pISSN 2071-4688
eISSN 2311-8709

Банковская деятельность

СЕГМЕНТАЦИЯ КЛИЕНТОВ РОЗНИЧНОГО БАНКА ПРИ МОДЕЛИРОВАНИИ ДЕФОЛТА КРЕДИТНОГО ТРЕБОВАНИЯ

Елена Владимировна ПАВЛОВА [a,•],
Владислав Владимирович РОСКОШЕНКО [b]

[a] кандидат экономических наук, доцент, экономический факультет,
МГУ им. М.В. Ломоносова,
Москва, Российская Федерация
lena.pavlova@gmail.com
ORCID: отсутствует
SPIN-код: отсутствует

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

• Ответственный автор

История статьи:
Рег. № 581/2020
Получена 15.10.2020
Получена в доработанном виде 29.10.2020
Одобрена 12.11.2020
Доступна онлайн 27.11.2020

УДК 336.71
JEL: G21

Ключевые слова:
кредитный скоринг, логистическая регрессия, эвристическая сегментация, ROC-кривые

Аннотация

Предмет. Сегментация клиентов розничного банка, способная повысить эффективность аппликативной скоринговой системы. Проблема выбора оптимального эвристического метода сегментации для задачи скоринга.

Цели. Определить оптимальный эвристический метод сегментации.

Методология. Использованы статистический метод исследования, анализ, контент-анализ источников.

Результаты. Сопоставление 33 эвристических методов сегментации клиентов розничного банка показало, что по величине метрики эффективности классификатора (AUROC) оптимальным оказался вариант сегментации по величине выданной ссуды, предложенный авторами. Метод заключается в дискретизации переменной «выданная величина ссуды» методом TreeR. Стоит отметить, что данное разбиение производилось отдельно в рамках каждого кредитного продукта, чтобы было выполнено регуляторное требование о сегментации портфелей ссуд. Для проведения исследования использованы данные по кредитам физическим лицам за 2017 г. (150 тыс. ссуд наличными и 29 тыс. ссуд кредитными картами) из портфеля банка топ-15 по состоянию на 1 октября 2018 г. одной из стран ЕАЭС.

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

Для цитирования: Павлова Е.В., Роскошенко В.В. Сегментация клиентов розничного банка при моделировании дефолта кредитного требования // Финансы и кредит. — 2020. — Т. 26, № 11. — С. 2594 — 2616.
https://doi.org/10.24891/fc.26.11.2594

Введение

При моделировании дефолта кредитного требования розничный банк сталкивается с проблемой сегментации заемщиков. Необходимо ли строить отдельные модели для каждого сегмента кредитных требований или достаточно одной общей модели? Каким образом выделять сегменты? На данные вопросы каждый банк находит собственный ответ.

В аппликативном кредитном скоринге сегментация клиентов заключается в разделении общего множества заявок на ссуды на подмножества, внутри которых имеет место уникальный набор факторов кредитного риска, характерный для конкретного профиля клиента банка. Для примера можно рассмотреть сегментацию по критерию запрашиваемого кредитного продукта. Очевидно, что множество ссуд в форме кредитных карт и в форме наличности имеют разный ожидаемый уровень дефолтности. Мы справедливо будем ожидать меньшую величину кредитного риска по кредитным картам, по которым банк имеет возможность уменьшения лимита счета в случае изменения кредитоспособности клиента.

В силу сегментации совокупное качество набора отдельных скоринговых моделей, построенных на специфических сегментах, может значимо превышать качество одной общей модели. В то же время такие факторы, как ошибочное выделение сегментов, небольшие объемы выборок и несбалансированность классов для отдельных сегментов, переобучение моделей на узких сегментах клиентов, могут привести к обратному результату.

В литературе можно встретить различные методы сегментации клиентов розничного банка в целях аппликативного скоринга. Н. Сиддики выделяет (*рис. 1*) статистические методы сегментирования и методы, основанные на эвристиках — упрощенных подходах, которые сформировались в индустрии на основе опыта и соображений операционной деятельности [1]. Многие из эвристических методов позволяют создавать потенциальные предикторы дефолта, специфичные для сегмента. Например, если выделить клиентов с кредитными картами, то для них возможно вычислить так называемую утилизацию лимита кредитной карты, которая часто демонстрирует высокую корреляцию с кредитоспособностью клиента.

