

**БИННИНГ ПЕРЕМЕННЫХ: КОМПРОМИСС МЕЖДУ ЭФФЕКТИВНОСТЬЮ
МОДЕЛИ И РЕГУЛИРОВАНИЕМ****Владислав Владимирович РОСКОШЕНКО**

аспирант, магистр экономики, экономический факультет,
МГУ им. М.В. Ломоносова, Москва, Российская Федерация
roskoshenkoeco@mail.ru
ORCID: отсутствует
SPIN-код: отсутствует

История статьи:

Получена 05.06.2019
Получена в доработанном
виде 19.06.2019
Одобрена 03.07.2019
Доступна онлайн
30.09.2019

УДК 336.71**JEL:** G21, G28**Аннотация**

Предмет. Дискретизация факторов дефолта кредитного требования. Банковская система России успешно переняла опыт западных розничных банков по моделированию различных аспектов поведения заемщиков — физических лиц. Моделирование кредитного риска остается единственной областью, регулируемой Банком России. Несмотря на наличие предписаний регулятора и Базель II, многие аспекты контролируются экспертным суждением. Нерешенной остается проблема выбора алгоритма дискретизации факторов дефолта кредитного требования.

Цели. Выделить множество валидных в отношении кредитного скоринга алгоритмов дискретизации и выбрать из него оптимальный алгоритм. Также необходимо показать, что дискретизация является неотъемлемой частью построения предиктивной модели в случае использования метода логистической регрессии.

Методология. Использованы: статистический анализ, контент-анализ источников.

Результаты. Показано, что среди алгоритмов, соответствующих требованиям Базель II (минимум 5% наблюдений в бакете, пропущенные значения переменной в отдельном бакете, наличие обоих типов наблюдений в бакете), а также устоявшихся в практике критериев (репрезентативность разбиения, монотонность целевой переменной по бакетам), оптимальным оказался предложенный автором алгоритм (TreeR). В его основе — разбиение непрерывной переменной в результате работы алгоритма построения деревьев решений для бинарной зависимой переменной.

Выводы. Данный алгоритм является принципиально новым решением задачи дискретизации непрерывной переменной. Важной особенностью TreeR выступает реализация в открытом программном обеспечении с опорой на общедоступные библиотеки.

Область применения. Результаты работы могут быть использованы в кредитном скоринге, равно как и в любом статистическом моделировании с использованием логистической регрессии.

Ключевые слова:

кредитный скоринг,
логистическая регрессия,
дискретизация,
предобработка данных,
непрерывная переменная

© Издательский дом ФИНАНСЫ и КРЕДИТ, 2019

Для цитирования: Роскошенко В.В. Биннинг переменных: компромисс между эффективностью модели и регулированием // *Финансы и кредит*. — 2019. — Т. 25, № 9. — С. 2040 — 2053.
<https://doi.org/10.24891/fc.25.9.2040>

Банковская система России успешно переняла опыт западных розничных банков по моделированию различных аспектов поведения заемщиков — физических лиц. Данный опыт по большей части аллоцировался на структурные подразделения банков, которые занимаются либо управлением кредитными рисками потребительского кредитования, либо

маркетингом, либо продажами кредитных продуктов и послепродажным обслуживанием.

В рамках управления взаимоотношениями с клиентами российские банки используют модели, позволяющие выбрать подходящий кредитный продукт для вторичного предложения, оптимальный канал коммуникации с клиентом, удобное время для оповещения о сформированном предложении.

В рамках управления кредитными рисками разрабатываются модели, которые делят клиентов на платежеспособных и потенциально выходящих в дефолт (так называемый аппликативный кредитный скоринг). Другие модели разделяют заемщиков с просроченной задолженностью на выгодных для собственного сбора задолженности заемщиков и тех, чей долг выгоднее продать в коллекторское агентство (скоринг просроченной задолженности). Модели фрод-скоринга разделяют входящий поток клиентов на потенциальных мошенников и добросовестных клиентов.

Моделирование кредитного риска остается единственной областью, регулируемой Банком России, так как именно эффективное управление рисками обеспечивает стабильность банковской системы. Рассмотрим, каким образом осуществляется регулирование в связи с влиянием кредитного риска на капитал банка.

Во-первых, банк при расчете достаточности капитала согласно Инструкции № 180-И Банка России выбирает правильный коэффициент риска в зависимости от типа заемщика и вида кредитных требований¹. Данный коэффициент позволяет взвесить задолженность и рассчитать суммарную величину кредитного риска — знаменатель формулы достаточности капитала. Инструкция Банка России прописывает соответствующие значения коэффициента риска.

Однако Положение № 483-П Банка России² предусматривает подход на основе внутренних рейтингов (ПВР), когда банк самостоятельно вычисляет коэффициенты риска. Для этого банк должен моделировать так называемые компоненты кредитного риска посредством построения моделей *PD* (среднегодовая вероятность дефолта), *EAD* (величина кредитных требований, подверженных кредитному риску на момент

дефолта) и *LGD* (доля невозвратных потерь в случае дефолта).

В России на данный момент только ПАО «Сбербанк» и АО «Райффайзенбанк» получили разрешение мегарегулятора на использование банковских методик управления кредитным риском и моделей количественной оценки для определения его величины с применением подхода на основе внутренних рейтингов (ПВР) в целях расчета нормативов достаточности капитала. Новый подход позволяет существенно экономить собственные средства банка, что и является стимулом для перехода на новую методику.

