МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

АЛГОРИТМЫ БИННИНГА В МОДЕЛИРОВАНИИ КРЕДИТНОГО РИСКА

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Севостьяновой Ирины Ильиничны

Научный руководитель
доцент, к. ф.-м. н., доцент _____ Н. Ю. Агафонова

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2022

Введение. Целью настоящей работы является исследование особенностей использования алгоритмов биннинга для оптимизации моделирования кредитного риска на базе метода логистической регрессии.

Для достижения заявленной цели работы были поставлены следующие **задачи**:

- изучение теоретических основ кредитного риска;
- исследование ключевых методов управления кредитными рисками;
- углубленное изучение языка программирования R в качестве инструмента практической реализации алгоритмов и построения моделей;
- изучение, реализация и оценка эффективности алгоритмов контролируемого и неконтролируемого биннинга в моделировании кредитного риска;
- построение и сравнительный анализ регрессионных моделей для оценки вероятности дефолта для нескольких независимых наборов данных.

Оценка кредитоспособности заемщиков является ключевой задачей в управлении кредитными рисками. Результаты оценки индивидуальных рисков служат основой для анализа рисков всего кредитного портфеля финансового учреждения.

Актуальность работы обусловлена существенным ростом кредитного риска для банковской сферы вследствие общей рецессии мировой экономики и нестабильности российской экономики, в частности, в условиях ввода ограничительных мер. Примерами наиболее значимых событий, оказавших беспрецедентное влияние на экономическое состояние мирового сообщества, могут служить пандемия коронавируса, объявленная в 2020 году, а в разрезе российской экономики – непрогнозированное повышение ключевой ставки до рекордного значения 20% в 2022 году для сдерживания инфляции, что незамедлительно привело к удорожанию банковских кредитов и другим финансовым последствиям.

В основе моделирования кредитного риска лежат статистические модели. Производительность разрабатываемых моделей является ключевым фактором, так как повышение точности прогнозов обеспечивает значительную экономию средств для финансовых учреждений. Поскольку зависимая переменная в задачах моделирования кредитного риска является чаще всего

бинарной, построение модели в большинстве случаев выполняется на базе логистической регрессии. Подобная закономерность обусловлена относительной легкостью интерпретации результатов моделирования, невысокой чувствительностью к выбросам, а также прямым моделированием вероятностей [1].

Работа имеет следующую **структуру**:

1. первый раздел содержит теоретические основы области управления кредитными рисками и описание методологии построения скоринговых моделей;
2. второй раздел посвящен процедуре биннинга как методу категоризации данных, его основным алгоритмам и соответствующим показателям для оценки информативности разбиения;
3. третий раздел, как и второй, описывает математический аппарат, на базе которого проводились практические исследования, а именно, логистическую регрессию в качестве ключевой статистической модели прогнозирования кредитного риска и метрикам ее оценки;
4. в четвертом разделе представлены этапы и результаты реализации и оценки качества моделей прогнозирования вероятности дефолта с применением алгоритмов биннинга.

Работа прошла апробацию на различных конференциях, в частности, на ежегодной студенческой конференции «Актуальные проблемы математики и механики», которую проводил механико-математический факультет СГУ в апреле 2022 года, в секции «Анализ данных», в X Международной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками», ноябрь 2021 года.

Результаты работы опубликованы в статьях [2], [3].

Основное содержание работы. В **первом** разделе рассматриваются теоретические основы моделирования кредитного риска, приведена задача управления рисками, описаны основные методы оценки и регулирования кредитного риска, а также определена методология скоринговых моделей для оценки кредитоспособности заемщика.

Кредитный риск – возможность возникновения убытков вследствие неоплаты или просроченной оплаты клиентом своих финансовых обязательств. Кредитному риску подвергается как кредитор, так и кредитозаемщик [4].

Оценка кредитоспособности заемщика – один из важнейших этапов процесса кредитования. Это вполне обоснованное действие со стороны финансовых учреждений, поскольку правильность оценки способности заемщика выплачивать кредит и проценты по нему непосредственно влияет на итоговую прибыль кредитной организации [5].

Кредитный скоринг можно опередить как метод начисления потенциальным заемщикам определенного количества баллов на основе информации о его социально-демографическом положении, кредитной истории, параметрах запрашиваемого кредита и принятие решения о выдаче или об отказе в кредите на основе набранного суммарного количества баллов. В основе построения скоринговых карт лежат статистические модели. Качество исходных статистических данных для построения статистической модели определяет ее точность прогнозирования и успех разработки скоринговой системы в целом [6].