В отличие от эвристических методов статистическая сегментация предполагает выделение групп клиентов на основе информации, заключенной непосредственно в данных о клиенте. Для этого применяются различные методы анализа данных и машинного обучения. Среди статистических методов Дж. Аурифелли разделяет описательные и регрессивные (или предиктивные) [2]. Описательные методы соотносят определенный уровень кредитного риска с клиентским профилем благодаря выделению групп клиентов, гомогенных с точки зрения клиентских характеристик. Наиболее известные описательные методы: LVQ (Learning Vector Quantization) и описательная кластеризация.

Регрессивные методы пытаются найти зависимости между целевой переменной и клиентскими характеристиками. Наиболее известные регрессивные методы: кластеризующие регрессии (Clusterwise Regression), а также одновременные и двушаговые деревья решений.

Ученый Н. Сиддики выделяет семейство эвристических методов сегментации клиентов, похожих по какой-либо социально-демографической характеристике (регион проживания (город/село), благополучие района), возраст, длина кредитной истории в БКИ или банке). Другое семейство эвристических методов основывается на типе кредитного продукта (класс кредитной карты, срок ипотеки, тип страховки, наличие залога, размер ссуды). Канал привлечения клиента также формирует отдельную группу методов сегментации клиентов (точка продажи, отделение банка, Интернет, дилеры, брокеры). Наличие данных также дает возможность выделить пул ссуд с похожим уровнем кредитного риска (объем кредитования клиента в прошлом, наличие или отсутствие негатива в БКИ, револьверные/транзактные клиенты для револьверных продуктов). Н. Сиддики выделяет семейство методов, основанных на типизации клиента (повторный/новый клиент, впервые взявший ипотеку или повторный опыт, профессиональные категории (инженеры, доктора и др.)). Последняя группа методов основана на продуктовой корзине клиента (какие кредитные продукты имеются на руках клиента, когда он обращается в банк).

Исследователь Л. Томас выделяет группу эвристических методов, когда сегментация клиентов совершается для облегчения бизнес-процессов банка [3]. Примером может служить разделение заявок по критерию кредитного продукта. Для каждого кредитного продукта, как правило, созданы разные кредитные стратегии, и заявки проходят разные этапы в кредитном конвейере, и наверняка будут иметь различные пороги отсечения на всех этапах, отдельная их обработка упрощает бизнес-процесс банка. К данной группе автор относит и выделение сегментов клиентов, для которых действующие скоринговые карты работают неэффективно.

Также Л. Томас предлагает группу методов на основе использования предиктора, который должен иметь, по мнению экспертов, сильное взаимодействие с остальными переменными (методы «ключевого предиктора»). Изначально данный метод был эвристическим, однако со временем переместился в группу статистических, поскольку именно на основе различных статистических методов производится оценка степени взаимодействия.

В современной российской практике в той или иной форме имеют место все подходы, описанные в литературе (*рис. 1*). Наиболее распространенный подход состоит в выделении кредитных продуктов, и уже в их рамках — отнесение клиентов к той или иной группе по критерию наличия кредитной истории у клиента в данном банке («новый» или «повторный» клиент для банка), а также по наличию кредитной истории в БКИ. Получаемые сегменты схематично можно отобразить следующим образом (*рис. 2*). Данный подход учитывает риск

финансового инструмента, который определяется либо спецификой самого инструмента, либо особенностью обеспечения. К последнему относится учет покрытия одним объектом обеспечения нескольких кредитных требований, отношения суммы кредита к стоимости обеспечения, наличия сезонности, наличия поручительств. Также данный подход учитывает кредитный риск заемщика, поскольку наличие кредитной истории в банке или в БКИ определяет величину доступных данных по заемщику, а, значит, и точность оценки его кредитного риска.

Автор Б. Бояринцев и др. отмечают, что частой отечественной практикой стала сегментация клиентов розничного банка по критерию величины дохода (mass retail, mass affluent, HNWI, UHNWI), а также по возрасту согласно теории поколений Н. Хоува, В. Штрауса (беби-бумеры, поколение X, миллениалы, поколение Z) [4].

Важным источником эвристических методов сегментации, по мнению Н. Сиддики, выступают потенциальные направления развития банка [1]. Для отечественных розничных банков Н. Галкина на фоне старения населения России выделяет такие перспективные сегменты клиентов, как молодежь, пенсионеры и жители небольших городов [5].

