f(x_1, x_2, \dots, x_n) \tag{1.1}$$

1.6. Описание процесса моделирования

1.6.1. Понятие выборки. Требование к формированию выборки.

Итак, при наличии достаточного объема различной информации, описывающей заёмщика, для того, чтобы уметь дифференцировать клиентов по степени их кредитного риска и принимать решения о кредитовании на

основе формализованных факторов, непосредственно связанных с этой величиной, необходимо построить математическую модель.

Для выполнения этой задачи, как правило, в первую очередь формируется выборка клиентов на основе ретроспективных данных, о которых уже можно сказать, «хорошие» это или «плохие заёмщики». То есть таких клиентов, которые уже имели опыт использования банковских продуктов и среди данных есть информация о том, выплатили ли такие клиенты кредит целиком или по их обязательствам по тем или иным причинам был реализован факт дефолта. Такие данные часто называют исторической выборкой. К такой выборке обязательно применяют условие ее репрезентативности – то есть способности в максимально полной мере отражать исследуемую генеральную совокупность.

Для проверки корректности работы модели и обобщающей способности из полученной выборки выделяется множество клиентов, не участвующее в обучении модели, которое будем называть тестовой выборкой, оставшихся же для обучения модели клиентов будем называть тренировочной выборкой или обучающей. Такое разделение, как правило, производят посредством механизма случайного отбора в соотношении 70:30 или 80:20, в зависимости от объема исходной исторической выборки.

Поскольку контрольная выборка призвана для того, чтобы показать, как модель будет работать на реальных данных в «боевых» условиях, важно обеспечить корректность ее формирования и репрезентативность генеральной совокупности. То же самое верно и для обучающей выборки, поскольку если отобрать из исторической выборки клиентов, сильно отличающихся от тех, которые представлены в генеральной совокупности, мы получим смещенные оценки весов факторов, а, следовательно, модель, не способную адекватно

ранжировать заёмщиков. Для осуществления корректности такого разбиения, как правило, обращают внимание на

Target rate по контрольной и обучающей выборкам – в идеальной ситуации они должны быть как можно ближе по значению. Такое разбиение называют стратифицированным по целевой переменной. Под target rate в случае решения задачи кредитного скоринга обычно понимается так называемый default rate – доля «плохих» заёмщиков в общей массе. Он оценивается согласно следующей формуле:

$$DR = \frac{bads}{goods + bads}, \tag{1.2}$$

где параметры bads и goods отражают количество дефолтных и недефолтных наблюдений соответственно

Стабильность факторов. Для расчета стабильности поведения факторов на обучающей и тестовой выборках можно использовать коэффициент стабильности популяции – PSI. В случае высокого значения этого коэффициента (выше 0.2) можно говорить об отсутствии свойства репрезентативности контрольной выборке и необходимости повторного разбиения. Данный коэффициент также можно рассчитывать отдельно для наблюдений с реализованным событием дефолта и для наблюдений, событие дефолта по которым реализовано не было.

Допустим, мы хотим рассчитать коэффициент стабильности популяций для двух переменных: y_1, y_2. Для расчета PSI необходимо сперва разделить переменные на некоторое количество интервалов, стандартной практикой является разделение на 10% или 5% перцентили. Тогда формула для расчета выглядит следующим образом:

$$PSI = \sum_{i=1}^{n} \left(y_1^i - y_2^i\right) * \ln(y_1^i - y_2^i) \quad (1.3)$$

где n – общее количество интервалов разбиения, y_1^i – количество значений первой переменной в i-ом интервале, y_2^i – количество значений второй переменной в i-ом интервале.

В ситуации, когда модель уже построена, можно также обратить внимание на стабильность распределения скорбалла на контрольной и обучающей выборках с помощью вышеупомянутой метрики PSI. В ситуации, когда значение коэффициента высоко, можно говорить о плохой обобщающей способности модели или отсутствии свойства репрезентативности контрольной выборки.

Хорошей практикой также является стратификация выборок по другим параметрам, описывающим заёмщика, например:

- Бизнес-сегмент заёмщика, который характеризует размер конкретной компании (например, величину выручки компании)
- Отрасль
- Период наблюдения

1.6.2. Методы классификации в задаче кредитного скоринга

На задачу скоринга можно смотреть как на задачу бинарной классификации, призванной наиболее точно разделить два множества клиентов на «плохих» и «хороших». Несмотря на некоторую «детскость» этих определений, эта терминология достаточно давно устоялась в области управления кредитным риском и до сих пор используется.

Сами методы классификации можно разделить на классические статистические модели (линейная регрессия), методы линейного программирования, а также на методы, знакомые из теории машинного обучения:

- Дерево решений – в основе алгоритма лежит некоторый информационный критерий (например, энтропийный критерий), который максимизируется про поиске разбиения на каждом шаге на основании какого-либо фактора. Из очевидных плюсов алгоритма в контексте кредитного скоринга следует отметить его прозрачность и интерпретируемость. Из минусов – способность легко переобучиться на конкретной выборке, «заучить» ее, и, как следствие, большой разброс получаемых оценок и неустойчивость алгоритма к входным данным.
- Градиентный бустинг – ансамблевой алгоритм – представляет собой последовательное построение решающих деревьев таким образом, чтобы минимизировать на каждом шаге не отклонение очередного построенного алгоритма от реальных значений целевой переменной, а отклонение между предсказанными значениями и ошибкой, полученной в результате отклонения уже построенных базовых алгоритмов от значений целевой переменной.
- Случайный лес – наряду с градиентным бустингом также является ансамблевым методом, в котором для построения каждого базового алгоритма используется сгенерированная с повторениями исходная выборка аналогичного размера, при этом в ходе поиска каждого разбиения используются не все признаки, поданные на вход модели, а некоторое случайное подмножество меньшего размера. Последняя идея носит название метода случайных подпространств. Классификация объектов при этом происходит методом голосования, а побеждает тот класс, за который проголосовало наибольшее число деревьев. Такой подход позволяет получить

максимально независимые базовые алгоритмы, ансамбль которых будет характеризоваться минимальным смещением и разбросом получаемых оценок.

•Нейронные сети – методы глубокого обучения позволяют моделировать более сложные, нелинейные зависимости, которые не могут уловить простые линейные модели, однако для построения таких моделей требуется немалое количество данных для конкретного сегмента заёмщиков, которым может похвастаться не каждый банк. Более того, нейросетевые методы являются, своего рода, «черным ящиком», что затрудняет интерпретацию их предсказаний для кредитных аналитиков.

Традиционными методами, полюбившимися в кредитном скоринге за их простоту и экономическую интерпретируемость, стали регрессионные методы, а, прежде всего, линейная регрессия. Так, при использовании многофакторной регрессии величина вероятности дефолта определяется следующим образом:

$$p = w_0 + w_1 x_1 + \cdots + w_n x_n \qquad (1.4)$$

Однако при использовании именно линейной регрессии очевиден ее недостаток, связанный с масштабами используемых факторов, которые не ограничены в своей области определения, в то время как величина вероятности дефолта ограничена в интервале $[0; 1]$.

Данный недостаток позволяет преодолеть использование метода логистической регрессии и устанавливает следующее соотношение:

$$\log\left(\frac{p}{1-p}\right) = w_0 + w_1 x_1 + \cdots + w_n x_n \qquad (1.5)$$

Стоит отметить, что все регрессионные методы достаточно чувствительны к наличию корреляции между факторами, поэтому факторы модели обязательно подвергаются проверке на мультиколлинеарность, для выявления которой используется коэффициент вздутия дисперсии (VIF – Variance Inflation Factor), определяемые по формуле:

$$VIF = \frac{1}{1 - R_j^2} \tag{1.6}$$

Где R_j^2 представляет собой коэффициент детерминации регрессии j-го фактора на все остальные, то есть регрессии вида:

$$x_i = w_0 + w_1 x_1 + w_{i-1} x_{i-1} + w_{i+1} x_{i+1} + \cdots + w_n x_n \tag{1.7}$$

В случае наличия мультиколлинеарности дисперсия оценок факторов регрессионной модели возрастает пропорционально данной величине, что делает их оценку нестабильной. Так, можно говорить о наличие мультиколлинеарности в случае, когда коэффициент вздутия дисперсии принимает значение, превышающее 5. Как правило, при выявлении таких факторов, они исключаются из модели.

1.6.3. Категоризация количественных переменных

Зачастую при построении скоринговой модели используется специфический для области анализа кредитных рисков метод категоризации – WOE-преобразования (англ. WOE – Weight Of Evidence). Данный метод категоризации позволяет упростить обработку выбросов, пропусков и экстремальных значений, а также существенно упростить интерпретацию получаемых оценок. Более того, по своей природе данный метод преобразования позволяет нивелировать недостаток линейных моделей,

поскольку позволяет уловить сложные нелинейные взаимосвязи между предиктором и зависимой переменной.

Алгоритм работы WOE-преобразования:

•Сперва каждый регрессор разбивается на n равных процентилей. Как правило, используют 5 процентили. После чего происходит вычисление значение WOE для каждой группы по следующей формуле:

$$WOE_i = \ln(\frac{d_i^{(1)}}{d_i^{(2)}}), \quad (1.8)$$

где величины $d_i^{(1)}$ и $d_i^{(2)}$представляют собой относительные частоты «плохих» и «хороших» заёмщиков в каждой группе соответственно

•Затем происходит объединение соседних групп, после чего расчет показателей WOE повторяется.

•При укрупнении категорий главным образом руководствуются следующими принципами:

- Размер группы не должен быть слишком мал (обычно говорят о размере на уровне 5% относительно всех наблюдений переменной)
- Внутри группы должны быть представлены как дефолтные наблюдения, так и недефолтные
- Доля «плохих» заёмщиков в каждой группе должна значимо отличаться между группами
- Значения WOE должно подчиняться условию монотонности – убывать или возрастать при переходе от одной группы к другой. Это требование явным образом обеспечивает интерпретируемость получаемых оценок.