В разделе **«Теоретическая база исследования. Алгоритмы биннинга в моделировании кредитного риска»** рассматривается процедура биннинга как метод категоризации непрерывных переменных для повышения качества моделирования кредитоспособности заемщика. Обобщенно приводятся основные алгоритмы контролируемого и неконтролируемого биннинга, а также критерии оценки качества категоризации.

Биннинг (от англ. binning) – форма дискретизации, способ сократить непрерывный диапазон значений в конечное число диапазонов, каждый из которых получает категориальное значение. Интервал наблюдаемых значений разбивается на меньшие группы, каждой из которых присваивается центральное значение, характеризующее эту группу. Все наблюдения, относящиеся к конкретной группе образуют связанный бин.

В качестве теоретического обзора существующих решений и литературы рассматриваются работы, посвященные проблеме моделирования кредитного риска для оценки кредитоспособности заемщика, использующие биннинг в качестве метода анализа и корректировки переменных для построения модели.

Важным показателем является **показатель веса категорий предиктора** WOE (Weight Of Evidence). Веса категорий предиктора помогают най-

ти по переменной «границы чувствительности» к появлению моделируемого события и провести оптимальным образом категоризацию количественных переменных. Показатели WOE для каждой категории рассчитываются по формуле:

$$WOE_i = \ln\left(\frac{d_i^{(1)}}{d_i^{(2)}}\right), \tag{1}$$

где $d_i^{(1)}$ и $d_i^{(2)}$ – относительные частоты «плохих» и «хороших» кредитов соответственно в i-том бине категоризованной переменной, $i = 1, \ldots, k$, k – число категорий переменной [7].

Далее полученные показатели весов категорий анализируются, происходит объединение соседних категорий и расчет показателей WOE повторяется, после чего рассчитывается **показатель информативности** (IV – Information Value) предиктора как взвешенная сумма WOE по всем бинам. Данный показатель отражает степень информативности предиктора для разделения «хороших» и «плохих» кредитов (значений целевой переменной) [8].

$$IV = \sum_{i=1}^{n} \left[WoE_i \left(d_i^{(1)} - d_i^{(2)} \right) \right]. \tag{2}$$

В **третьем** разделе приведено описание метода логистической регрессии в качестве модели классификации, которая используется в эмпирическом исследовании, а также ключевые метрики качества моделей.

Логистическая регрессия – это разновидность множественной регрессии, общее назначение которой состоит в анализе связи между несколькими независимыми переменными и зависимой переменной. Для определения зависимости между переменными, одна из которых категориально зависима, а другие независимы применяется логистическая функция.

Высокая точность получаемых прогнозных моделей обеспечивает частое использование метода в качестве эталонного во всех направлениях банковской деятельности. Отдельно следует отметить простоту интерпретации результатов моделирования как весомое преимущество метода, так как нередко наглядное объяснение принятых решений о выдаче денежных ссуд служит основной в решении споров между финансовыми учреждениями и их клиентами.

Раздел **«Реализация и оценка эффективности моделей прогнозирования вероятности дефолта с применением алгоритмов биннинга»** посвящен описанию этапов исследования эффективности алгоритмов контролируемого и некотролируемого биннинга, а также разработки модели и программного продукта для прогнозирования вероятности дефолта по кредиту.

Формальная постановка задачи работы. Пусть данные представлены в виде $Y, X_1, \ldots, X_n$, где Y (target) – целевая бинарная переменная $\{0,1\}$.

$X_1, X_2, \ldots, X_n$ – предикторы, которые могут быть как непрерывными, так и категориальными переменными (feature).

Требуется оценить вероятность дефолта с помощью метода логистической регрессии с учетом категоризации данных, проведенной по итогам процедуры биннинга.

Для решения поставленной задачи были выделены **два ключевых этапа**:

1. реализация, оптимизация и оценка базовых алгоритмов контролируемого и неконтролируемого биннинга;
2. построение и сравнительный анализ моделей прогнозирования дефолта с помощью метода логистической регрессии с учетом категоризации данных, проведенной по итогам процедуры биннинга.

