12TH INTERNATIONAL GSOM ECONOMY&MANAGEMENT CONFERENCE 2025 (EMC 2025)

Special track «Financial innovation and risk management»

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ТРАНЗАКЦИОННЫХ МОДЕЛЕЙ И МОДЕЛЕЙ ОКРУЖЕНИЯ НА КРЕДИТНЫЕ РИСКИ ЗАЕМЩИКОВ - ЮРИДИЧЕСКИХ ЛИЦ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Пермяков Н.П., Моргунов А.В.

МФТИ

Проблема / исследовательский вопрос

Традиционные методы кредитного скоринга не всегда эффективны, поскольку они могут не учитывать сложные и динамичные паттерны поведения заемщиков, что ведет к повышенным финансовым рискам для банков и других кредитных институтов.

Необходимость снижения финансовых рисков для банка, рассматривая не только индивидуальные показатели заёмщиков, но и экономические связи между ними.

МФТИ

Цель исследования

Разработка интегральной модели на основе трёх модулей для оценки кредитоспособности юридических лиц

Объект исследования:

Кредитные риски.

Предмет исследования:

Алгоритмы и методы машинного обучения, направленные на анализ и оценку кредитоспособности юридических лиц.

Новизна

Заключается в создании модели, интегрирующей анализ кредитоспособности юридических лиц с оценкой вероятности дефолта на основе анализа групповых экономических связей и методов машинного обучения. Разработка такой модели актуальна для коммерческих банков и иных финансовых институтов, поскольку повышение точности кредитного скоринга и минимизация рисков дефолта являются ключевыми задачами финансового рынка. Модульная архитектура модели позволяет использовать её компоненты как автономно, так и в комплексе, в зависимости от доступности данных и качества метрик отдельных подмоделей, Такой проект отвечает текущим требованиям финансового рынка к повышению точности кредитного скоринга и снижению рисков дефолта юридических лиц.

МФТИ

Актуальность

Обусловлена возрастающим спросом на экономически эффективные и надежные методы оценки кредитоспособности в условиях роста кредитных рисков, вызванного динамикой рыночной среды. Определение кредитоспособности заёмщиков критически важно для бизнеса, так как позволяет минимизировать финансовые потери от невозврата кредитов и оптимизировать управление капиталом.

Задачи исследования

- Анализ существующих подходов к оценке кредитоспособности юридических лиц.
- Разработка отдельных модулей для анализа финансовых данных и данных по исковым делам.
- Построение моделей машинного обучения для каждого из разработанных модулей.
- Интеграция модулей в единую систему с учетом взвешивания значимости каждой модели.
- Оценка эффективности разработанной системы.

МФТИ

Методология исследования

- Сбор и предобработка данных: очистка, заполнение пропусков, нормализация, и добавление дополнительных данных из дополнительных источников.
- Анализ данных: исследование структуры данных, выявление закономерностей и взаимосвязей.
- Разработка и тестирование моделей машинного обучения для каждого модуля на основе выбранных алгоритмов.
- Взвешивание значимости моделей и их интеграция в единую систему оценки кредитоспособности.
- Оценка эффективности системы на основе сравнения с традиционными методами и анализа её способности прогнозировать кредитные риски, а так же с текущими участниками рынка

Дизайн эксперимента

Цель исследования:

- Подтверждение эффективности комплексного подхода

Этапы исследования:

- Сбор и предобработка данных (финансовые, транзакционные, исковые данные).

Разработка модулей:

- Финансовый модуль (данные из Спарк-Интерфакс).
- Транзакционный модуль (закрытые данные от банков).
- Исковой модуль (данные с ресурса «Право.ru»).
- Построение моделей машинного обучения (логистическая регрессия, CatBoost).
- Интеграция модулей в единую систему.
- Оценка эффективности (метрика Gini, бутстрап-анализ).