•Итоговое разбиение подбирается таким образом, чтобы оно доставляло максимум выбранной метрике качества, вычисленной на регрессии,

построенной на конкретном преобразованном факторе. В качестве такой метрики, как правило, используется коэффициент Джини.

Так, для поиска такого разбиения можно использовать дерево решений, при построении которого на каждом шаге можно использовать коэффициент Джини, проверяя, соответствует ли полученное разбиение требованиям, связанным с монотонностью преобразованной переменной, размером групп и тд.

1.6.4. Методы и способы включения регрессоров в модель.

Как правило, решение о включении очередного фактора в модель или же его исключения принимается на этапе так называемого однофакторного анализа, в ходе которого строится регрессионная модель по каждому преобразованному фактору отдельно, и он анализируется независимо от остальных. Обычно решение принимают на основании следующих критериев:

- Статистическая значимость полученной оценки коэффициента регрессии. Для проверки гипотезы о незначимости фактора используется тест Стьюдента, в ходе которого тестируется гипотеза о равенстве истинного значения коэффициента генеральной совокупности нулю. Рассчитывается t-статистика, для которой определяют связанной с ней значение P-value – вероятность получить такое значение коэффициента случайным образом. Как правило, в случае, когда полученное значение превышает 5% уровень значимости, фактор исключается из модели.
- Скоррелированность фактора. Для оценки корреляций между факторами используются коэффициенты корреляции Спирмена и Пирсона. В случае, когда два фактора имеют высокое значение коэффициента корреляции (скажем, больше 0.7), исключается тот, который имеет меньшую предсказательную силу в смысле выбранной метрики. Отсутствие

мультиколлинеарности полученного множества факторов можно проверить с помощью метрики VIF.

- Стабильность поведения на обучающей и контрольной выборке. Для определения используется коэффициент стабильности популяции, который может быть рассчитан на всех наблюдениях на обучающей и тестовой выборках, а также отдельно на наблюдениях, описывающих как «плохих», так и «хороших» заёмщиков. В случае, когда на одном из множеств данный коэффициент принимает высокое значение, фактор исключается из модели.
- Ранжирующая способность фактора. Для каждой регрессии, построенной на конкретном факторе, вычисляется метрика, описывающая ранжирующая способность модели (как правило, в качестве этой метрики выступает коэффициент Джини). В случае малого значения этой метрики а также существенного уменьшения ее величины на контрольной выборке относительно обучающей фактор исключается из модели.

Далее к полученному множеству факторов применяется процедура Stepwise Selection, которая будет описана далее в главе, посвященной практической реализации.

1.6.5. Критерии качества оценки модели

Важной характеристикой построенной модели является ее способность адекватно ранжировать заёмщиков – отличать «хороших» от «плохих». Для оценки ранжирующей способности модели строят ROC-кривую, которая показывает зависимость количества верно классифицированных положительных исходов от количества неверно классифицированных отрицательных исходов. Для сравнения нескольких моделей используют показатель, называемый ROC-AUC (Area Under ROC-Curve) – площадь под ROC-кривой. Этот показатель измеряется от 0.5 до 1, где значение 0.5

соответствует плохому качеству модели, сравнимому с качеством случайного классификатора, который присваивал бы каждому из двух классов ту или иную метку с вероятностью 0.5. Значение 1 соответствует классификатору, способному безошибочно разделить два класса.

Стоит сказать, что в качестве метрики качества построенной модели чаще используется коэффициент Джини, получаемый применением линейного преобразования к рассчитанному показателю ROC-AUC по следующей формуле:

$$Gini = AUC * 2 - 1 \quad (1.9)$$

2. ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1. Описание программных продуктов и технологий, используемых для моделирования

Далее в ходе практической реализации на всех этапах используется язык программирования Python 3.5 в среде программирования JupyterLab.

Для выгрузки данных использовался фреймворк pyspark, предоставляющий удобное для разработчика API для работы с данными на движке Spark. Сами данные физически располагаются на Hadoop-кластере под управлением менеджера задач Yarn.

Для расчета статистических метрик были использованы библиотеки numpy, scipy, statsmodels. Для визуализации результатов – библиотеки matplotlib и seaborn.

Для подготовки данных для моделирования был использован фреймворк pandas. Построение модели происходило при помощи соответствующих классов в библиотеках statsmodels, sklearn.

2.1. Описание формирования выборки и определение целевой переменной

2.1.1. Формирование исторической выборки для разработки модели

Для формирования выборки была использована специальная форма, которая содержит информацию о:

- Действующих кредитных договорах
- Действующих кредитных продуктах
- Суммах балансовых и внебалансовых задолженностей
- Плановых и фактических сроках их погашения

Форма составляется на первое число каждого месяца.

Из формы были выгружены данные за период с 2015 по 2019 год. Далее были отобраны только те записи, которые соответствуют следующим срезам:

- На 1 января
- На 1 апреля
- На 1 июля
- На 1 октября

Это было сделано для того, чтобы подобрать необходимую частоту наблюдения клиента в течение года. Так, наблюдая клиента помесячно, есть вероятность получить смещенные оценки, поскольку временного интервала в 1 месяц недостаточно для того, чтобы сделать вывод о том, что кредитная оценка заёмщика изменилась. Так, например, финансовую отчетность клиенты сдают поквартально, информация об исках обновляется еще реже, в связи с чем связанные с этими событиями факторы не будут обновлены, а наблюдение в модели будет «лишний раз» учтено, что послужит смещением получаемой для фактора оценки коэффициента регрессии.

Далее временные точки, характеризующие конкретные срезы данных на определенный момент времени, будем называть точками мониторинга или точками наблюдения.

Затем для полученной совокупности наблюдений были выгружены дополнительные данные, с целью формирования целевого сегмента и отсечения нецелевых наблюдений, а именно:

- Риск-сегмент
- Бизнес-сегмент

- Вид деятельности (код)

Поскольку выборка формируется за некоторый исторический промежуток времени и часть информация, привязанная к конкретному клиенту в определенный момент времени могла измениться, важно учесть ретроспективность данных, то есть для каждого наблюдения выгружать такие данные, которые характеризовали бы заёмщика именно в наблюдаемый момент времени. Для этого данные выгружались в соответствие со следующим алгоритмом:

- Для каждого наблюдения определялась последняя релевантная для него запись, соответствующая конкретному клиенту до точки наблюдения.
- В случае, когда такая запись не была найдена (это возможно в случае, когда для строки данных неверно определена дата, в соответствие с которой эту запись можно считать актуальной), бралась первая релевантная для наблюдения запись сразу же после даты наблюдения.

Для определения целевого сегмента модели достаточно взять наблюдения, соответствующие определенным риск- и бизнес-сегментам, а также исключить клиентов, занимающихся определенными видами деятельности. Однако, в силу определенных особенностей АС, не всегда удается определить риск-сегмент заёмщика, поэтому было принято решение доопределить понятие целевого риск-сегмента с помощью использования кредитных рейтингов.

Для этой цели из автоматизированной системы, содержащей расчеты по скоринговым моделям, были выгружены те кредитные рейтинги заёмщиков, которые были рассчитаны по моделям, разработанным для целевого сегмента. Так, в случаях, когда у клиента отсутствовал утвержденный риск-сегмент,

наблюдение все равно считалось целевым в том случае, если клиент имел утвержденный рейтинг, рассчитанный по релевантной для сегмента модели.

Также были получены данные о выявленных эпизодах мошенничества клиентов. Эту информацию важно учесть при разработке модели. Так, например, попадание в выборку клиента, который был замечен в фальсификации финансовой отчетности с целью завышения кредитного рейтинга, вызовет смещение получаемых оценок коэффициентов регрессии для факторов, использующих данные финансовой отчетности клиентов. Поэтому таких клиентов важно исключить из выборки для разработки.

В результате произведенных операций была собрана, насчитывающая порядка 92 тысяч наблюдений.

2.1.2. Определение целевого события.

Для определения целевой переменной из внутренних автоматизированных систем были выгружены данные о событиях дефолта клиентов банка. Целевое событие представляет собой дефолт заёмщика в течение года после даты мониторинга. При этом важно выделить в отдельный сегмент тех заёмщиков, которые уже находились в дефолте на дату наблюдения. Необходимо исключить такие наблюдения, поскольку их присутствие в выборке может вызвать переобучение модели и «ложные» срабатывания.

2.1.3. Выделение выборки для валидации.

Для выделения контрольной части выборки был написан специальный алгоритм, который позволяет:

- Получить непересекающиеся сегменты клиентов в обучающей/контрольной выборках
- Стратифицировать полученное разбиение по:
    - Целевой переменной
    - Году наблюдения
    - Бизнес-сегменту

Первый пункт особенно важен, поскольку при формировании выборок важно разделить их таким образом, чтобы тестирование осуществлялось на клиентах, которых модель еще «не видела», так как это влияет на оценку обобщающей способности модели.

Второй пункт не менее важен, поскольку при выделении выборки для валидации необходимо, чтобы полученные выборки обладали свойством репрезентативности, то есть в достаточной мере описывали генеральную совокупность.

Наборы данных разделялись на обучающую и тестовую выборки в следующем соотношении

- Обучающая выборка – 80% наблюдений
- Контрольная выборка – 20% наблюдений
- Фрод-наблюдения формируют отдельную выборку

Был осуществлен перебор случайных подвыборок общей выборки для разработки, при которых достигается приемлемая стратификация по соотношению разделения количества наблюдений в разрезе анализируемых переменных в каждом фолде относительно всей выборки.