Ожидается, что Банк России, разрешая некоторым банкам ПВР, предъявляет жесткие требования к соответствующему моделированию. Так, Положение № 483-П регламентирует критерии качества построенной модели, а также особенности жизненного цикла модели, постулируя необходимость регулярной внутренней валидации модели посредством тестов на устойчивость функционирования и измерения качества прогнозирования путем сравнения с фактическими исходами³. Банк России отмечает право банка на использование экспертного суждения в форме профессионального суждения подразделения, однако подчеркивает необходимость в таком случае отражать во внутренних документах порядок сочетания экспертного суждения и результатов моделирования. Кроме того, Положение № 483-П, регламентирует описание методологии моделирования, которое должно включать теоретические предпосылки и допущения, используемые источники данных, а также статистические методы валидации устойчивости модели за пределами периода построения.

Во-вторых, кредитный риск воздействует на капитал банка в ходе определения резервов под обеспечение ссуд. Согласно Положению № 590-П Банка России⁴ в зависимости от

¹ Инструкция Банка России от 28.06.2017 № 180-И «Об обязательных нормативах банков».

² Положение Банка России от 06.08.2015 № 483-П «О порядке расчета величины кредитного риска на основе внутренних рейтингов».

³ Там же.

⁴ Положение Банка России от 28.06.2017 № 590-П «О порядке формирования кредитными организациями резервов на возможные потери по ссудам, ссудной и приравненной к ней задолженности».

финансового положения и качества обслуживания долга банк классифицирует ссуды по категориям качества. Для каждой категории качества кредитного требования Положение определяет минимальный размер резерва, финальное значение которого банк устанавливает по своему усмотрению не ниже данного порога.

Очевидно, компоненты кредитного риска также могут использоваться банком для определения оптимальной величины резерва для каждого портфеля однородных ссуд. Соответственно, банк может использовать свои собственные методики и модели (в первую очередь PD-модели), чтобы оценить финансовое положение заемщика. Тем не менее требований к таким моделям Положение № 590 не содержит.

Можно сделать вывод о том, что на данный момент мегарегулятор российской банковской системы сформулировал требования к моделированию только в рамках управления кредитным риском. Более того, имеющиеся требования справедливы только по отношению к банкам, получившим разрешение на ПБР.

Тем не менее для всех банков действует требование Банка России иметь возможность объяснить причину отказа в ссуде по запросу мегарегулятора. В силу этого основным инструментом моделирования дефолта заемщика остается логистическая регрессия, несмотря на взятый банковской системой курс на использование машинного обучения для моделирования поведения розничных заемщиков, так как именно логистическая регрессия является наиболее интерпретируемым классификатором из используемых. Посредством данного права Банк России в явном виде склоняет банковскую систему к консервативному подходу в отношении выбора методов моделирования.

Проблема сдерживания активного стремления банков перейти на более сложные, но менее интерпретируемые алгоритмы раскрыта в некоторых работах. Д.М. Томджак в своей работе поддерживает требование

интерпретируемости и подчеркивает крайнюю необходимость использования в кредитном скоринге ограниченного круга подходов к моделированию: логистическая регрессия, деревья решений и правила [1].

Ученый Д. Гиган отмечает наличие конфликта философии больших данных, текущей регуляторной системы и действующего законодательства [2]. Иными словами, автор утверждает, что возникшая проблема несоответствия не ограничивается только лишь алгоритмами моделирования дефолтов заемщиков. Дело в том, что в условиях реализации концепции больших данных, когда поток данных, доступных экономическому субъекту, многократно увеличивается, необходимы принципиально новые подходы к анализу, добыче, хранению, очистке, передаче, визуализации и обеспечению безопасности информации. Таким образом, от регуляторных и законодательных институтов требуется внедрение новых технологий, разработка методологий, перестройка философии и культуры экономических субъектов. Автор прогнозирует, что традиционная консервативная политика центральных банков по сдерживанию использования новых подходов и стремление к унификации моделирования не будет удачной.

Исследователь Ю. Чиа отмечает, что ограничения на внедрение неинтерпретируемых алгоритмов регуляторами привел к активному использованию кредитными организациями данных по клиентам из социальных сетей [3]. Таким образом, под угрозой оказались личные данные множества клиентов. Авторы считают разумным убрать ограничения на использование банками «черных ящиков» в пользу защиты личной жизни клиентов.

Тем не менее ведутся работы по созданию возможности интерпретации современных алгоритмов машинного обучения. Так, Р. Флорес-Лопес предлагает новый подход к ансамблированию (CADF) [4]. Данный подход находит баланс между высокой точностью ансамбля и возможностью интерпретации решений классификатора. Авторы отмечают преимущества данного подхода: CADF снижает размерность финальной модели,

превращая ансамбль из «черного ящика» в «белый ящик», структуру которого возможно рассмотреть и дать интерпретацию его решениям.

Российский банковский рынок демонстрирует консервативность в выборе методов моделирования кредитных рисков, отдавая предпочтение логистической регрессии. Тем не менее многие аспекты построения таких моделей по-прежнему жестко не детерминированы, поэтому банковские аналитики руководствуются общепринятыми практиками либо опираются на собственное экспертное суждение.

Другим источником требований к статистическому моделированию дефолта для кредитных требований выступает документ (Базель II) Базельского комитета по банковскому надзору⁵. Стоит отметить, что требования данного комитета являются (в отличие от требований, выдвигаемых Банком России) необходимыми к исполнению для участников банковской системы РФ лишь в случае намерения банка пройти проверку на соответствие требованиям данного Комитета. Соответствие требованиям Базельского комитета по банковскому надзору, в частности, позволяет отечественным банкам получить доступ к более дешевому западному фондированию.