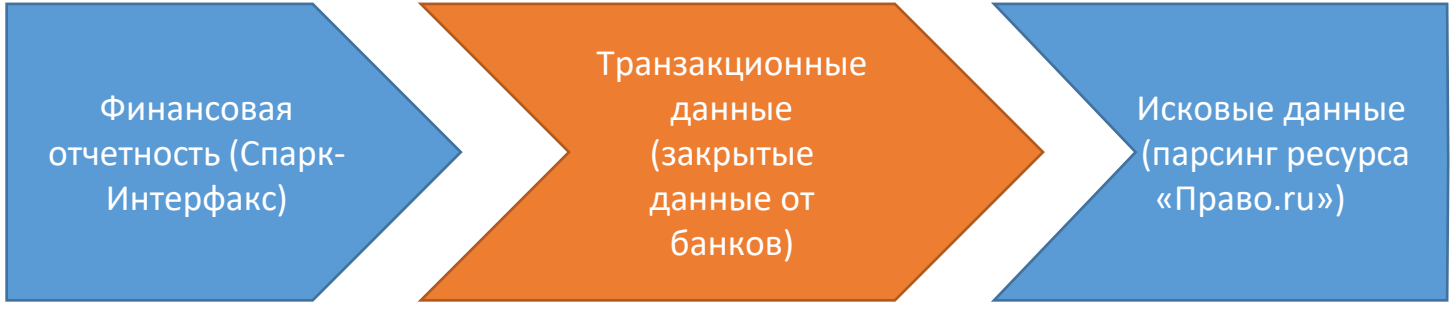
Методология:

- WOE-трансформация и анализ монотонности (для улучшения предсказательной силы каждой из модели по отдельности).
- Тест Вальда для проверки значимости коэффициентов.

МФТИ

Источники данных для исследования

Данные были собраны по клиентам сегмента малого и среднего бизнеса (МСБ) с годовой выручкой от 30’000’000 руб. до 150’000’000 руб. за период с 2013 по 2024 гг. включительно.



МФТИ

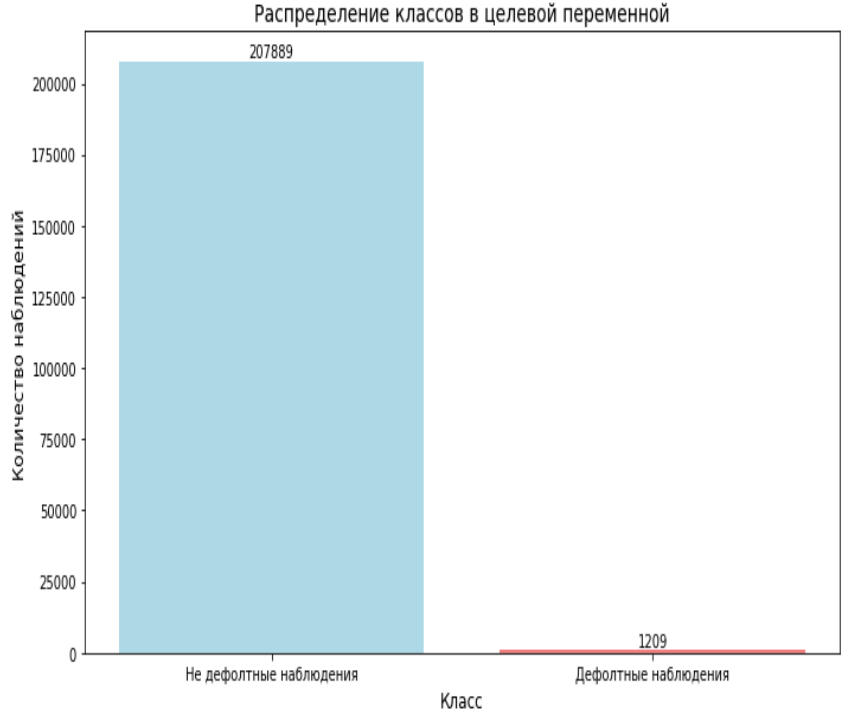
Распределение данных

Из представленной диаграммы видно, что у нас присутствует дисбаланс классов. При этом, важно понимать, что уровень дефолтов в выборке для разработки модуля, основанного на финансовых данных, а также других модулей соответствует фактическому уровню дефолтности на рынке по клиентам МСБ.

Данный набор данных включал в себя 363 столбца, среди которых присутствовала целевая переменная, и охватывал 209098 наблюдений.