Стратификация набора данных по случайных подвыборкам производилась таким образом, чтобы отношение суммы весов наблюдений в каждом из фолдов к сумме весов в общей выборке максимально соответствовало заданному соотношению, а именно 1:4.

Для процедуры стратификации была использована библиотека StratifiedGroupKFold, которая является собственной надстройкой над классом sklearn сочетающим функционал StrtifiedKFold и GroupKFold.

Поиск репрезентативного разбиения заключается в поиске такой подвыборки, которая давала бы оптимальный в смысле среднеквадратического отклонения результат стратификации для всех признаков сразу. Соответственно, в качестве метрики была использована сумма квадратов разностей между целевым соотношением весов для фолда и соотношением, полученным при разбиении.

Для этого в разрезе подсегментов, факторов и подвыборки, каждая запись присваивалась в одну из пяти групп так, чтобы четыре из них образовывали обучающую часть выборки, а оставшаяся часть – тестовую. Далее, по всем группам и значениям признака считается сумма квадратов разностей между целевым соотношением весов для тестовой выборки и полученным на данной случайно подвыборке соотношением.

2.2. Описание данных

2.2.1. Источники данных

Для моделирования были использованы следующие источники данных:

- Информация об административных и гражданских арбитражных делал юридических лиц РФ

- Информация о финансовой отчетности юридических лиц РФ

2.2.2. Сбор данных о судебных исках

Для разработки модели для каждого наблюдения были выгружены данные по судебным искам до даты мониторинга заёмщика.

В ходе анализа выгруженных данных было обнаружено несколько проблем. В частности, в источнике имеются некоторые проблемы в восстановлении ретроспективности по:

- Данным в части определения состава участников судебного дела
- Данным в части определения суммы иска

В первом случае была проанализирована частота смены ролей участников арбитражных дел, доля таких дел и условная вероятность смены роли участия в деле при условии нахождения в определенном статусе участника. Иными словами, была предпринята попытка ответить на следующий вопрос: «Если указано, что клиент занял определенную сторону участия в судебном процессе, какова вероятность, что она изменится?». Для понимания природы подобных явлений были проанализированы конкретные судебные разбирательства, проведена консультация с экспертом.

Во втором случае было установлено, что доля дел, в которых изменяется первоначальная сумма требований – 8%. В результате был разработан следующий алгоритм приоритезации данных:

1. Для каждого наблюдения были выгружены все судебные дела, зарегистрированные на горизонте 2 лет до даты мониторинга. При этом дела о банкротствах рассматривались отдельно и не были

ограничены глубиной в 2 года, а были взяты безусловно все банкротные дела до даты мониторинга.

2. Для определения релевантного состава участников был придуман следующий механизм приоритезации:
    a. По каждой тройке (ИНН, Дата мониторинга, Судебное дело) берется ближайшая сторона участия до даты мониторинга. В случае отсутствия такой записи, берется ближайшая сторона участия «из будущего».
    b. В случае, если согласно указанному в предыдущем пункте алгоритму определено несколько релевантных сторон участия в судебном деле, итоговая сторона участия определяется в соответствие со следующей приоритетностью сторон:
        i. Ответчик
        ii. Если дело банкротное – Третье лицо, иначе – Истец
        iii. Если дело банкротное – Истец, иначе – Третье лицо.
3. Для определения суммы иска был использован следующий алгоритм приоритезации:
    a. По каждой тройке (ИНН, Дата мониторинга, Судебное дело) берется ближайшая сумма до даты мониторинга. В случае отсутствия такой записи, берется ближайшая сумма «из будущего».
    b. В случае, когда согласно указанному в предыдущем пункте алгоритму определено несколько релевантных сумм иска, приоритезация происходит в порядке убывания суммы иска. Также считаем, что пропуски в поле «Сумма иска» тождественны нулевой сумме.
4. По полученным данным далее формируются агрегаты, участвующие в формировании признаков.

2.2.3. Сбор финансовой отчетности

При разработке модуля были также использованы данные по финансовой отчетности клиентов. При этом выгружалась не вся финансовая отчетность, а та отчетность, которая прошла проверки на качество. Для проверки финансовой отчетности используются следующие факты:

- Единица измерения отчетности известна
- Сумма показателей по отдельным строкам внутри разделов есть сумма ИТОГО, представляющая собой отдельную строку финансовой отчетности
- Ввиду экономического смысла, отдельные строки финансовой отчетности всегда есть величины положительные (в отдельных случаях - неотрицательные)

При этом, для выгрузки самой отчетности, было использовано два источника, один из которых является внутренней автоматизированной системой и имеет меньшую задержку данных, второй же представляет собой внешний источник данных и имеет большую задержку.

В качестве более приоритетного источника ввиду большей актуальности данных была использована внутренняя АС. В случаях, когда данные в первом источнике не были найдены, они выгружались из второго источника.

Поскольку наблюдения имеют квартальную структуру внутри года, а внутренняя автоматизированная система содержит квартальные отчетности с накопительным итогом в течение года, итоговая величина по каждой строке отчетности дополнительно рассчитывалась по принципу Last-Twelve-Months – с учетом последних 12 месяцев. Алгоритм расчета LTM, или алгоритм аннуализации-строк отчетности представлен далее:

1. Если текущая актуальная отчетность – годовая отчетность, то LTM-показатель представляет собой показатель согласно этой отчетности
2. Если текущая актуальная отчетность не является годовой: LTM-показатель = Показатель согласно текущей актуальной отчетности + Показатель согласно актуальной годовой отчетности – Показатель за аналогичный актуальному отчетный период годом ранее
3. Если же показатель за аналогичный отчетный период годом ранее не был найден, LTM-показатель аппроксимируется ближайшей годовой отчетностью
4. В случае, когда и ближайшая годовая отчетность не была найдена, LTM-показатель аппроксимируется мультипликацией пропорционально периоду текущей отчетности.

Также на данных внутренней АС дополнительно был рассчитан мультипликатор EBITDA – прибыль компании до вычета процента по кредитам, налога на прибыль и амортизации по основным нематериальным активам.

Дополнительно были отсечены данные с отрицательной результирующей аннуализированной выручкой.

2.2.4. Качество данных

Данные по финансовой отчетности содержат лишь незначительное число пропусков, которые были обработаны отдельно. Пропуски обусловлены отсутствием актуальной финансовой отчетности, согласно описанному в предыдущем пункте механизму приоритезации, а также ограничением на дату актуальности финансовой отчетности.

При этом пропусков в данных по арбитражным искам нет и не может быть, поскольку такая ситуация приравнивается к отсутствию исков, и, соответственно, нулевым значениям в соответствующих агрегатах, таких как, например, сумма или количество исков.

Информация о пропусках в разрезе типа выборки описана в Таблице 1:

Таблица 1. Информация о пропусках в выборке для разработки

Выборка	**Количество наблюдений**	**Количество пропусков**	**% пропусков**
Обучающая	75699	287	0,38%
Контрольная	18930	112	0,59%

На этапе применения модели информация по судебным искам и финансовой отчётности будет доступна по всем клиентам.

2.3. Однофакторный анализ.

2.3.1. Длинный список факторов

На этапе однофакторного анализа совместно с экспертом в прикладной области был сформирован, так называемый, длинный список факторов модели. Длинный список факторов – это список таких факторов, которые в ходе однофакторного анализа будут подвержены ряду статистических проверок, результат которых определит, является ли фактор пригодным для включения в модель, или же он будет исключен из дальнейшего рассмотрения. В целях избежания раскрытия информации о коммерческой разработке внутри организации, считаем далее, что состояние заёмщика в конкретный момент времени описывают обезличенные риск-факторы: фактор 1, фактор 2, фактор 3, …, фактор n. Первоначально был сформирован длинный список, состоящий из 45 факторов.

2.3.2. Преобразование факторов

Для каждого фактора из длинного списка было проведено разбиение наблюдений выборки на группы. Количество групп и их границы подбирались по принципу максимизации коэффициента Gini на обучающей выборке. Также полученное преобразование должно отвечать ряду требований:

- Монотонность разбиения
- Статистическая значимость на 5% уровне. Под статистической значимостью в этом случае понимается значимое отличие средних значений между соседними группами.

Для расчета коэффициента Gini и дальнейшей обработки факторов было применено WOE-преобразование (Weight of Evidence). Данная процедура позволяет поставить в соответствие группам значений фактора некоторый скоринговый балл. Значение балла для i-ой группы фактора рассчитывалось по формулам 2.1, 2.2:

$$WOE_i = \ln\left(\frac{N_{good}(i)/N_{good}}{N_{bad}(i)/N_{bad}}\right), \text{если } N_{good}(i) \neq 0, N_{bad}(i) \neq 0 \quad (2.1)$$

$$WOE_i = \ln\left(\frac{(N_{good}(i) + 0.5)/N_{good}}{(N_{bad}(i) + 0.5)/N_{bad}}\right), if\ N_{good}(i) = 0 \text{ или } N_{bad}(i) = 0 \quad (2.2)$$

Где $N_{good}(i), N_{good}$ – количество недефолтных наблюдений в i-ой группы и на всей выборке соответственно, $N_{bad}(i), N_{bad}$ – количество дефолтных наблюдений в i-ой группе и на всей выборке соответственно.

Преобразованный фактор представляет собой ступенчатую функцию, где диапазону значений фактора соответствует WOE-значение этого диапазона.

На рис.2.1 приведен пример WOE-преобразованного фактора, отвечающего требованию монотонности – с ростом значения фактора уровень дефолтов внутри каждой группы монотонно возрастает как на тестовой, так и на тренировочной выборках. Уровень дефолта при этом в последней группе, выделенной по принципу отсутствующей информации о значении фактора, не учитывается. Уровень дефолтов на обучающей выборке представлен синей кривой, на обучающей – красной кривой. Однако такого удается достичь не всегда. Дело в том, что разбиение «обучается» на тренировочной выборке, следовательно, монотонное поведение на тестовой выборке не гарантируется. На рис.2.2 приведен пример фактора, не отвечающего требованию монотонности.