Документ Базель II регламентирует процесс построения модели и ее жизненный цикл. В первую очередь требуется учет всех возможных драйверов кредитного риска: характеристики на уровне розничного заемщика, переменные, описывающие текущие и исторические просрочки клиента, а также особенности кредитного продукта, залога, наличие гарантий и страхования. Так, факторы кредитного риска на уровне розничного заемщика (например, тип занятости) в достаточной степени позволяют оценить чувствительность заемщика к неблагоприятным экономическим условиям или неожиданным событиям. Документ также требует значимого распределения заемщиков и

кредитных требований на уровне пула договоров по различным классам риска⁶.

Особую группу представляют требования к данным, используемым банком при моделировании. В частности, предварительная валидация новой модели должна показывать репрезентативность использованных при построении данных относительно актуального потока заемщиков банка. Банк должен собирать и хранить данные по заемщикам как для обеспечения внутренних процессов риск-менеджмента, так и для составления отчетности по требованию регулятора. Кроме того, банк должен собирать как можно более подробные данные для возможности ретроспективного назначения рисковых классов на случай создания более сложной рейтинговой системы⁷.

Помимо описанных основных направлений в зону внимания документа попадают многие аспекты непосредственно построения логистической регрессии. В частности, модель должна включать в себя объясняющие (вероятность дефолта) переменные, трансформированные из условно непрерывной формы в разделенные на так называемые бакеты⁸. Допустим, что одна из переменных модели — возраст заемщика, тогда неким алгоритмом данная переменная будет преобразована в категориальную с уровнями (бакетами) «до 30 лет», «от 30 до 40 лет», «от 40 до 50 лет», «более 50 лет».

Подобная дискретизация условно непрерывных переменных призвана снизить влияние шума, присутствующего в сырых данных. Источником этого шума выступают либо ошибки ввода данных (ошибки операторов при переносе анкетных данных, некорректная идентификация данных Бюро Кредитных Историй), либо намеренное введение в заблуждение банка клиентом при заполнении анкетных данных. Наиболее яркий пример последнего — завышение располагаемого

⁶ Там же.

⁷ Там же.

⁸ Бакетирование или биннинг — синонимы, означающие разбиение переменной на сегменты. Бакет или бин — это результаты данной процедуры.
URL: https://en.wikipedia.org/wiki/Data_binning

⁵ Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework.
URL: <https://www.bis.org/publ/bcbs107.htm>

дохода при потребительском кредитовании без подтверждения дохода заемщика. В научных работах можно встретить исследования влияния предварительной дискретизации на точность классификации алгоритмов, чувствительных к размерности предикторов. Так, Д. Салем отмечает рост точности наивного байесовского классификатора в случае предварительного биннинга переменных [5].

Биннинг позволяет повысить устойчивость модели за счет объединения в одну группу редких значений непрерывной переменной. Наконец, использование непрерывной переменной с большим числом уникальных значений в логистической регрессии может привести к переобучению, иными словами, модель будет неспособна обобщить зависимости и будет показывать сильное падение качества на тестовой выборке данных. Тем не менее дискретизация, как правило, ведет к потере информации.

При этом Базель II требует, чтобы каждый выделенный бакет включал в себя минимум 5% наблюдений из выборки построения модели. Данное требование должно обеспечить устойчивость модели во времени, так как 5-процентный размер бакета на выборке построения с высокой вероятностью обеспечит значимую долю от входящего потока заемщиков в будущих поколениях выдач.

Также документ предписывает выделять пропущенные значения переменной в отдельный бакет. Со временем это требование было расширено: необходимо выделять разные типы пропусков в данных: «естественные» пропущенные значения (например, отсутствие кредитной истории по клиенту в Бюро кредитных историй (БКИ)), «ошибочные» пропуски, появившиеся в силу ошибок записи, идентификации и хранения информации кредитной организацией.

Последнее требование относительно дискретизации свидетельствует о необходимости выбора таких бакетов, чтобы в каждом были обязательно представлены оба типа заемщиков. Иными словами,

некорректным будет разбиение, при котором будет выделена группа наблюдений только «хороших» или только «плохих» заемщиков.

Помимо предписаний Банка России и Базель II имеются требования специфичные для самой сферы кредитного скоринга, которые условно можно определить как общепринятые практики⁹. Среди таковых можно выделить требование монотонности уровня дефолта по бакетам. Иными словами, необходимо чтобы при движении от одного бакета к другому уровень дефолта либо монотонно возрастал, либо монотонно убывал. Данное требование является следствием предпосылки моделирования, согласно которой «хорошая» переменная-предиктор способна разделять высокорисковый сегмент от низкорискового даже в случае глобальной немонотонности зависимости между переменной-предиктором и предсказываемой переменной (флагом дефолта).

Вторым специфичным требованием является репрезентативность дискретизации. Разбиение переменной на бакеты можно считать репрезентативным в случае, если оно максимизирует корреляцию между переменной-предиктором и предсказываемой переменной (флагом дефолта). Другими словами, необходимо выбирать то разбиение из множества возможных вариантов, при котором утрачивается наименьшее количество информации о зависимости между исходной непрерывной переменной и флагом дефолта заемщика.

Последнее предписание является техническим и подчеркивает необходимость удовлетворения алгоритмом разбиения условий, зафиксированных в документе Базель II (минимальная доля каждого бакета на уровне 5%, выделение в отдельную группу пропущенных значений и отсутствие бакетов, содержащих только один тип заемщиков).

Несмотря на наличие предписаний регуляторов, сформулированных представлений о свойствах дискретизации, специфичных

⁹ Mironchik P., Tchistiakov V. Monotone Optimal Binning Algorithm for Credit Risk Modeling.
URL: <http://doi.org/10.13140/RG.2.2.31885.44003>

кредитному скорингу, нерешенной остается проблема выбора алгоритма дискретизации факторов дефолта для кредитных требований. Этот аспект статистического моделирования компонентов кредитного риска мы рассмотрим в данной работе.