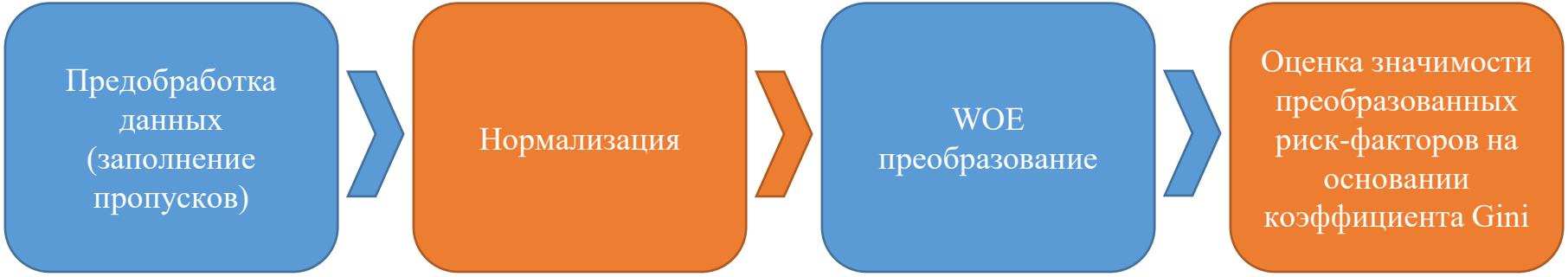
Сама же целевая переменная является бинарной и была получена из следующих источников:

- Единого федерального реестра сведений о банкротстве;
- Газета «Коммерсантъ»



МФТИ

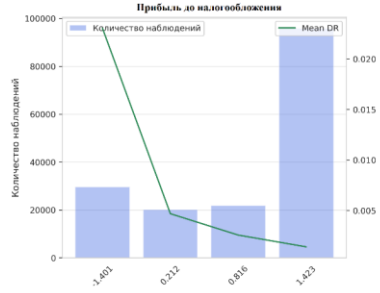
Обработка данных



1. Удаление данных на основании пропусков: финансовый модуль порог 40%, транзакционный модуль – 60%.
2. Был проведен однофакторный анализ. Для каждого фактора был рассчитан коэффициент Gini: он позволяет оценить предсказательную силу и значимость взаимосвязи между риск-фактором и целевой переменной.
3. Пропуски заполнялись следующим образом: все наблюдения группировались по клиентам, если в группе было хотя бы 1 наблюдение, то заполнялись средними значениями, в противном случае нулями.
4. Был проведен многофакторный анализ на основе WOE-преобразования.
5. Был проведен тест на монотонность риск-факторов и отобраны итоговые признаки для построения модели.

МФТИ

Пример проверки данных на монотонность



МФТИ

Разработка модулей (транзакционный модуль)

Данные для построение модели были подпадают под NDA.

Принцип действия был схож с принципом модуля на основе финансовых данных.

1. На основе однофакторного анализа был проведен первоначальный отбор факторов и произведено заполнение пропусков.
2. Построена модель на основе CatBoost с метрикой качества Gini на валидационной выборке 67,5%
3. При помощи деревьев решения была произведена WOE-трансформация риск-факторов
4. Затем был произведен многофакторный анализ. Сначала риск-факторы были сгруппированы в смысловые группы. После этого был установлен порог на максимальное количество риск-факторов и перебор всех возможных комбинаций для получения лучшей модели методом логистической регрессии. В итоге получилось 3 модели, из которых мы выбрали одну с наибольшей метрикой в 63,56% Gini. Также была выбрана модель Logistic Regression.
5. По итогу в модуль вошли следующие риск-факторы: Сумма открытых картотек/Транзакционная Выручка за 12 мес. (вес – 40%), Доля выплат на погашение начисленных процентов во всех выплатах по кредитам за 12 мес. (вес – 30%), Доля списаний ТОП-10 крупнейшим поставщикам (вес – 20%) , Расходы на аренду и оплату по договору/Транзакционная Выручка за 12 мес. (вес – 10%).

Разработка модулей (финансовый модуль)

Данные для построение модели были получены из открытого источника Спарк-Интерфакс.