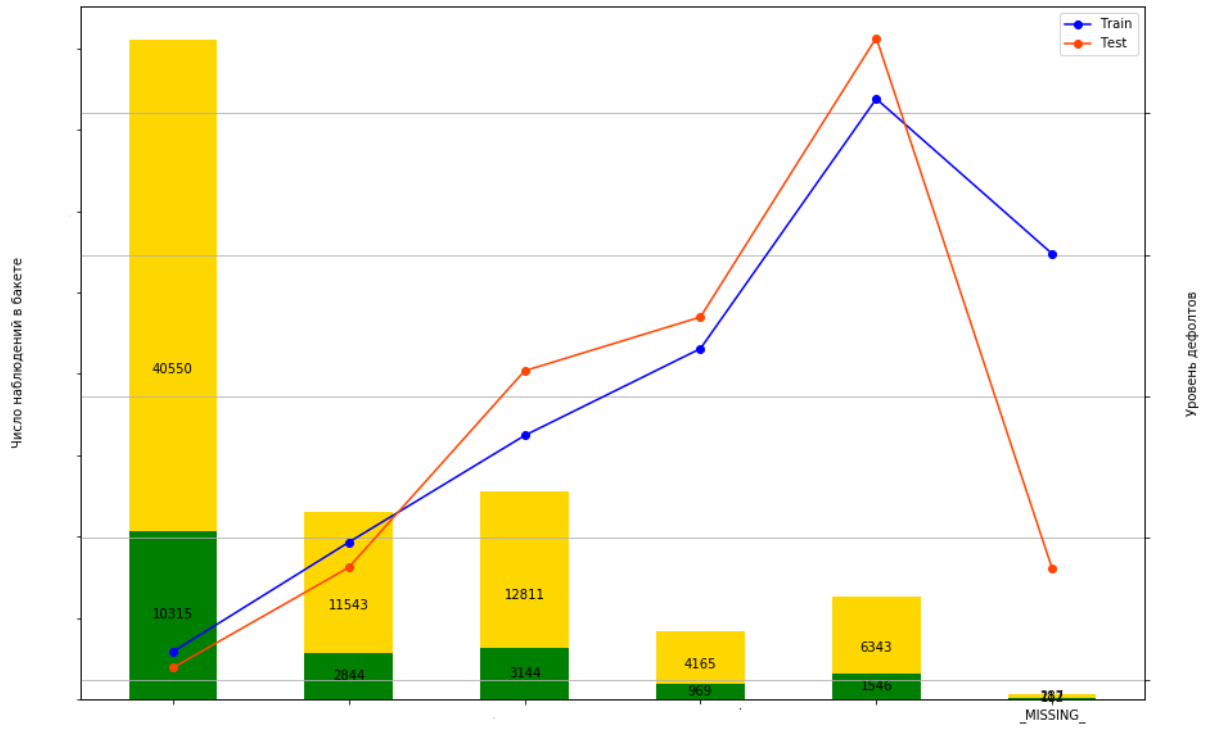


Рис. 2.1 Пример корректного WOE-разбиения фактора

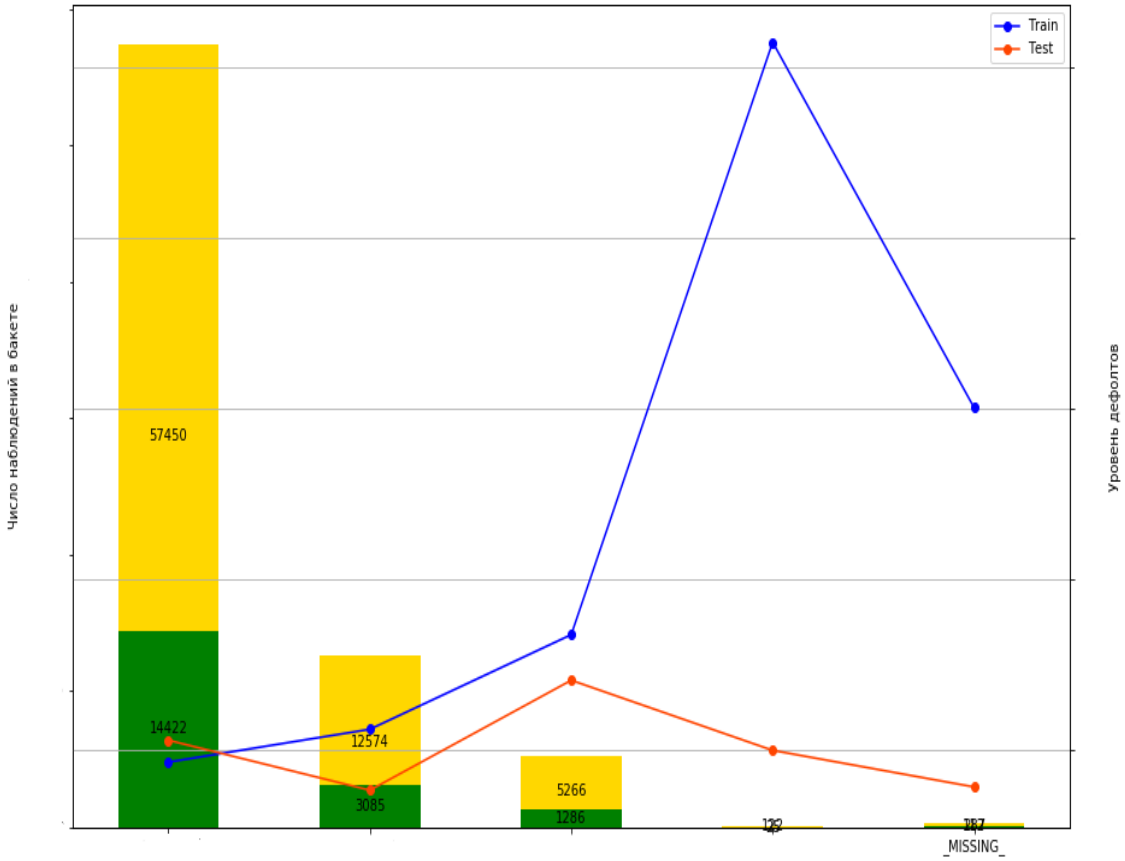


Рис 2.2 Пример плохого разбиения фактора

2.3.3. Основные статистические метрики

Для каждого фактора были рассчитаны следующие статистические показатели:

1. Коэффициент Джини:
    a. На обучающей выборке
    b. На контрольной выборке
2. Индекс стабильности популяции (PSI):
    a. Отдельно только по дефолтным наблюдениям
    b. Отдельно только по недефолтным наблюдениям

   c. По всем наблюдениям без разделения по целевому признаку
3. Статистическая значимость фактора (P-value):
   a. На обучающей выборке
   b. На контрольной выборке
4. Абсолютное и относительное изменение на контрольной выборке относительно обучающей
5. Коэффициент вариации коэффициента Джини по годам:
   a. на обучающей
   b. и контрольной выборках
6. Значение коэффициента Джини за последние 12 месяцев

Дополнительно проверялось, что:

1. Коэффициент перед фактором не меняет знак на контрольной выборке относительно обучающей.
2. Не наблюдается промежутков немонотонности WOE-преобразованного фактора, или, по крайней мере, если таковые имеются, они не попадают в «красную зону» согласно актуальной методике валидации.
3. Наличие монотонной зависимости уровня риска от значений фактора и соблюдение бизнес-логики.

Далее были установлены критерии исключения факторов из длинного списка. Невыполнение хотя бы одного условия влечет исключение фактора из короткого списка:

1. Коэффициент Джини на обучающей выборке > 5 (в абсолютном выражении)
2. Коэффициент вариации коэффициента Джини < 30

3. Абсолютное снижение коэффициента Джини на контрольной выборке менее 5 или аналогичное относительное снижение менее 15%
4. Статистическая значимость фактора, а именно рассчитанное P-value, на контрольной и обучающей выборках не превышает уровень значимости в 5%.

Для проверки всех условий был написан фреймворк, позволяющий рассчитать и визуализировать все вышеупомянутые метрики. На рис.2.3 представлен фактор, удовлетворяющий всем требованиям. На рис.2.4 показан фактор, не обладающий должной ранжирующей способностью. Кроме того, фактор оказывается незначим на контрольной выборке. Полученное значение P-value сильно превышает порог.

	Значение
Влияние фактора	—
Смена знака	Нет
Gini Train	29.07
Gini Train LTM	31.23
Gini Train VAR	5.82
Gini Test	33.51
Gini Test LTM	32.45
Gini Test VAR	5.73
Gini Increase	4.44
P-value Train	0.000
P-value Test	0.000
PSI All	0.002
PSI Good	0.002
PSI Bad	0.000

Рис. 2.3 Фактор, удовлетворяющий всем статистическим метрикам

	Значение
Влияние фактора	—
Смена знака	Нет
Gini Train	4.32
Gini Train LTM	-2.42
Gini Train VAR	111.75
Gini Test	-0.64
Gini Test LTM	8.81
Gini Test VAR	502.90
Gini Increase	-4.97
P-value Train	0.000
P-value Test	0.691
PSI All	0.001
PSI Good	0.001
PSI Bad	0.000

Рис. 2.4 Фактор, показывающий плохое качество ранжирования

В результате анализа с учетом порогов отсечения факторов, приведенных ранее был получен так называемый короткий список, состоящий из 18 факторов, которые приведены в Таблице 2:

Таблица 2. Ранжирующая способность шорт-листа факторов

Фактор	**Gini (%)**
Фактор 1	22,69
Фактор 2	22,32
Фактор 3	30,34
Фактор 4	30,52

Продолжение таблица 2.