Эмпирической основой исследования является набор данных ONE-YEAR PD, представленный в [9]. Объем выборки составляет 25906 наблюдений. Было принято решение в качестве целевой переменной использовать фиктивную бинарную переменную, которая принимает значение 1 (далее будет использоваться термин «Good Rate» – кредит был возвращен заемщиком, рейтинг заемщика – высокий) в случае, если у клиента отсутствует задолженность по кредитному счету, клиент не является банкротом и кредитный счет клиента прошел первоначальный срок погашения без положительной остаточной задолженности. В противном случае переменной присваивается значение 0 (т.е. «Bad Rate» – кредит не был возвращен, рейтинг заемщика – низкий).

В рамках настоящей работы исследование выполнялось с учетом пре-

дикторов:

- ежегодный доход (annual_income);
- максимальное количество месяцев с просрочкой платежа за последние 12 месяцев (max_arrears_12m);
- кредитный балл (bureau_score – рассчитывается по специальной модели исходя из данных о кредитной истории и текущем финансовом состоянии заемщика, хранящихся в базе данных кредитного учреждения, характеризует вероятность возврата долга заемщиком);
- средний кредитный балл за 6 месяцев (avg_bureau_score_6m);
- величина текущей просрочки платежа в месяцах (arrears_months);
- доля использования текущего кредитного счета (cc_util).

Построение и оценка эффективности алгоритмов биннинга. Для оценки эффективности методов **неконтролируемого биннинга** были выбраны алгоритм разбиения на интервалы равной длины (equal-width) и алгоритм разбиения на равные по количеству наблюдений интервалы (equal-size). По результатам построения классических и модифицированных алгоритмов (модификация состояла в добавлении этапа аналитической обработки бинов, содержащих только одинаковые значения для целевой переменной), сделан вывод о низкой эффективности алгоритмов вследствие несоответствия базовым критериям качества биннинга.

В разрезе частных заключений, характерных для рассмотренного набора данных, следует отметить:

- достижение более высокого уровня IV при использовании алгоритма разбиения на интервалы равной длины по сравнению с алгоритмом разбиения на равные по количеству наблюдений интервалы;
- отсутствие монотонности WOE для алгоритма equal-size в преимущественной части проведенных экспериментов, свидетельствующее о некачественном биннинге;
- высокая информативность предиктора, содержащего данные о максимальном количестве месяцев с просрочкой платежа за последний год.

В качестве исследуемого метода **контролируемого биннинга** был выбран алгоритм монотонного биннинга. Преимущество данного алгоритма заключается в более эффективной обработке несимметрично распределенных

наборов данных. Для оценки качества биннига алгоритм был выполнен для 75% наблюдений исходного набора данных (обучающая часть), после чего по найденным точкам осуществлялось разбиение оставшихся 25% (тестовая часть). Показатели IV и WOE были рассчитаны как для тестовой части, так и для обучающей.

Оценка производительности алгоритма монотонного биннинга позволяет сделать вывод, что алгоритм монотонного биннинга и метод условных деревьев решений (алгоритм является одним из наиболее актуальных и активно применяется в банковских учреждениях для решения задачи построения скоринговых карт) демонстрируют относительно равную степень эффективности, однако результирующее число бинов может существенно различаться. Однако монотонный биннинг в отличии от метода CTree не гарантирует прохождение 5%-го барьера для результирующих бинов, что обуславливает необходимость дополнительной аналитической обработки полученного разбиения.

Построение и анализ моделей прогнозирования вероятности дефолта. Второй и ключевой этап исследования состоял в построении и сравнительном анализе прогнозных моделей для определения значения зависимой переменной.

Цель проведения экспериментов заключалась в оценке эффективности использования процедуры биннинга как одного из инструментов повышения качества модели. Также особое внимание было уделено проверке гипотезы о результативности категоризации отдельных переменных в противовес практике стандартизации всех количественных переменных, которая активно применяется в литературе.

Качество модели оценивалось на основе показателя точности прогнозных значений (*accuracy*) и матрицы неточностей (confusion matrix). Дополнительно были рассчитаны показатели *precision*и *recall*. Показатели *precision* и *recall* рассчитаны для класса «Bad rate», т.е. для инвертированной матрицы неточностей, где положительным событием считается фиксация прогностической моделью дефолта по кредиту.

Предварительно для исследуемых независимых переменных были отобраны наиболее информативные разбиения, на базе которых уже проводилось

построение регрессионных моделей.