1. На основе однофакторного анализа был проведен первоначальный отбор факторов и произведено заполнение пропусков.
2. Построена модель на основе CatBoost с метрикой качества Gini на валидационной выборке 82,2%
3. При помощи деревьев решения была произведена WOE-трансформация риск-факторов
4. Затем был произведен многофакторный анализ. Сначала риск-факторы были сгруппированы в смысловые группы. После этого был установлен порог на максимальное количество риск-факторов и перебор всех возможных комбинаций для получения лучшей модели методом логистической регрессии (Logistic Regression). В итоге получилось 5 моделей, из которых мы выбрали одну с наибольшей метрикой в 78,99% Gini.
5. По итогу в итоговую модуль вошли следующие риск-факторы: Логарифм Активов (вес – 25%), Денежные средства/Краткосрочные обязательства (вес – 40%), Себестоимость продаж/Кредиторская задолженность (вес – 20%), Динамика Активов (вес – 15%).

Разработка модулей (исковой модуль)

Данные для построение модели были получены из открытого источника Право.ру и Спарк-Интерфакс.

1. На основе однофакторного анализа был проведен первоначальный отбор факторов и произведено заполнение пропусков.
2. Построена модель на основе CatBoost с метрикой качества Gini на валидационной выборке 84,2%
3. При помощи деревьев решения была произведена WOE-трансформация риск-факторов
4. Затем был произведен многофакторный анализ. Сначала риск-факторы были сгруппированы в смысловые группы. После этого был установлен порог на максимальное количество риск-факторов и перебор всех возможных комбинаций для получения лучшей модели методом логистической регрессии. Наиболее точные модели логистической регрессии давали Gini на валидационной выборке менее 70%, по этой причине в качестве итоговой модели была выбрана модель на основе CatBoost.
5. По итогу в модуль вошли следующие риск-факторы: Сумма активных исков в качестве ответчика/Выручка (вес – 50%), Сумма активных исков в качестве ответчика/Капитал (вес – 30%), Сумма активных исков в качестве ответчика/EBIT (вес – 20%).

Разработка интегральной модели

Интегральная модель оценки кредитоспособности основана на комбинации транзакционной модели и моделях окружения, представленных в модульной системе.

Благодаря подобной системе можно не только анализировать текущее финансовое состояние компании,

но и учитывать потенциальные риски, связанные с ее деятельностью в динамике.

Интегральная модель может быть представлена в виде формулы:

$$PD_{integral} = f(PD_{module1} * W_{module1} + PD_{module2} * W_{module2} + \ldots + PD_{moduleN} * W_{moduleN})$$

где

$PD_{integral}$ – вероятность дефолта по интегральной модели;

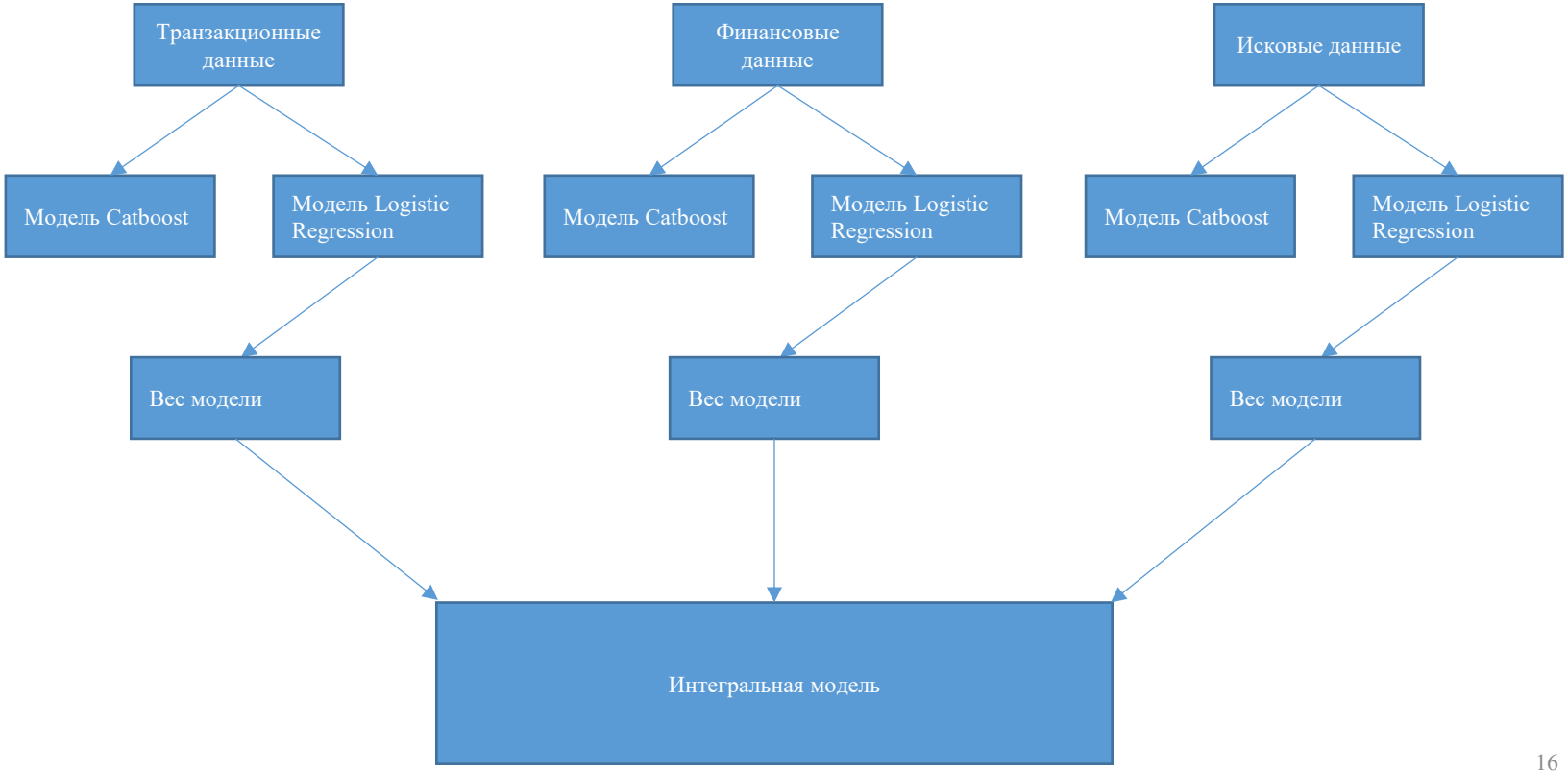
$PD_{moduleN}$ – вероятность дефолта по N-му модулю;