Фактор 5	37,07
Фактор 6	31,66
Фактор 7	33,97
Фактор 7	35,45
Фактор 8	28,20
Фактор 9	29,10
Фактор 10	26,30
Фактор 11	27,07
Фактор 12	29,07
Фактор 13	26,12
Фактор 14	27,16
Фактор 15	28,36
Фактор 16	25,19
Фактор 17	25,67

2.4. Многофакторный анализ

В процессе моделирования было принято решение строить модель только на наблюдениях без пропусков в данных, поскольку совместное моделирование переменных с несутевыми пропусками, таких как, например, отсутствие данных, может привести к получению смещенных оценок. Оценка вероятности дефолта наблюдений с пропусками данных отдельно

рассчитывалась как средний Default Rate по наблюдениям с пропусками на обучающей выборке. Количество таких наблюдений составило 287. Средний Default Rate по наблюдениям с пропусками составил 8%, что выше, чем средний Default Rate по портфелю.

Так, для наблюдений без пропусков в данных выборка для моделирования имеет структуру, описанную в Таблице 3:

Таблица 3. Количество наблюдений в выборке для разработки

Выборка	**Количество наблюдений**
Обучающая [Train]	75412
Контрольная [Out-Of-Sample]	18818

2.4.1. Оценка корреляций

Для анализа взаимозависимости были рассчитаны корреляционные матрицы для преобразованных факторов. Порог, начиная с которого коэффициент корреляции считался высоким, был зафиксирован на уровне 0.7. Так, из пар факторов, показывающих абсолютное значение корреляции более 0.7, один фактор исключался. Выбор исключаемого фактора основывался на сравнении ранжирующей способности в смысле коэффициента Джини на обучающей части выборки.

Дополнительно для оценки корреляций для каждого фактора был расчитан коэффициент вздутия дисперсии VIF Пороговое значение для отсечения фактора было установлено на уровне VIF равном 3, однако коэффициенты вздутия дисперсии по всем факторам не превосходили заданное значение.

2.4.2. Выбор модельной формы связи

Далее многофакторный анализ проводился с использованием метода логистической регрессии, который позволяет определить веса объясняющих факторов, обеспечивающих наибольшую предсказательную силу модели. В качестве целевой переменной выступал флаг-индикатор дефолта, в качестве объясняющих переменных – факторы из короткого списка после исключения коррелятов.

2.4.3. Отбор признаков в модель

Для отбора факторов в модель к результирующему списку факторов был применен алгоритм StepwiseSelection. Процедура отбора признаков StepwiseSelection представляет собой следующий алгоритм:

1. *Инициализация*: список включенных в модель переменных объявляется пустым. Уровень значимости для включения в модель – 1%. Уровень значимости для исключения – 5%.
2. *Шаг включения*: далее для каждой переменной строится модель логистической регресии и подсчитывается значение P-value. Из всех не включенных в модель переменных выбирается та, которая имеет наименьшее значение. Далее это значение сравнивается с пороговым значением на включение в модель. В случае, если полученное значение P-value не превосходит пороговое значение, переменная включается в модель.
3. *Шаг исключения*: из всех включенных в модель переменных выбирается та, которая имеет наибольшее значение P-value. Это значение сравнивается с пороговым значением на исключение из модели. В случае, если расчетное значение P-value превосходит пороговое, переменная исключается.
4. Шаги 2-3 выполняются итеративно до тех пор, пока состав включенных в модель переменных изменяется внутри одной итерации. Если в

результате шагов 2-3 список включенных в модель факторов не изменился, алгоритм завершается.

В результате процедуры отбора признаков в модели осталось 3 фактора.

Для модельного списка факторов также проверялся тот факт, что статистическая значимость факторов, а именно рассчитанное значение P-value не превышает 0.05. Статистическая значимость модельных факторов (значение P-value) представлена в Таблице 4:

Таблица 4. Статистическая значимость модельных факторов

Фактор	**P-value**
Фактор 1	0.00
Фактор 2	0.00
Фактор 3	0.00

Все факторы отвечают требованию статистической значимости полученных оценок.

2.4.4. Итоговая модель

В результате была построена модель логистической регрессии с тремя независимыми переменными. Коэффициент Джини модели на выборке для валидации составил 43.7%, что превышает аналогичное значение на обучающей выборке на 5 пунктов. Это говорит о хорошей обобщающей способности модели и отсутствии переобучения.

Основные метрики, по которым оценивалось качество модели, приведены в Таблице 5:

Таблица 5. Характеристики финальной модели

Характеристики финальной модели на WOE-преобразованных факторах	
Джини на обучающей выборке	38.77%
Джини на контрольной выборке	43.75%
PSI Скорбалла	0.003
Коэффициент вариации Gini на обучающей выборке	3.27%
Коэффициент вариации Gini на контрольной выборке	12.58%

Дополнительно исследовалось, что модель ведет себя стабильно на всех временных срезах. С этой целью были вычислены коэффициенты вариации коэффициента Джини, значения которых представлены в Таблице. Также был построен график, отражающий изменение метрики на каждом временном срезе с 95% доверительным интервалом. График представлен на рис.2.5.

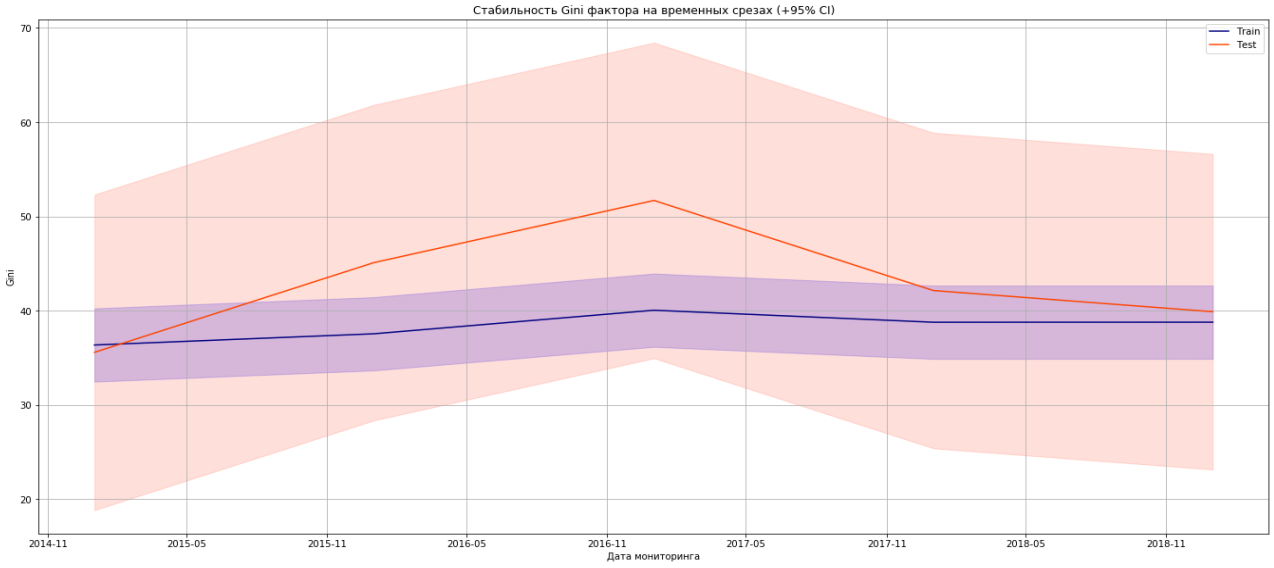


Рис. 2.5 Стабильность ранжирования модели на временных срезах

Исходя из расчетных значений коэффициента вариации и визуализации, можно сделать вывод, что модель ведет себя стабильно на всем временном отрезке.

На рис.2.6 приведена ROC-кривая модели. Зеленая кривая была построена на обучающей выборке, красная – на контрольной выборке.