Наиболее подробная и обширная классификация алгоритмов дискретизации представлена в работе С. Гарсии [6]. Также таксономию подходов к дискретизации можно найти в работе С. Котсянтиса [7]. Возможно выделить целый ряд характеристик для классификации алгоритмов дискретизации.

В первую очередь алгоритмы можно соотнести в зависимости от метода, используемого для поиска точек разбиения: методы на основе ошибок (работа Р. Кохави [8]), статистические методы (работа М. Булле [9]) и методы на основе энтропии информации (работа У. Файяд [10]).

Также подходы можно разделить на статические и динамические. Критерием в данном случае выступает независимость и момент взаимодействия с обучающимся алгоритмом (например, логистической регрессией), для которого переменные подвергаются дискретизации. Статические алгоритмы действуют до начала работы обучающихся алгоритмов и являются, таким образом, независимыми от них. Динамические действуют в момент построения модели, располагают частичным объемом данных, и создают дискретизацию переменной, которая зависит от обучающегося алгоритма.

Можно выделить одномерные и многомерные подходы к дискретизации. Многомерные алгоритмы одновременно учитывают наличие всех объясняющих переменных для определения оптимального разбиения каждой из них. Они также способны выделять бакеты для каждой переменной по очереди, учитывая при этом взаимодействие данного разбиения с остальными переменными. Одномерные алгоритмы, напротив, дискретизируют переменные по отдельности.

Также подходы дискретизации можно разделить по критерию использования

информации о принадлежности наблюдений к одному из классов (например, к дефолтникам или благонадежным заемщикам). Большинство имеющихся подходов используют данную информацию. Среди второй группы можно выделить такие подходы, как разбиение на равные интервалы и разбиение с равной частотой наблюдений в бакетах.

Другим критерием может выступить подход к созданию алгоритмом новых бакетов. Один подход, который условно можно обозначить как «сверху-вниз», начинает дискретизацию с разбиения всего диапазона значений переменной на две части и в дальнейшем ищет точки разбиения на каждом из образовавшихся интервалов. Другой подход, который условно можно обозначить как «снизу-вверх», на вход получает большое количество интервалов и постепенно объединяет смежные на основе какого-либо показателя гомогенности этих бакетов.

Алгоритмы дискретизации можно разделить на глобальные и локальные. Критерием в данном случае выступит объем информации, доступный алгоритму. Глобальные алгоритмы используют всю доступную информацию, в то время как локальные — часть информации. От локальных стоит отделять подходы «сверху-вниз», где дискретизация на разных «ветках» иерархии происходит независимо друг от друга, а также все динамические алгоритмы, которые по определению используют неполный набор данных.

Наконец, подходы к разбиению переменных также можно разделить на прямые и иерархические. Первые одновременно разбивают весь диапазон переменной на k частей, требуя указания параметра k . Данная группа включает не только алгоритмы, отрабатывающие в один шаг, но и те, которые выбирают несколько точек разбиения за один шаг. Иерархические алгоритмы, напротив, начинают с довольно простого разбиения и усложняют его с каждым шагом. Данная группа требует указания критерия остановки процесса эволюции дискретизации переменной.

Очевидно, разные алгоритмы будут с разным успехом подходить под конкретные задачи аналитиков, так как во многом качество дискретизации зависит от имеющихся данных. Для сравнения качества дискретизации алгоритмов можно выделить ряд критериев.

В первую очередь алгоритмы отличаются количеством бакетов, которые в среднем они склонны выделять. Для задач кредитного скоринга требуется выделять небольшое количество бакетов, каждый из которых содержит значимую долю входящего потока клиентов — от 5%. Для оценки данного аспекта дискретизации можно использовать индекс Херфиндала—Хиршмана. Низкое значение индекса будет свидетельствовать о равномерности распределения выборки клиентов по бакетам переменной.

Вторым критерием выступает точность. Хороший алгоритм дискретизации способен выделить бакеты таким образом, чтобы прогностическая сила модели не снижалась значительно на тестовой выборке. Для оценки точности возможно обучать модель на одной переменной и вычислять точность прогнозирования обученной модели на тестовой выборке. Другой вариант — оценивать разделяющую силу переменной после ее дискретизации также на тестовой выборке. Для последнего необходимо вычислить так называемый показатель информационной ценности (information value) переменной, который показывает информативность переменной с точки зрения ее способности разделять «хороших» и «плохих» заемщиков. При вычислении информационной ценности переменной используется показатель WoE (weight of evidence), который отражает уровень кредитного риска для каждого бакета переменной.

Кроме того, критерием сравнения алгоритмов дискретизации может выступить машинное время выполнения. Для самообучающихся моделей, используемых в высокодинамичных процессах (например, формирование контекстной рекламы), критичными могут стать значительные временные затраты на дискретизацию переменных-предикторов.

Однако в задачах кредитного скоринга такой критерий не находится на первых позициях.

В литературе встречаются работы по сравнению подходов к дискретизации на основе перечисленных критериев. Помимо таксономии подходов к дискретизации, в работе С. Гарсии представлен сравнительный анализ 30 алгоритмов дискретизации наиболее актуальных и широко используемых в последние годы [6]. В рамках эмпирического анализа были использованы шесть методов моделирования бинарной классификации из пяти различных типов (два алгоритма из семейства деревьев решений (C 4.5, PUBLIC), метод кластеризации KNN, методы генерации правил (DataSqueezer, Ripper), а также вероятностный классификатор Naive Bayes). Можно заметить отсутствие логистической регрессии среди рассмотренных методов моделирования. Результаты использования разных комбинаций алгоритма дискретизации и метода классификации усреднялись на 40 различных наборах данных. В качестве критериев сравнения алгоритмов использовались число бакетов и точность прогнозирования. Авторы отмечают ряд алгоритмов дискретизации, показавших высокий уровень точности для всех рассмотренных типов моделирования: FUSINTER (предложен Д. Зикхедом [11]), ChiMerge (предложен Р. Кербером [12]), CAIM (предложен Л. Курганом [13]), и Modified Chi2 (предложен Ф. Тейем [14]).