По результатам сравнительного анализа полученных моделей были сделаны следующие выводы:

— использование в модели категоризированных предикторов, являющихся информативными с точки зрения значения показателя IV, обеспечивает повышение точности модели;

— замена исходных значений на веса категорий для всех предикторов не позволяет обеспечить улучшение обобщающей способности модели, даже при условии значимости используемых независимых переменных, данная особенность объясняется наличием предикторов, для которых были выявлены явные признаки некачественного биннига (в данном случае cc_util);

— для большинства моделей характерно преобладание ошибок I рода – когда заемщик фактически возвратил кредит, а модель выявила дефолт. Для рассматриваемой предметной области полученный результат может считаться преимуществом модели, так как отказ в выдаче кредита потенциально ненадежному клиенту является предпочтительным вариантом поведения по сравнению с ростом процента выдаваемых кредитов без относительно высокой гарантии возврата;

— алгоритмы неконтролируемого биннинга не продемонстрировали рост эффективности моделирования, однако разбиение на интервалы равной длины можно считать предпочтительным по сравнению с разбиением на равные по количеству наблюдений интервалы, как по результатам тестирования моделей, так и с точки зрения логики.

После предварительного подтверждения выдвинутой гипотезы было принято решение дополнительно оценить качество построенных моделей с использованием показателя AUC (площади под ROC-кривой), а также провести серию исследований для другого набора данных.

По итогам сравнения значений показателей качества построенных моделей были выведены следующие заключения:

— значение показателя AUC для моделей с использованием неконтролируемого биннинга существенно ниже и явно выделяется из группы построенных моделей, так, например, если для показателя обобщающей

способности разница между наиболее эффективными моделями и моделями equal-width и equal-size составляет ≈ 0.01, то для показателя AUC эта величина составляет уже от 0.1 до 0.2;

— показатель AUC чувствителен к включению в модель совокупности переменных с некачественным разбиением;

— выявленная для показателя точности прогнозных значения тенденция к улучшению качества модели при включении информативных категоризированных предикторов не является характерной для показателя AUC, например, включение в модель переменных со средней информативностью разбиения (annual_income) или «переобученных» переменных (cc_unil) не является фактором для выраженного снижения эффективности;

— наиболее эффективные модели по показателю AUC были получены при использовании алгоритма монотонного биннинга для всех значимых переменных, а также при категоризации только переменной cc_util, разбиение для которой, как уже было отмечено выше, не является информативным, однако модель получила высокую оценку вследствие склонности к более частому совершению ошибок I рода по сравнению с другими моделями в серии экспериментов.

В качестве исследуемой выборки для **второй серии экспериментов** был выбран набор данных «Loan Default Prediction» (прогноз дефолта по кредиту), представленный на платформе Kaggle в рамках открытого конкурса, объявленного компанией Geekbrains [10]. Целевая переменная – бинарная переменная «Credit Default», которая принимает значение 0 в случае, если кредит не был возвращен и зафиксирован дефолт (т.е. «Bad Rate»), и значение 1, если кредит был возвращен (т.е. «Good Rate»). Объем выборки составляет 7500 экземпляров.

Исследование выполнялось с учетом предикторов:

— ежегодный доход (Annual_Income);

— текущая сумма задолженности (Current_Loan_Amount);

— кредитный балл (Credit_Score – рассчитывается по специальной модели исходя из данных о кредитной истории и текущем финансовом состоянии заемщика);

— величина ежемесячного платежа (Monthly_Debt);
— максимальная величина задолженности по кредиту (Maximum_Open_Credit);
— текущий кредитный баланс (Current_Credit_Balance).

Выводы по результатам второй серии экспериментов (для набора данных «Loan Default Prediction»):

— гипотеза о повышении эффективности модели при использовании биннинга только для информативных с точки зрения IV предикторов подтверждается при оценке модели с помощью показателя точности;
— все модели с использованием алгоритмов категоризации переменных (контролируемый биннинг) обеспечили достижение более высоких показателей AUC (средний прирост эффективности составил ≈ 0.1);
— показатель AUC вновь не продемонстрировал корреляцию с использованием в модели категоризации всех переменных или переменных исключительно с наиболее эффективным разбиением;
— отдельно следует отметить, что для рассматриваемого набора данных алгоритм разбиения на интервалы равной длины показал достаточно высокие результаты (для лучшей модели $accuracy = 0.7744$, для модели equal-width $accuracy = 0.7696$), что объясняется равномерностью выборки;
— как и для первой серии экспериментов, для большинства прогностических моделей характерно преобладание ошибок I рода (в среднем около 420 объектов классифицируются как кредиты с высоким риском невозврата, при этом согласно исходному набору данных кредит был выплачен).