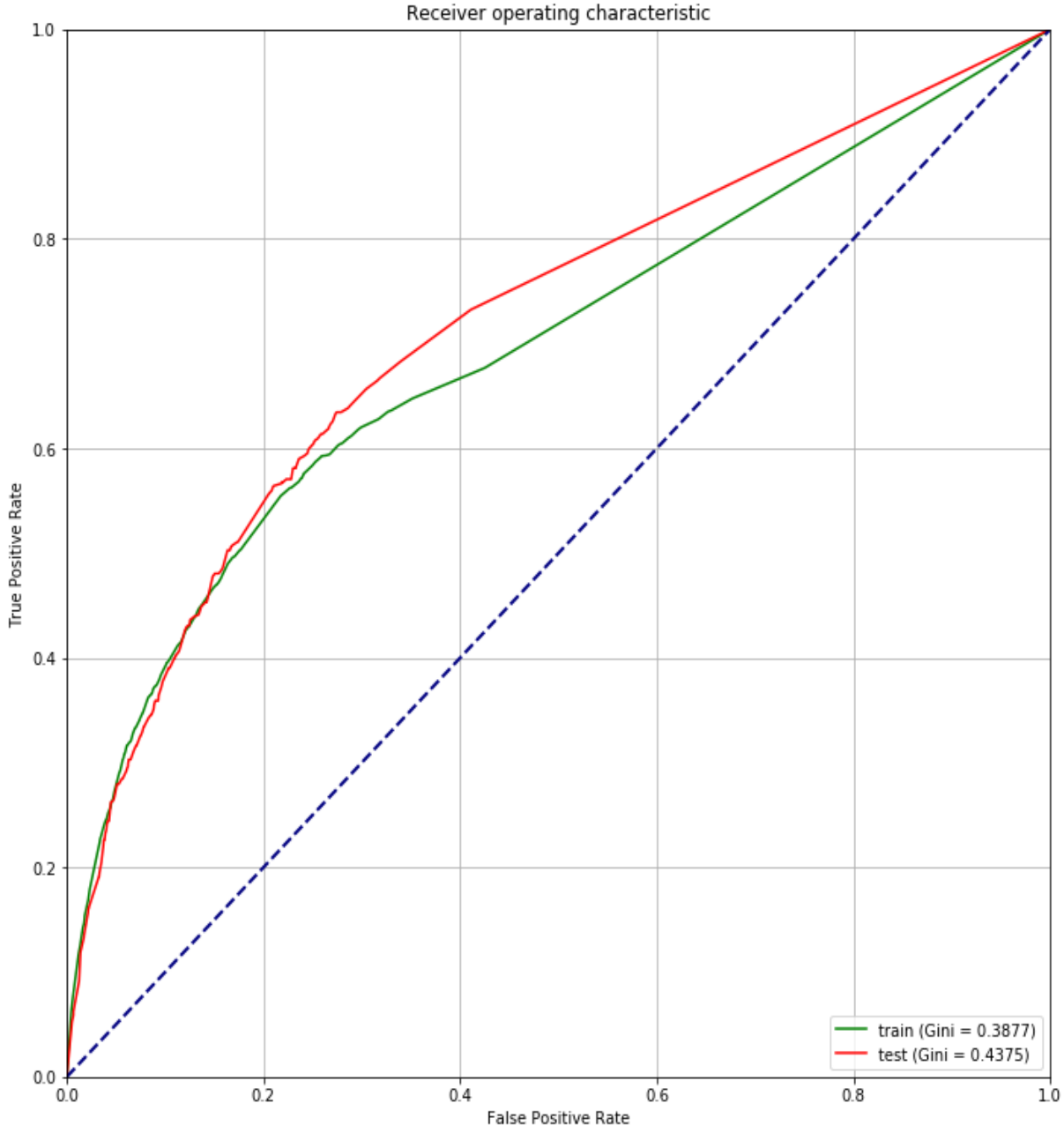


Рис. 2.6 ROC-кривая модели

Дополнительно проверялось, что все модельные факторы соответствуют критериям, предъявляемым к факторам на этапах однофакторного и многофакторного анализа. Так, все модельные факторы удовлетворяют необходимым требованиям. Расчетные значения статистических метрик представлены на рис.2.7.

	Смена знака	P-value Train	P-value Test	VIF Train	VIF Test
Фактор 1	Нет	0.000	0.000	2.081	2.127
Фактор 2	Нет	0.000	0.000	1.947	2.044
Фактор 3	Нет	0.000	0.000	1.277	1.282

Рис. 2.7 Сводная статистика по модельным факторам

2.4.5. Альтернативное моделирование

В ходе процесса моделирования были также предприняты попытки построить лучшую модель, используя иные алгоритмы классификации, такие как решающее дерево, случайный лес, градиентный бустинг, перцептрон Розенблатта. Однако ни один из этих алгоритмов не показал аналогичную алгоритму логистической регрессии ранжирующую способность при схожей сложности модели. Под сложностью модели подразумевается ее структура, количество параметров.

Так, результат, достигнутый посредством использования алгоритма логистической регрессии всего на трех переменных, достигается при применении алгоритма градиентного бустинга только лишь при использовании 12 факторов, при этом теряется важнейшее свойство модели

для последующего анализа: способность прозрачной интерпретации влияния факторов модели на итоговый скорбалл.

ЗАКЛЮЧЕНИЕ

Модель оценки вероятности дефолта позволяет в автоматическом режиме оценивать уровень кредитного риска заёмщика в процессе принятия решения по выдаче кредита, тем самым исключив из нее экспертную составляющую, что позволяет значительно ускорить принятие решения.

В данной работе реализована модель оценки кредитного риска, использующая информацию всего лишь о трех риск-факторах.

С помощью экспертов в предметной области был сформирован список факторов, описывающих событие дефолта заёмщика, которые были подвергнуты ряду статистических и экономических проверок, позволивших отсеять непригодные для моделирования факторы.

С использованием алгоритма последовательного включения факторов в модель был сформирован итоговый список факторов и оценены коэффициенты модели логистической регрессии.

Модель показала высокие результаты точности, заметно превышающие результаты предыдущей модели, используя при этом значительно меньшее количество информации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Деан Фантаццини. Управление кредитным риском // Прикладная эконометрика 4(12) 2008: 84-137.

[2] Сорокин А.С. Построение скоринговых карт с использованием модели логистической регрессии // Интернет-журнал «Науковедение» 2(3) 2014: 1-29.

[3] Ковалев М., Корженевская В. Методика построения банковской скоринговой модели для оценки кредитоспособности физических лиц // Банки Казахстана 1 2008: 43-48

[4] Карминский А.М. Кредитные рейтинги и их моделирование. М.: НИУ ВШЭ, 2015.

[5] Карминский А.М., Лозинская А.М. Оценка кредитного риска при ипотечном жилищном кредитовании // XV Апрельская международная научная конференция по проблемам развития экономики и общества: в 3 кн. Кн. 1 / Отв. ред.: Е.Г. Ясин. М.: Изд. дом НИУ ВШЭ. 2015. С. 353-366.

[6] Елфимова И.Ф. Управление кредитными рисками коммерческого банка. https://cyberleninka.ru/article/n/upravlenie-kreditnymi-riskami-kommercheskogo-banka-2

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)»

Институт *№ 8 «Информационные технологии и прикладная математика»* **Кафедра** 810Б

Направление подготовки 02.04.02 ФИИТ **Группа** М8О-203М-19

Квалификация (степень) *магистр*

РАЗДАТОЧНЫЙ МАТЕРИАЛ
К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ МАГИСТРА
(МАГИСТЕРСКОЙ ДИССЕРТАЦИИ)

На тему: «Применение методов машинного обучения при моделировании вероятности дефолта в задаче оценки кредитного риска»

Автор диссертации Лопатенко Валентин Васильевич _____
(Фамилия, имя, отчество) (подпись)

Научный руководитель Абгарян Каринэ Карленовна _____
(Фамилия, имя, отчество) (подпись)

Рецензент Думин Павел Николаевич _____
(Фамилия, имя, отчество) (подпись)

К з а щ и т е д о п у с т и т ь

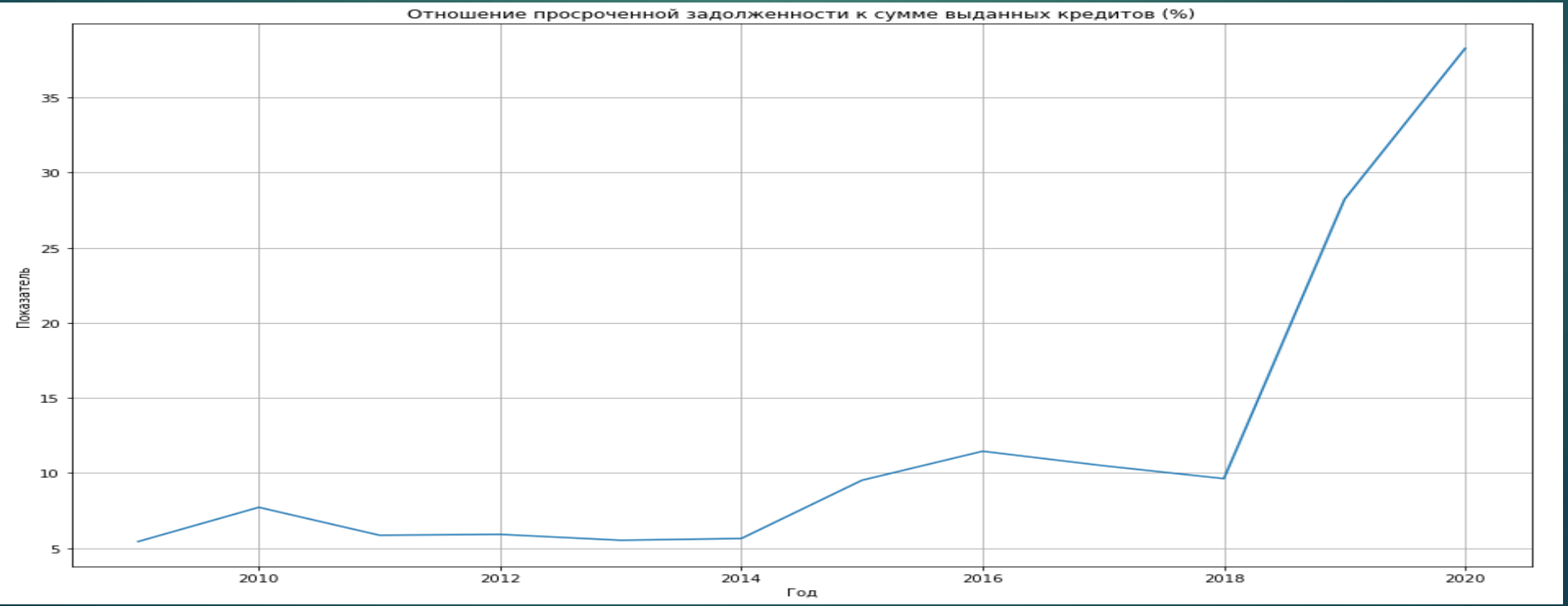
Зав. кафедрой 810Б Абгарян Каринэ Карленовна _____
(№ каф.) (фамилия, имя, отчество полностью) (подпись)

“ 24 ” мая 2021г.