$W_{moduleN}$ – вес N-го модуля в интегральной модели;

f – функция агрегации, включающая в себя логистическую регрессию.

МФТИ

Схема интегральной модели



Интегральные модели

Было построено 3 модели:
1. На основе финансового и транзакционного модулей ;
2. На основе финансового модуля и модуля на исковых данных;
3. На основе финансового и транзакционного модулей и модуля на исковых данных.

Модель	Модули	Коэффициенты модулей	Веса модулей	P-значения	Метрика Gini базовая	Метрика Gini интегральная
Модель № 1	Финансовый, Транзакционный	0.79652499, 0.51310223	60.8%, 39.2%	$9.28e^{-53}$, $3.38e^{-20}$	78.8%	83.4%
Модель № 2	Финансовый, Исковой	0.19324893, 0.98244	16.5%, 83.5%	$7.15e^{-30}$, $2.28e^{-22}$	75.4%	85.3%
Модель № 3	Финансовый, Транзакционный, Исковой	0.14911329, 0.10666997, 0.87119495	13.2%, 9.5%, 77.3%	$3.7e^{-2}$, $9.5e^{-2}$, $3.38e^{-20}$	71.1%	82.4%

МФТИ

Светофорная схема для валидации

Светофорная схема № 1. Пороги эффективность ранжирования всей модели.

gini на уровне модели	Сигнал
Выше 60%	
От 40% до 60%	
<40%	

Светофорная схема № 2. Пороги эффективности ранжирования отдельных факторов.

gini на уровне переменных	Сигнал
Меньше 5%	
От 5% до 10% или наблюдалось снижение ниже 10% на какой-либо из выборок	
Выше 10%	

Светофорная схема № 3. Пороги для коэффициента Gini с использованием доверительных интервалов.

Значение gini	Сигнал
Больше 50%	
От 40% до 50%	
Меньше 40%	

Светофорная схема № 4. Пороги для сравнение наблюдаемого и модельного PD на уровне портфеля.