В работе П. Мирончука и В. Чистякова можно встретить сравнение подходов к дискретизации в условиях симуляции кредитного скоринга и именно для логистической регрессии как метода моделирования¹⁰. Авторы использовали публичные данные компании FICO об открытых кредитных линиях, обеспеченных залогом в виде недвижимости. Дискретизации подлежала переменная — отношение одобренного лимита к стоимости залога. С одной стороны, подходы были сопоставлены на основе критерия

¹⁰ Mironchuk P., Tchistiakov V. Monotone Optimal Binning Algorithm for Credit Risk Modeling.
URL: <http://doi.org/10.13140/RG.2.2.31885.44003>

равномерности наполнения бакетов. Для этого были рассмотрены количество выделенных бакетов и значение индекса Херфиндаля—Хиршмана. Лучшими алгоритмами здесь стали один из алгоритмов построения деревьев решений (Conditional Inference Trees) (предложен К. Хорником [15]), метод рекурсивной дискретизации на основе принципа снижения энтропии информации (Multi-interval discretization) (предложен У. Файядом [10]), классический метод разбиения на равные по доле наблюдений бакеты, а также предложенный авторами метод монотонного биннинга (Optimal Binning). С другой стороны, подходы были сопоставлены на основе точности. В данном случае среди лучших не оказалось метода Conditional Inference Trees, но добавился метод объединения смежных интервалов на основе статистики Хи-квадрат Р. Кербера (Chi-merge) [12]. С учетом того, что монотонность дискретизации среди лучших из рассмотренных авторами методов достигается только у авторского, метод Optimal Binning рекомендовался к использованию в кредитном скоринге. Ограничением данной работы является отсутствие реализации авторского алгоритма на распространенных языках программирования.

Опираясь на имеющиеся прикладные работы по сравнению методов дискретизации, осуществим собственное исследование. Возьмем основные успешные подходы к дискретизации из рассмотренных работ, алгоритмы распространенные в практике, но не рассмотренные в работах, а также добавим в список собственный подход. Дадим краткое описание этим алгоритмам.

Chi-Merge — алгоритм, явившийся развитием классических методов дискретизации, таких как дискретизация на равные интервалы и разбиение на равные по численности наблюдений интервалы (рассмотрен Р. Кербером [12], Ю. Сангом [16]). Цель подхода — выделить бакеты, в наибольшей степени отличающиеся друг от друга уровнем риска (для случая кредитного скоринга). В начале имеется определенное количество бакетов переменной — от 20 до 100. Далее для

каждой пары смежных интервалов вычисляется статистика Хи-квадрат для тестирования подобия интервалов в паре. Выбрав наибольшее значение статистики Хи-квадрат, алгоритм сравнивает его с пороговым значением. Если статистика Хи-квадрат ниже этого значения, то бакеты объединяются и весь процесс повторяется.

Описанный алгоритм предполагает выбор аналитиком порогового значения, что лишает процесс дискретизации автоматизации. В работе было предложено решение, в соответствии с которым пороговое значение подбиралось автоматически на основе данных. Кроме того, был предложен критерий остановки дискретизации на основе уровня несогласованности в данных. Данный алгоритм получил название Chi2. Тем не менее обновленный алгоритм требовал установки порогового значения теперь уже для уровня несогласованности, а также не учитывал неточности в расчете статистики Хи-квадрат (некорректно вычислялась степень свободы). В работе Ф.Е. Тэй данные моменты были исправлены и обновленный подход получил название Modified Chi2 [14].

Классический подход (Equal width) — разбиение переменной на предустановленное количество интервалов равной ширины. Данный алгоритм редко используется на практике, но зачастую выступает базовым решением для первичного анализа имеющихся данных.

Multi-interval discretization — подход к дискретизации, основанный на эвристике снижения энтропии при разбиении непрерывной переменной (предложен У. Файядом [10]). Схема дискретизации рекурсивная (сверху вниз): изначально весь диапазон значений переменной разделяется на два интервала, впоследствии для каждого из них повторяется такой шаг. Энтропия информации зависит от разнородности объектов в бакете: в более рисковом она высокая, и наоборот.

Алгоритм деревьев условного вывода (Conditional Inference Trees) реализует рекурсивную дискретизацию непрерывной

переменной. Для выбора точки разбиения используется p -значение отличия двух потенциальных бакетов, рассчитанное на основе перемешивания (permutation test) ярлыков принадлежности элементов выборки к тому или иному классу. Такой критерий остановки роста дерева позволяет избежать использования кросс-валидации и ручного усеечения листьев дерева. Данный подход был предложен К. Хорником [15].

САСС — алгоритм, в основе которого лежит расчет коэффициента сопряженности между классами и непрерывной переменной, разбитой на бакеты. Данный подход был предложен Ч.-Дж. Тсай [17]. САСС в классификации подходов к дискретизации располагается в группе статичных, глобальных, прямых и использующих информацию о распределении ярлыков принадлежности элементов выборки к классам.

Ameva — схожий с предыдущим подход, максимизирующий коэффициент сопряженности (по сути, корреляцию) между бакетированной переменной и флагом принадлежности наблюдения к одному из классов («хорошие» или «плохие») на основе статистики Хи-квадрат и минимизирующий количество полученных бакетов. Метод предложен Л. Госалес-Абрилом [18]. Данный алгоритм работает автономно и не требует указания желаемого количества бакетов.