Москва 2021

Актуальность темы исследования

1. Необходимость в модели автоматического скоринга
2. Необходимость учитывать новые дефолты, порожденные пандемией COVID-19 в 2020 году
3. Периодическая валидация показала, что разработанная ранее модель неспособна корректно ранжировать новые наблюдения



Постановка задачи

- По имеющейся информации о заёмщике – данным о финансовой отчетности и судебных исках, требуется оценить вероятность дефолта на горизонте в 1 год. Иначе говоря, необходимо восстановить зависимость вида:

$$P(y = 1) = \mathrm{f}(\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{n}), \text{где}$$

- X_i- факторы модели
- y – целевое событие – дефолт заёмщика в течение 1 года после наблюдения

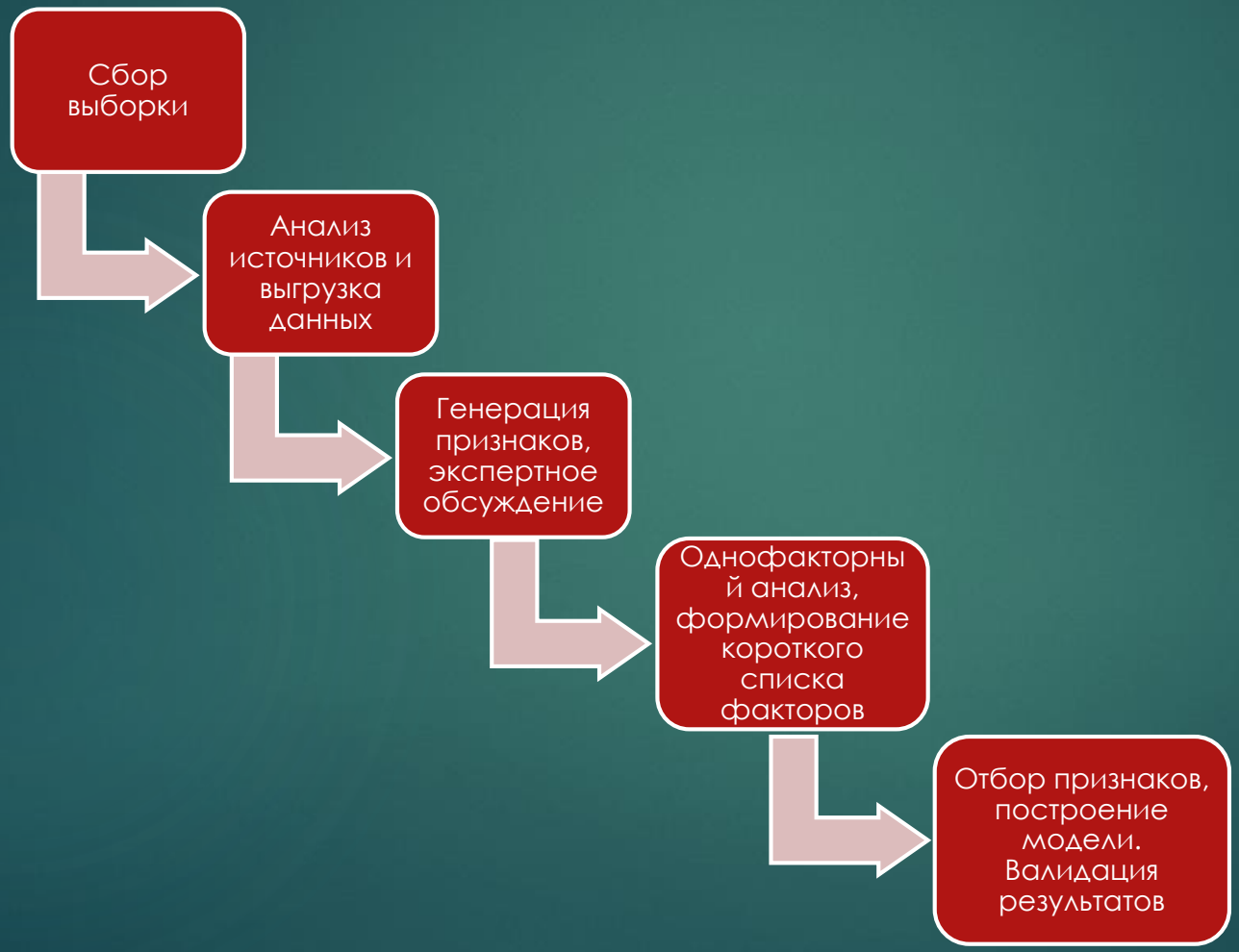
Методы решения

- Основной метод решения – метод логистической регрессии с применением дискретного монотонного преобразования факторов – WOE-преобразования. Метрика качества – коэффициент Джини.
- Достоинства метода:
  - Интерпретируемость и прозрачность с экономической точки зрения
  - Полное соответствие бизнес-логике работы факторов
  - Статистическая значимость получаемых оценок
  - Богатый арсенал верификации полученных результатов
- Недостатки:
  - Более трудоемкая процедура построения (в сравнении с остальными моделями ML)
  - Линейность

Программная реализация

- Разработка велась на Hadoop-кластере
- Язык разработки – Python 3.5.3
- Основной фреймворк для работы с данными – pyspark
- Основные библиотеки, используемые при моделировании:
  - statsmodels
  - pandas
  - sklearn
- Среда разработки – Jupyter HUB
- Система контроля версий – git (BitBucket)

Этапы построения модели



Методы решения: WOE-преобразование

- Переменная разбивается на интервалы
- Для каждого интервала считается значение WOE по формуле:

$$WOE_i = \ln(\frac{freq\_goods_i}{freq\_bads_i}), \text{где}$$

$freq\_goods$ - доля наблюдений внутри интервала, для которых не наступило целевое событие

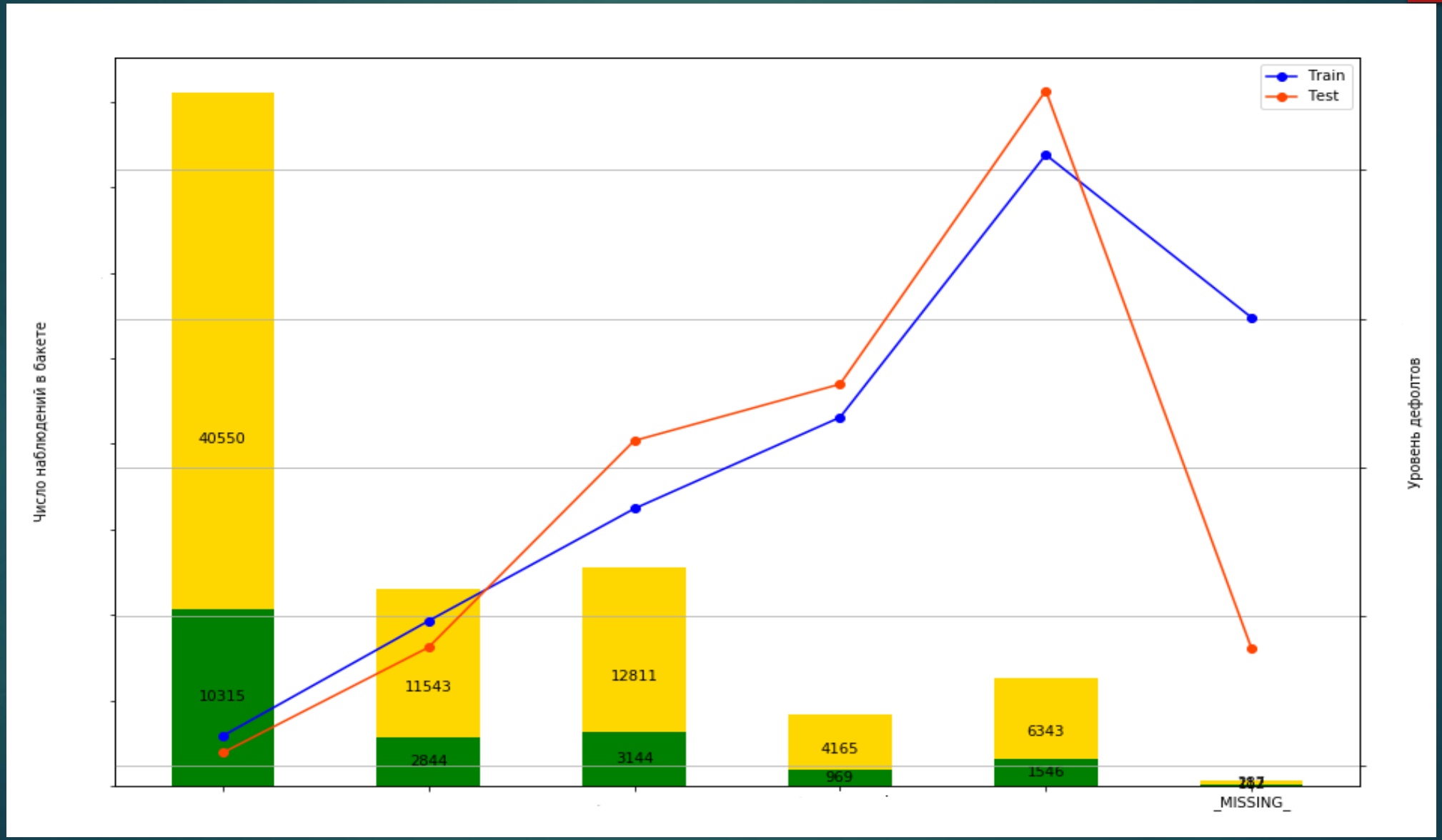
$freq\_bads$ - доля наблюдений внутри интервала, для которых целевое событие наступило

- Проблема – найти оптимальные границы интервалов так, чтобы полученное преобразование отвечало следующим требованиям:

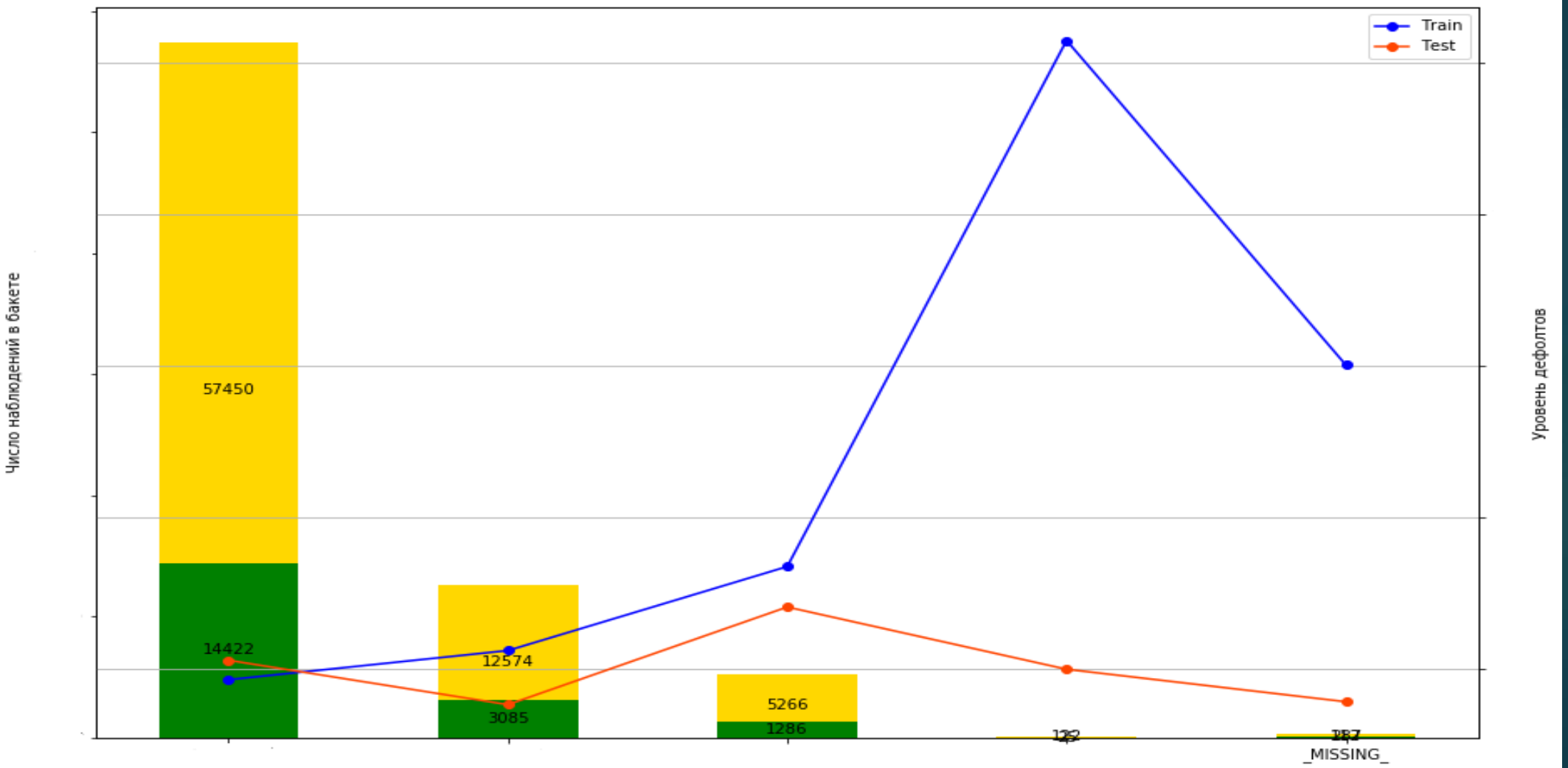
1. Монотонность
2. Оптимальность с точки зрения целевой метрики

- Решение – использование дерева решений для поиска оптимального разбиения

Пример: WOE-преобразование



Пример: WOE-преобразование



Пример: расчет метрик

	Значение
Влияние фактора	—
Смена знака	Нет
Gini Train	29.07
Gini Train LTM	31.23
Gini Train VAR	5.82
Gini Test	33.51
Gini Test LTM	32.45
Gini Test VAR	5.73
Gini Increase	4.44
P-value Train	0.000
P-value Test	0.000
PSI All	0.002
PSI Good	0.002
PSI Bad	0.000

	Значение
Влияние фактора	—
Смена знака	Нет
Gini Train	4.32
Gini Train LTM	-2.42
Gini Train VAR	111.75
Gini Test	-0.64
Gini Test LTM	8.81
Gini Test VAR	502.90
Gini Increase	-4.97
P-value Train	0.000
P-value Test	0.691
PSI All	0.001
PSI Good	0.001
PSI Bad	0.000

Многофакторный анализ

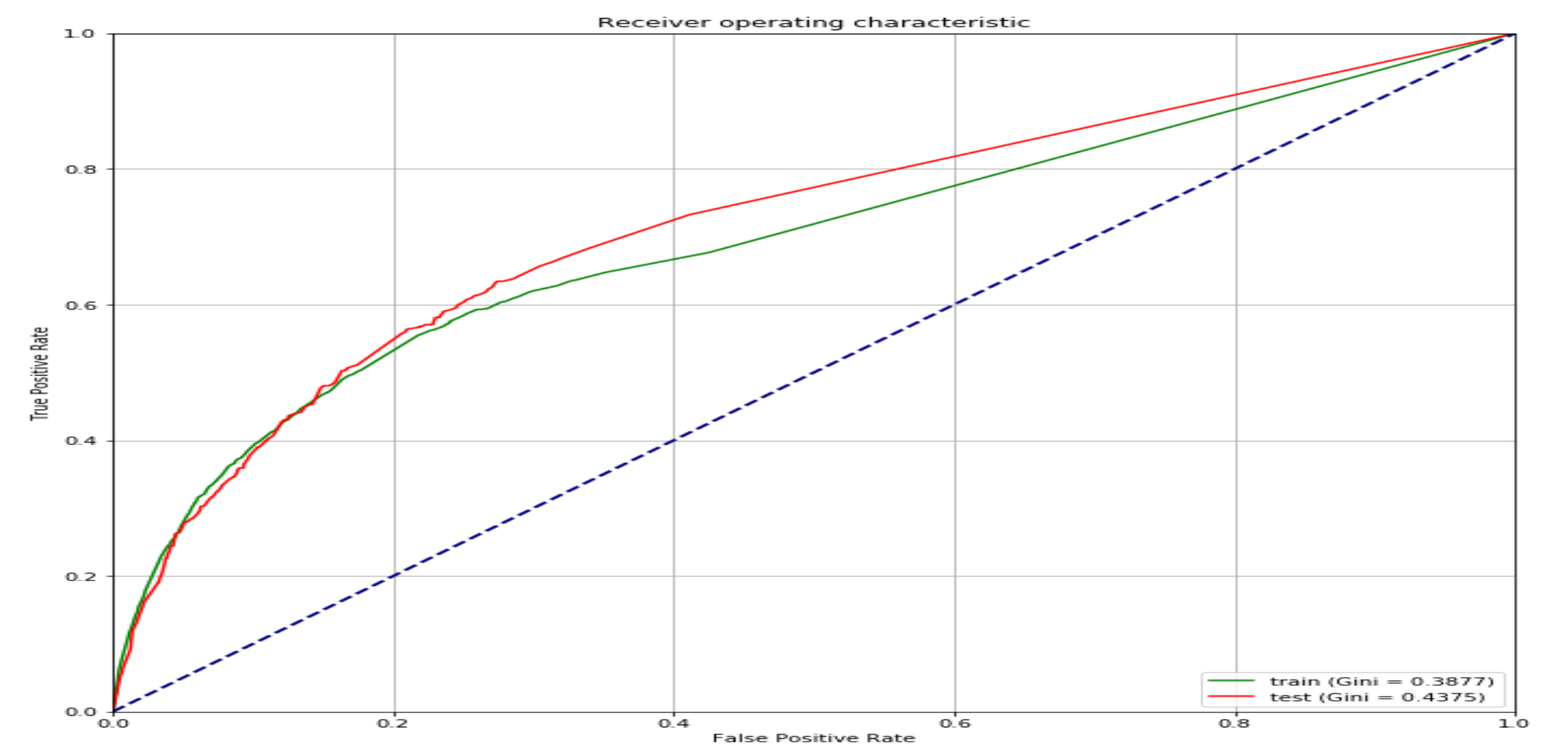
- Количество наблюдений в выборке:

Выборка	Количество наблюдений
Обучающая [Train]	75412
Контрольная [Out-Of-Sample]	18818

- Итоговые характеристики модели:

Характеристики финальной модели на WOE-преобразованных факторах	
Джини на обучающей выборке	38.77%
Джини на контрольной выборке	43.75%
PSI Скорбалла	0.003
Коэффициент вариации Gini на обучающей выборке	3.27%
Коэффициент вариации Gini на контрольной выборке	12.58%

ROC-кривая


Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
train (Gini = 0.3877)
test (Gini = 0.4375)

Резюме - новые подходы

1. Вместе с экспертом проработан ряд новых гипотез. Результат – полностью обновленный длинный список факторов.
2. Написана библиотека автоматизированного расчета статистических показателей для оценки свойств отдельных факторов на этапе однофакторного анализа.
3. Иной подход к отбору факторов – использование алгоритма StepwiseSelection. Результат – сокращение количества модельных факторов до 3 с сохранением статистических свойств и сильной ранжирующей способности.
4. Кумулятивный результат: Качество модели возросло на 14п. в терминах коэффициента Gini.

Спасибо за внимание!