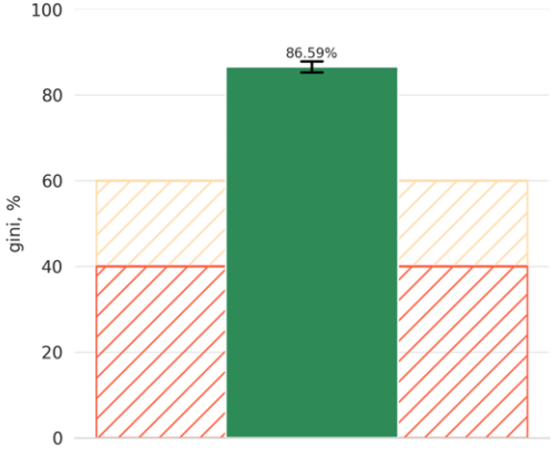
Наблюдаемый средний PD	Сигнал
Между квантилем 5% и квантилем 95%	
Меньше квантиля 5% или больше квантиля 95%	
Меньше квантиля 1% или больше квантиля 99%	

Светофорная схема № 5. Пороги для концентраций наблюдений по рейтингам в динамике.

Значение HHI	Сигнал
Меньше 0.2%	
Больше 0.2% и меньше 0.3%	
Больше 0.3%	

МФТИ

Валидация модели № 3



РАНЖИРОВАНИЕ	
gini, %	86.59
Светофор	Зеленая зона

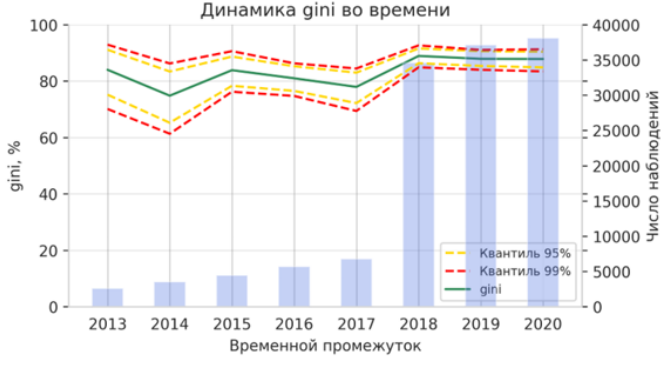
Эффективность ранжирования отдельных факторов.

Фактор	gini, %	Светофор
Финансовый модуль	80.44	Зеленая зона
Транзакционный модуль	64.4	Зеленая зона
Исковой модуль	84.24	Зеленая зона

МФТИ

Валидация модели № 3

Динамика коэффициента Джини за период с 2013 по 2020 год с применением метода бутстрап для определения доверительных интервалов.



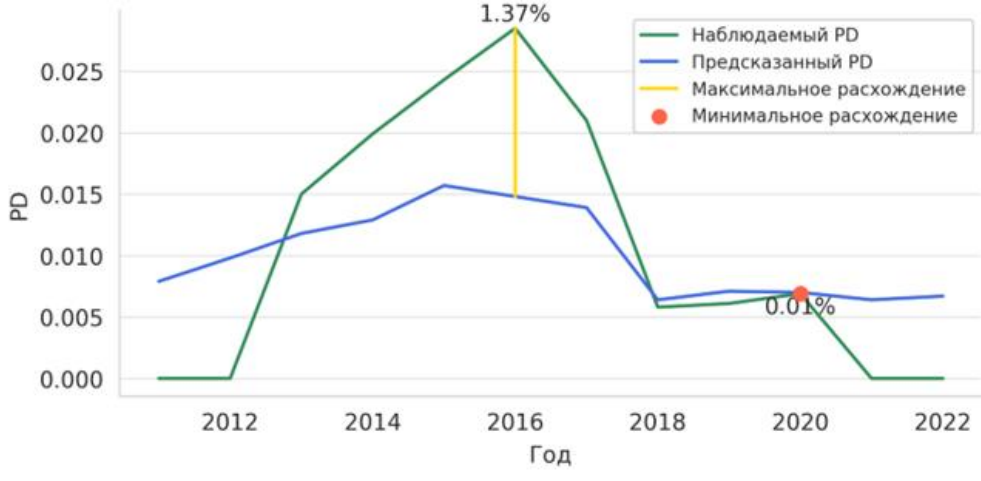
Доверительные интервалы, представленные в таблице, оцененные для каждого года, достаточно узкие, что свидетельствует о стабильности коэффициента Gini и, следовательно, предсказательной силы модели.