САИМ также максимизирует взаимозависимость класса и бакетов переменной. Метод был разработан Л. Курганом [13].

Дадим описание нескольким подходам, используемым на практике в банковском секторе. Эти алгоритмы в своей основе содержат описанные алгоритмы, доработанные с учетом специфических требований к дискретизации в кредитном скоринге.

Smbinning — алгоритм, в основе которого деревья условного вывода (Conditional Inference Trees). Из особенностей — более доступная возможность установки минимального размера бакета, а также встроенный конвертер сырых данных в дискретизированный вид.

Varbin.monotonic — подход, который выполняет все требования к кредитному скорингу: монотонность уровня целевой переменной по бакетам, минимальный размер бакетов и отдельная обработка пропущенных значений. В основе алгоритма — использование деревьев решений и монотонной регрессии для достижения компромисса между требованиями и максимизацией разделяющей способности переменной.

Logibin — подход, который частично выполняет все требования к кредитному скорингу: минимальный размер бакетов и отдельная обработка пропущенных значений. Монотонности уровня целевой переменной по бакетам также возможно добиться при использовании дополнительных функций, которые принудительно достигают монотонности посредством объединения смежных бакетов. В основе алгоритма — деревья условного вывода (Conditional Inference Trees). Из ограничений подхода можно отметить неоптимальность принуждения к монотонности посредством объединения бакетов после того, как бакеты уже сформированы. Такая последовательность шагов вероятнее приводит к большей потере разделяющей способности переменной. Другое ограничение — неспособность достичь монотонности для всех переменных.

Мы предлагаем собственный алгоритм для дискретизации — TreeR. В его основе — разбиение непрерывной переменной в результате работы алгоритма построения деревьев решений для бинарной зависимой переменной (класс DecisionTreeClassifier из sklearn.tree библиотеки Python). В качестве параметров для DecisionTreeClassifier передается ограничение на минимальный размер бакета в 5%. Для переменной строятся четыре дерева решений разной глубины — от 1 до 4. Пропущенные значения принудительно выделяются в отдельный бакет и не участвуют в построении деревьев. В итоге выбирается разбиение с максимальным значением показателя информационной ценности (IV) среди вариантов, удовлетворяющих условию монотонности уровня целевой переменной по бакетам.

Для сравнения описанных алгоритмов дискретизации были взяты данные о потребительском кредитовании физических лиц одним из крупнейших розничных финансовых институтов на российском рынке. Данные содержат информацию по 150 тысячам выданных ссуд. Среди переменных, описывающих заемщиков, мы выбрали следующие: средняя утилизация кредитных карт или кредитных линий (отношение баланса по необеспеченным кредитам к одобренному лимиту), возраст заемщика, число вхождений клиентом в просрочку 30–59 дней за последние два года, отношение долговой нагрузки (месячные выплаты по имеющимся ссудам, алименты и стоимость проживания) к ежемесячному валовому доходу.

Согласно *табл. 1* пригодными для использования в кредитном скоринге подходами являются Varbin.monotonic и TreeR. Как свидетельствуют результаты эмпирического тестирования (*табл. 2*), TreeR превзошел конкурирующий алгоритм. При сопоставлении TreeR со всеми остальными подходами можно заметить, что Equal Width и Smbinning для некоторых переменных оказались сильнее с точки зрения разделяющей способности разбитой переменной. Тем не менее данные варианты дискретизации не подлежат использованию в кредитном скоринге в силу немонотонности уровня целевой переменной по полученным бакедам.

Таблица 1
Удовлетворение требований кредитного скоринга алгоритмами

Table 1
Performance of credit scoring requirements through algorithms

Алгоритмы	Реализация	Монотонность	Минимальный размер в 5%	Отдельная обработка пропущенных значений
Chi-Merge	R (discretization)	–	–	–
Modified Chi2	R (discretization)	–	–	–
Equal Width	R (discretization)	–	+	+
Multi-interval discretization	R (discretization)	–	–	–
Conditional Inference Trees	R (partykit)	–	+	+
CACC	R (discretization)	–	–	–
Ameva	R (discretization)	–	–	–
CAIM	R (discretization)	–	–	–
Smbinning	R (smbinning)	–	+	+
Varbin.monotonic	R (varbin)	+	+	+
Logibin	R (logibin)	–	+	+
TreeR	Python	+	+	+

Источник: авторская разработка

Source: Authoring

Таблица 2
Результаты эмпирического сопоставления алгоритмов (Gini)

Table 2
The results of the empirical comparison of algorithms (Gini)

Алгоритм	Возраст	Средняя утилизация по необеспеченным кредитам	Число просрочек 30 — 59 дней за последние два года	Отношение долговой нагрузки к месячному валовому доходу
Chi-Merge	26,22	—	35,47	—
Modified Chi2	0	—	21,4	—
Equal Width	26,76	54,84	37,23	3,38
Multi-interval discretization	26,41	55,71	37,2	10,43
Conditional Inference Trees	27,6	—	37,21	—
CACC	21,08	41,76	37,1	10,43
Ameva	21,08	41,76	35,65	9,8
CAIM	21,08	41,76	35,65	7,14
Smbinning	27,35	—	35,47	10,63
Varbin.monotonic	23,47	40,1	35,47	8,25
Logibin	27,22	—	35,47	10,42
TreeR	27,6	54,73	37,62	7,1

Примечание. Прочерки означают ненахождение алгоритмом значимого разбиения переменной.