Ученые С. Карпова и др. выделяют ряд традиционных и новых подходов к сегментации клиентов банка, которые используются как в отечественной банковской системе, так и за рубежом [6]. Авторы указывают сегментацию по уровню дохода (малообеспеченные клиенты, средний класс, клиенты с высоким доходом). Другой подход — сочетание возраста и семейного положения клиента (молодежь (16—25 лет), молодые семьи, работающие лица среднего возраста (25—57 лет), семьи «со стажем», работающие лица предпенсионного возраста (50—62 лет), пенсионеры (57 и старше) — по состоянию на 2020 г.). Также авторы приводят модель сегментации Королевского банка Канады (RBC Royal Bank) на основе исследований жизненного цикла клиента: «молодежь» (до 18 лет), «начало работы» (18—35 лет), «строители» (35—50 лет), «аккумуляторы» (50—60 лет), «хранители» (60 лет и старше). С. Карпова и др. отмечают широкое использование сегментации на основе теории поколений Н. Хоува и В. Штрауса.

Проблема предварительной сегментации для моделирования розничных дефолтов кредитного требования разрешается регулирующим органом (Банком России) в требованиях Положения от 06.08.2015 № 483 «О порядке расчета величины кредитного риска на основе внутренних рейтингов». Однако стоит отметить, что данное положение необходимо к исполнению только кредитным организациям, претендующим на использование ПВР (подхода количественной оценки рисков на основе внутренних рейтингов). В данном Положении регулятор отмечает необходимость разработки единой иерархической системы классов, подклассов и сегментов кредитных требований, а также соотнесение сегментов с применяемыми моделями количественной оценки риска (формирование карты моделей). Важным понятием данного документа является

рейтинговая шкала, которая представляет собой набор разрядов (с сортировкой по возрастанию либо убыванию кредитного риска). Каждый разряд сформирован портфелем однородных (по кредитному риску) кредитных требований. В описании устройства рейтинговой системы (шкалы)[1] можно видеть позицию Банка России относительно сегментирования заемщиков: «при распределении кредитных требований в портфели однородных кредитных требований (разряды рейтинговой шкалы) учитываются факторы, определяющие как риск заемщика (например, тип заемщика, его демографические характеристики), так и риск финансового инструмента (например, характеристики продукта и (или) обеспечения…)». С учетом того, что отнесение кредитных требований к разным уровням кредитного риска происходит на основе PD-моделей, которые и отражают риск заемщика, можно понять, что регулятор указывает на необходимость только лишь продуктовой сегментации кредитных требований. Более того, предвидя сложности банков с выделением незначимых (по численности) для их профиля продуктовых портфелей, Банк России не дает жесткой группировки по возможным финансовым инструментам.

В документе Базельского комитета по банковскому надзору Basel II сегмент определялся как «пул ссуд с однородным кредитным риском». Таким образом, данная версия документа не давала конкретных предписаний для сегментации. Если обратиться к последним соглашения Базельского комитета по банковскому надзору Basel IV, а именно к предложениям в новой редакции IRB, которые после согласования с банковским сообществом будут приняты в 2022 г., можно увидеть разделение розничного кредитования на три сегмента: ипотечные займы, кредитные карты и остальные кредиты. Тем не менее новые предложения могут быть отредактированы после согласований, и не обязательно будут отражены в отечественных регламентах.

Наша работа представит новый эвристический метод сегментации клиентов розничного банка. Для обоснования новизны метода, помимо классификации данных методов, которая очерчивает области потенциальных критериев сегментации, необходимо рассмотреть исследования, в которых предложены конкретные эвристические методы сегментации клиентов розничного банка. Последнее предполагает наличие обоснования допустимости использования методов, а также демонстрацию их превалирующей эффективности для какой-либо задачи. При этом учитывая, что объектом сегментации выступают клиенты розничного банка, мы не ограничиваем тематику работ проблемами скоринга. Таким образом, будут рассмотрены все работы по эвристической сегментации клиентов розничного банка, в том числе маркетинговые. Основные исследования и их выводы представлены в *табл. 1*.

В данной статье будут сопоставлены представленные в литературе эвристические методы сегментации клиентов розничного банка, методы, широко используемые в практике, а также предложенные авторами. Сопоставление будет

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

осуществляться по критерию роста эффективности кредитного скоринга. Стоит отметить, что в литературе нет единства мнений относительно того, что построение отдельных скоринговых карт для сегментов (набор скоринговых карт) дает прирост эффективности скоринговых систем розничного банка (*табл. 2*).

В следующем разделе будут рассмотрены использованные данные и методология исследования, а в последнем разделе — результаты исследования.

Данные и методология исследования

Для проведения исследования использованы данные по кредитам физическим лицам за 2017 г. (150 тыс. ссуд наличными и 29 тыс. ссуд кредитными картами) из портфеля банка топ-15 по состоянию на 1 октября 2018 г. одной из стран ЕАЭС.