Год	Число наблюдений	gini	Светофор	q01	q99	q05	q95
2013	2598	84.02	Зеленая зона	70.11	92.93	75.19	91.14
2014	3522	74.83	Зеленая зона	61.33	86.26	65.22	83.41
2015	4446	83.88	Зеленая зона	76.24	90.62	78.34	88.7
2016	5656	81.0	Зеленая зона	74.73	86.3	76.52	85.27
2017	6750	77.95	Зеленая зона	69.41	84.47	72.2	83.01
2018	34577	88.93	Зеленая зона	84.91	92.68	86.34	91.59
2019	37128	87.91	Зеленая зона	84.02	91.07	85.38	90.6
2020	38096	87.86	Зеленая зона	83.45	91.3	84.86	90.43

МФТИ

Валидация модели № 3

Анализ расхождений между наблюдаемым PD (Probability of default) и предсказанным PD.



МФТИ

Валидация модели № 3

Анализа концентрации наблюдений по рейтингам в динамике с использованием индекса Херфиндаля-Хиршмана (HHI).

Таблица сопровождающая график, предоставляет годовые значения скорректированного HHI.

Год	Значение индекса HHI adj	Результат проверки индекса HHI adj
2011	0.0703	Зеленая зона
2012	0.0688	Зеленая зона
2013	0.0727	Зеленая зона
2014	0.0679	Зеленая зона
2015	0.0618	Зеленая зона
2016	0.0655	Зеленая зона
2017	0.0662	Зеленая зона
2018	0.1175	Зеленая зона
2019	0.1171	Зеленая зона
2020	0.115	Зеленая зона
2021	0.124	Зеленая зона
2022	0.1151	Зеленая зона

МФТИ

Рейтинговая мастер-шкала для калибровки

Вероятность дефолта			Внутренние рейтинги	
Дискретное значение вероятности дефолта (PD)	Нижняя граница	Верхняя граница	Цифровое значение	Рейтинг
0.03%	0.00%	0.04%	20	RU 20
0.05%	0.04%	0.07%	19	RU 19
0.08%	0.07%	0.11%	18	RU 18
0.14%	0.11%	0.20%	17	RU 17
0.25%	0.20%	0.30%	16	RU 16
0.36%	0.30%	0.44%	15	RU 15
0.52%	0.44%	0.63%	14	RU 14
0.74%	0.63%	0.90%	13	RU 13
1.07%	0.90%	1.30%	12	RU 12
1.53%	1.30%	1.87%	11	RU 11
2.20%	1.87%	2.68%	10	RU 10
3.16%	2.68%	3.85%	9	RU 9
4.54%	3.85%	5.54%	8	RU 8
6.53%	5.54%	7.96%	7	RU 7
9.38%	7.96%	11.43%	6	RU 6
13.48%	11.43%	16.43%	5	RU 5
19.37%	16.43%	23.61%	4	RU 4
27.84%	23.61%	33.92%	3	RU 3
39.80%	33.92%	100.00%	2	RU 2
100.00%	100.00%	100.00%	1	RU 1

МФТИ

Калибровка модели (Point-in-Time)

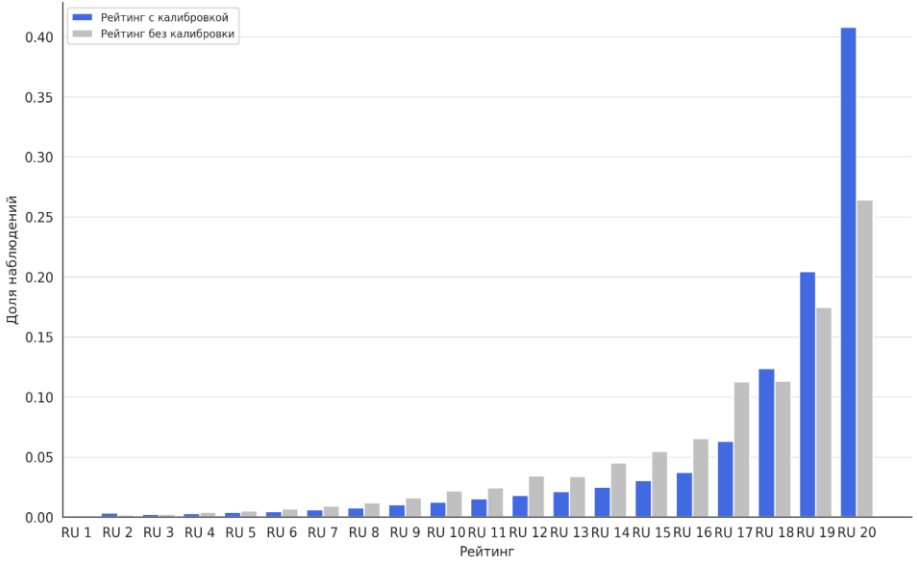
Для калибровки была построена модель логистической регрессии с макроэкономическими факторами и скор-баллом интегральной модели. Gini откалиброванной модели составляет 90.3%