Источник: авторская разработка

Source: Authoring

Список литературы

1. Tomczak J.M., Zięba M. Classification Restricted Boltzmann Machine for Comprehensible Credit Scoring Model. *Expert Systems with Applications*, 2015, vol. 42, iss. 4, pp. 1789–1796. URL: <https://doi.org/10.1016/j.eswa.2014.10.016>
2. Guégan D., Hassani B. Regulatory Learning: How to Supervise Machine Learning Models? An Application to Credit Scoring. *The Journal of Finance and Data Science*, 2018, vol. 4, iss. 3, pp. 157–171. URL: <https://doi.org/10.1016/j.jfds.2018.04.001>
3. Xia Y., Liu C., Da B., Xie F. A Novel Heterogeneous Ensemble Credit Scoring Model Based on Bstacking Approach. *Expert Systems with Applications*, 2018, vol. 93, pp. 182–199. URL: <https://doi.org/10.1016/j.eswa.2017.10.022>
4. Florez-Lopez R., Ramon-Jeronimo J.M. Enhancing Accuracy and Interpretability of Ensemble Strategies in Credit Risk Assessment. A Correlated-Adjusted Decision Forest Proposal. *Expert Systems with Applications*, 2015, vol. 42, iss. 13, pp. 5737–5753. URL: <https://doi.org/10.1016/j.eswa.2015.02.042>
5. Salem D. Supervised Versus Unsupervised Discretization for Improving Network Intrusion Detection. *International Journal of Computer Science and Information Security (IJCSIS)*, 2016, vol. 14, iss. 10, pp. 583–590.
6. García S., Luengo J., Saéz J.A. et al. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, 2013, vol. 25, no. 4, pp. 734–750. URL: <https://doi.org/10.1109/TKDE.2012.35>
7. Kotsiantis S.B., Kanellopoulos D. Discretization Techniques: A Recent Survey. *GESTS International Transactions on Computer Science and Engineering*, 2006, vol. 32, iss. 1, pp. 47–58.

8. Kohavi R., Sahami M. Error-Based and Entropy-Based Discretization of Continuous Features. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, AAAI Press, 1996, pp. 114–119.
URL: <http://robotics.stanford.edu/users/sahami/papers-dir/kdd96-disc.pdf>
9. Boulle M. Khiops: A Statistical Discretization Method of Continuous Attributes. *Machine Learning*, 2004, vol. 55, iss. 1, pp. 53–69.
URL: <https://doi.org/10.1023/B:MACH.0000019804.29836.05>
10. Fayyad U.M., Irani K.B. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. International Joint Conferences on Artificial Intelligence. AAAI Press, 1993, vol. 2, pp. 1022–1027.
11. Zighed D.A., Rabaséda S., Rakotomalala R. FUSINTER: A Method for Discretization of Continuous Attributes. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 1998, vol. 06, no. 03, pp. 307–326. URL: <https://doi.org/10.1142/S0218488598000264>
12. Kerber R. ChiMerge: Discretization of Numeric Attributes. Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose, California, AAAI Press, 1992, pp. 123–128.
URL: <https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1992-Kerber-ChimErge-AAAI92.pdf>
13. Kurgan L.A., Cios K.J. CAIM Discretization Algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 2004, vol. 16, iss. 2, pp. 145–153.
URL: <https://doi.org/10.1109/TKDE.2004.1269594>
14. Tay F.E.H., Shen L. A Modified Chi2 Algorithm for Discretization. *IEEE Transactions on Knowledge and Data Engineering*, 2002, vol. 14, iss. 3, pp. 666–670.
URL: <https://doi.org/10.1109/TKDE.2002.1000349>
15. Hothorn T., Hornik K., Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 2006, vol. 15, iss. 3, pp. 651–674.
URL: <https://doi.org/10.1198/106186006X133933>
16. Yu Sang, Heng Qi, Keqiu Li et al. An Effective Discretization Method for Disposing High-Dimensional Data. *Information Sciences*, 2014, vol. 270, pp. 73–91.
URL: <https://doi.org/10.1016/j.ins.2014.02.113>
17. Tsai C.J., Lee C. I., Yang W.P. A Discretization Algorithm Based on Class-Attribute Contingency Coefficient. *Information Sciences*, 2008, vol. 178, iss. 3, pp. 714–731.
URL: <https://doi.org/10.1016/j.ins.2007.09.004>
18. Gonzalez-Abril L., Cuberos F.J., Velasco F., Ortega J.A. Ameva: An Autonomous Discretization Algorithm. *Expert Systems with Applications*, 2009, vol. 36, iss. 3, part 1, pp. 5327–5332.
URL: <https://doi.org/10.1016/j.eswa.2008.06.063>

Информация о конфликте интересов

Я, автор данной статьи, со всей ответственностью заявляю о частичном и полном отсутствии фактического или потенциального конфликта интересов с какой бы то ни было третьей стороной, который может возникнуть вследствие публикации данной статьи. Настоящее заявление относится к проведению научной работы, сбору и обработке данных, написанию и подготовке статьи, принятию решения о публикации рукописи.

BINNING OF VARIABLES: A COMPROMISE BETWEEN THE EFFICIENCY OF THE MODEL AND REGULATION**Vladislav V. ROSKOSHENKO**Lomonosov Moscow State University (MSU), Moscow, Russian Federation
roskoshenkoeco@mail.ru
ORCID: not available**Article history:**Received 5 June 2019
Received in revised form
19 June 2019
Accepted 3 July 2019
Available online
30 September 2019**JEL classification:** G21, G28**Keywords:** credit scoring,
logistic regression,
discretization, data
preprocessing, continuous
variable**Abstract****Subject** The research discretizes default factors of claims to loan repayment, as an aspect of statistical modeling in banking.**Objectives** The research identifies multiple valid discretization algorithms for credit scoring and choose the most appropriate one. It also demonstrates that discretization is necessary for building a predictive model given the logistic regression method is applied.**Methods** For purposes of research, I conducted the statistical analysis, content analysis of sources.**Results** As part of statistical analysis, the proposed algorithm (TreeR) was proved to be most appropriate among algorithms that are compliant with Basel II requirements and existing criteria. TreeR splits the continuous variable as a result of the algorithm raising decision trees for a binary dependent variable. The algorithm is a brand new solution to the discretization of the continuous variable. What distinguishes TreeR is that it sits on the open access software and relies upon publicly available libraries.**Conclusions and Relevance** The findings can be used for credit scoring as well as for statistical modeling based on the logistic regression.