Заключение. В ходе работы исследованы особенности использования алгоритмов биннинга для оптимизации моделирования кредитного риска на базе метода логистической регрессии.

Были решены следующие задачи:

— изучение теоретических основ кредитного риска;
— исследование ключевых методов управления кредитными рисками;
— углубленное изучение языка программирования R в качестве инструмента практической реализации алгоритмов и построения моделей;

— изучение, реализация и оценка эффективности алгоритмов контролируемого и неконтролируемого биннинга в моделировании кредитного риска;
— построение и сравнительный анализ регрессионных моделей для оценки вероятности дефолта для нескольких независимых наборов данных.

Результатом исследования служит вывод о повышении качества оценки значения целевой переменной в случае проведения процедуры биннинга для информативных предикторов с точки зрения показателя IV. В разрезе рассмотренных алгоритмов биннинга следует отметить, что алгоритм монотонного биннинга является эмпирическим и требует дальнейшей аналитической обработки, как машинной, так и логической. Алгоритм демонстрирует относительно высокую эффективность по показателям IV, WOE, но не гарантирует удовлетворение главному условию качественного биннинга – прохождение каждым бином 5%-го порога. В свою очередь, алгоритмы неконтролируемого биннинга не обеспечивают достижения оптимального разбиения, однако могут служить базой для тестирования алгоритмов контролируемого биннинга.

Выдвинутая гипотеза о повышении качества прогностической модели в зависимости от включения в модель только информативных предикторов с точки зрения IV является истинной в случае базирования оценки на показателе *accuracy*. С другой стороны, закономерность не подтверждается для достаточно популярного в литературе показателя AUC, который не обладает чувствительностью к точечному включению в модель параметров исключительно с информативным разбиением.

Эффективность применения статистических и эконометрических методов к проблеме моделирования кредитного риска с использованием современных технологий интеллектуального анализа данных позволяет гарантировать своевременное получение результатов для принятия решения о кредитовании. Перспектива применения новых моделей может послужить катализатором роста конкуренции на рынке банковского кредитования, что, с одной стороны, приведет к снижению процентных ставок по кредитам и повышению их привлекательности для граждан, а с другой стороны, позволит коммерческим банкам существенно сократить риск невозврата заемных средств.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Kraus, A.* Recent Methods from Statistics and Machine Learning for Credit Scoring / A. Kraus. — Munchen: Cuvillier, 2014.

2 *Севостьянова, И.И.* Алгоритмы биннинга в моделировании кредитного риска / И.И. Севостьянова // *Математическое и компьютерное моделирование в экономике, страховании и управлении рисками.* — 2021. — № 6. — С. 147–151.

3 *Севостьянова, И.И.* Алгоритмы биннинга в моделировании кредитного риска / И.И. Севостьянова // *Научные исследования студентов Саратовского государственного университета.* — 2021. — С. 7–8.

4 *Порошина, А.М.* Обзор подходов к моделированию кредитного риска на портфельном уровне / А.М. Порошина // *Финансовая аналитика: проблемы и решения.* — 2013. — № 3.

5 *Ханнанова, Е. А.* Теоретические основы оценки кредитоспособности / Е. А. Ханнанова // *Вестник науки и образования.* — 2016.

6 *Сорокин, А. С.* Построение скоринговых карт с использованием модели логистической регрессии / А. С. Сорокин // *Науковедение.* — 2014.

7 *Mironchyk, P.* Monotone optimal binning algorithm for credit risk modeling / P. Mironchyk, V. Tchistiakov. — 2017.

8 *Siddiqi, N.* Credit risk scorecards: developing and implementing intelligent credit scoring / N. Siddiqi. — John Wiley and Sons, Inc., Hoboken, New Jersey, 2006.

9 *Bellini, T.* IFRS 9 and CECL Credit Risk Modelling and Validation: A Practical Guide with Examples Worked in R and SAS / T. Bellini. — Academic Press, 2019. — P. 654.

10 Loan Default Prediction [Электронный ресурс]. — URL: https://www.kaggle.com/competitions/credit-default-prediction-ai-big-data/overview (Дата обращения 20.03.2022). - Загл. с экрана. - Яз. англ.