	коэффициент	Вес, %	std err	p-value	Gini, %
ВВП в рыночных ценах, млрд.руб.	-0,7245	32,2	0,036	0	50.61
Платные услуги физический рост, % от предущего года	-0,3144	14	0,037	0	43.01
ТТС скор-балл	1,208	53,8	0,013	0	86.59

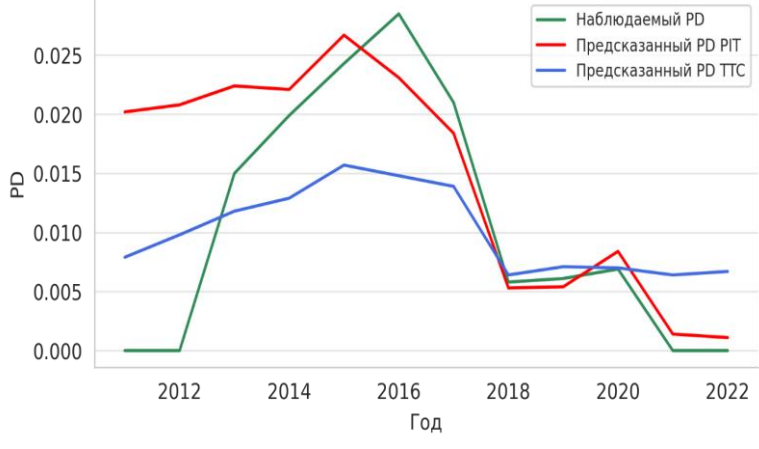
МФТИ

Тестирование калибровки

Распределение рейтингов по интегральной модели TTC (Through-the-Cycle) и по модели с калибровкой PIT (Point-in-Time)



Рейтинг PD TTC более сглаженный и менее зависим от текущей фазы макроэкономического цикла. PD PIT повторяет пиковые значения наблюдаемой частоты дефолтов.



МФТИ

Список ключевых источников

- Моргунов А. В. Методы оценки кредитных рисков инвестиционных проектов: дис. … канд. эконом. наук 08.00.10 / Моргунов Алексей Владимирович
- Фёдорова Е.А., Довженко С.Е., Фёдоров Ф.Ю. Модели прогнозирования банкротства российских предприятий: отраслевые особенности // Проблемы прогнозирования.
- Тотьмянина К. М. Обзор моделей вероятности дефолта // Управление финансовыми рисками. – 2011.
- Попов В. Б., Кадыров Э. Ш. Анализ моделей прогнозирования вероятности банкротства предприятий // Ученые записки Крымского федерального университета имени ВИ Вернадского. Экономика и управление.
- С. Голицын. Транзакционный скоринг как средство идентификации и оценки клиента. [Электронный ресурс]. – FeatureBanking.

МФТИ

Практическая значимость работы и дальнейшие шаги

Разработка модели оценки кредитоспособности на основе машинного обучения представляет интерес для коммерческих банков и возможно иных структур так как на текущий момент для финансового рынка крайне важна точность кредитного скоринга и снижение рисков дефолта юридических лиц.

Дальнейшие шаги – разработка модуля окружения заемщиков на основе транзакционных данных и его учет в интегральной модели, включает в себя следующие шаги:

1) Выделение групп-окружений заемщиков с учетом наличия поручительств между заемщиками или существенным объемам встречных транзакций с учетом имеющейся информации по юридическим группам заемщиков (в периметре консолидации МСФО или РСБУ) из открытых источников.

2) Формирование агрегированных показателей по группам-окружениям заемщиков и разработка отдельного модуля на основании данных показателей по имеющейся статистике.

3) Учет полученного модуля на основе агрегированных показателей по группам-окружениям заемщиков в интегральной модели.

Спасибо
за внимание!

МФТИ