© Publishing house FINANCE and CREDIT, 2019

Please cite this article as: Roskoshenko V.V. Binning of Variables: A Compromise Between the Efficiency of the Model and Regulation. *Finance and Credit*, 2019, vol. 25, iss. 9, pp. 2040–2053.
<https://doi.org/10.24891/fc.25.9.2040>**References**

1. Tomczak J.M., Zięba M. Classification Restricted Boltzmann Machine for Comprehensible Credit Scoring Model. *Expert Systems with Applications*, 2015, vol. 42, iss. 4, pp. 1789–1796. URL: <https://doi.org/10.1016/j.eswa.2014.10.016>
2. Guégan D., Hassani B. Regulatory Learning: How to Supervise Machine Learning Models? An Application to Credit Scoring. *The Journal of Finance and Data Science*, 2018, vol. 4, iss. 3, pp. 157–171. URL: <https://doi.org/10.1016/j.jfds.2018.04.001>
3. Xia Y., Liu C., Da B., Xie F. A Novel Heterogeneous Ensemble Credit Scoring Model Based on Bstopping Approach. *Expert Systems with Applications*, 2018, vol. 93, pp. 182–199. URL: <https://doi.org/10.1016/j.eswa.2017.10.022>
4. Florez-Lopez R., Ramon-Jeronimo J.M. Enhancing Accuracy and Interpretability of Ensemble Strategies in Credit Risk Assessment. A Correlated-Adjusted Decision Forest Proposal. *Expert Systems with Applications*, 2015, vol. 42, iss. 13, pp. 5737–5753. URL: <https://doi.org/10.1016/j.eswa.2015.02.042>
5. Salem D. Supervised Versus Unsupervised Discretization for Improving Network Intrusion Detection. *International Journal of Computer Science and Information Security (IJCSIS)*, 2016, vol. 14, iss. 10, pp. 583–590.

6. García S., Luengo J., Saéz J.A. et al. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, 2013, vol. 25, no. 4, pp. 734–750. URL: <https://doi.org/10.1109/TKDE.2012.35>
7. Kotsiantis S.B., Kanellopoulos D. Discretization Techniques: A Recent Survey. *GESTS International Transactions on Computer Science and Engineering*, 2006, vol. 32, iss. 1, pp. 47–58.
8. Kohavi R., Sahami M. Error-Based and Entropy-Based Discretization of Continuous Features. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, AAAI Press, 1996, pp. 114–119. URL: <http://robotics.stanford.edu/users/sahami/papers-dir/kdd96-disc.pdf>
9. Boulle M. Khiops: A Statistical Discretization Method of Continuous Attributes. *Machine Learning*, 2004, vol. 55, iss. 1, pp. 53–69. URL: <https://doi.org/10.1023/B:MACH.0000019804.29836.05>
10. Fayyad U.M., Irani K.B. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. International Joint Conferences on Artificial Intelligence. AAAI Press, 1993, vol. 2, pp. 1022–1027.
11. Zighed D.A., Rabaséda S., Rakotomalala R. FUSINTER: A Method for Discretization of Continuous Attributes. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 1998, vol. 6, no. 3, pp. 307–326. URL: <https://doi.org/10.1142/S0218488598000264>
12. Kerber R. ChiMerge: Discretization of Numeric Attributes. Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose, California, AAAI Press, 1992, pp. 123–128. URL: <https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1992-Kerber-ChimErge-AAAI92.pdf>
13. Kurgan L.A., Cios K.J. CAIM Discretization Algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 2004, vol. 16, iss. 2, pp. 145–153. URL: <https://doi.org/10.1109/TKDE.2004.1269594>
14. Tay F.E.H., Shen L. A Modified Chi2 Algorithm for Discretization. *IEEE Transactions on Knowledge and Data Engineering*, 2002, vol. 14, iss. 3, pp. 666–670. URL: <https://doi.org/10.1109/TKDE.2002.1000349>
15. Hothorn T., Hornik K., Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 2006, vol. 15, iss. 3, pp. 651–674. URL: <https://doi.org/10.1198/106186006X133933>
16. Yu Sang, Heng Qi, Keqiu Li et al. An Effective Discretization Method for Disposing High-Dimensional Data. *Information Sciences*, 2014, vol. 270, pp. 73–91. URL: <https://doi.org/10.1016/j.ins.2014.02.113>
17. Tsai C.J., Lee C.I., Yang W.P. A Discretization Algorithm Based on Class-Attribute Contingency Coefficient. *Information Sciences*, 2008, vol. 178, iss. 3, pp. 714–731. URL: <https://doi.org/10.1016/j.ins.2007.09.004>
18. Gonzalez-Abril L., Cuberos F.J., Velasco F., Ortega J.A. Ameva: An Autonomous Discretization Algorithm. *Expert Systems with Applications*, 2009, vol. 36, iss. 3, part 1, pp. 5327–5332. URL: <https://doi.org/10.1016/j.eswa.2008.06.063>

Conflict-of-interest notification

I, the author of this article, bindingly and explicitly declare of the partial and total lack of actual or potential conflict of interest with any other third party whatsoever, which may arise as a result of the publication of this article. This statement relates to the study, data collection and interpretation, writing and preparation of the article, and the decision to submit the manuscript for publication.