Данные содержат информацию из трех источников: анкетные данные, данные Бюро кредитных историй (БКИ), а также данные о предыдущей кредитной истории в данном банке (КИ). Анкетные данные содержат информацию, которую все клиенты представляют в процессе подачи заявки на ссуду. Эти БКИ представляют данные обо всех предоставленных в бюро фактах выдачи кредитов другими финансовыми организациями, а также поданных заявок на ссуды. Кроме того, такие БКИ содержат помесячные балансы обслуживания взятых ссуд. Указанные КИ представляют информацию о предыдущих заявках на ссуды от клиентов в данном банке, необязательно выданные. В силу этого, помимо выданных ссуд, возможно определить, в каких ссудах банк отказал клиенту в прошлом, от каких клиент отказался сам, а также какими клиент не воспользовался. По выданным кредитам имеются помесячные балансы обслуживания на каждый отчетный период. Также представлена информация по обслуживанию выставленных платежей: даты очередных взносов, фактические даты платежей, размер внесенных средств.

При расчете переменных-предикторов дефолта кредитного требования использовалось два разделяемых варианта «пустых» значений (в соответствии со стандартами Базель II): отсутствие значений в данных источника, отсутствие данных по клиенту в источнике. Например, «количество закрытых ссуд наличными в БКИ» может не иметь значений из-за того, что клиент не брал такие ссуды, но имел другие, а может не иметь значений из-за отсутствия кредитной истории клиента в БКИ. Разные виды «пустых» значений отражают различные группы клиентов, которые, как правило, значимо различаются по уровню кредитного риска.

Выделение сегментов для числовых переменных, например «размер выданной ссуды», базируется на дискретизации методом TreeR, ранее предложенным В. Роскошенко[2]. Для получения выборок достаточного размера для

[2] *Роскошенко В.В.* Биннинг переменных: компромисс между эффективностью модели и регулированием // Финансы и кредит. 2019. Т. 25. № 9. С. 2040—2053.

моделирования дефолта кредитного требования, метод TreeR использовался с параметром минимального листа дерева решения в 20% для 150 тыс. ссуд наличными, а также 35% для 29 тыс. ссуд кредитными картами. Более высокая доля минимального бакета для кредитных карт объясняется меньшим размером исходной выборки (всего 29 тыс.). С учетом выделения 20% наблюдений при формировании отложенной выборки, а также распределения оставшихся данных между выборкой построения и тестирования в соотношении 80%/20% размер выборки построения составлял 64% от каждого выделенного сегмента. Соответственно, для ссуд наличными выборка построения не опускалась ниже 19,2 тыс. наблюдений, а для кредитных карт — 6,5 тыс. Исключение составляют сегменты, сформированные «пустыми» значениями числовых переменных, которые выделяются в отдельный бакет методом TreeR, а также сегменты, полученные ручным разбиением, например, по наличию источников данных или на основе теории поколений Хоува—Штрауса. Тем не менее выборка построения таких сегментов не опускалась ниже 4 тыс. наблюдений, а при условии близости уровней дефолтности с соседними сегментами такое разбиение корректировалось посредством объединения.

Производилось моделирование дефолта кредитного обязательства. Базовый вариант — набор скоринговых карт, построенных на данных кредитных продуктов. Такой подход позволяет выполнить регуляторное требование по разделению кредитных продуктов. Остальные варианты — построение моделей в разрезе кредитных продуктов и сегментов, полученных согласно конкретному эвристическому методу. В качестве классификатора использовалась логистическая регрессия. Схема обучения алгоритма классификации представлена на *рис. 3*.

Все заявленные на *рис. 3* отсечения предикторов дефолта ссуды зафиксированы, использовались с одними значениями для всех построений. Процесс построения модели содержит эвристики, широко используемые на практике, и является автоматическим (не требует вмешательства человека). Все вместе должно обеспечивать объективность сопоставления различных методов сегментации.

В практике анализа данных для построения моделей исходные наблюдения разбиваются на выборки обучения (train), тестирования (test) и отложенную выборку (oot). При этом последняя никак не используется при моделировании помимо оценки того, насколько финальная модель «промахивается» на неизвестных для нее данных. Такой подход гарантирует, что модель не будет переобучена под конкретные данные, и на новом потоке клиентов мы увидим примерно те же результаты эффективности. Соответственно, принято выбирать конкретный алгоритм классификации (логистическая регрессия, случайный лес, машина опорных векторов и др.), параметры модели, а также конкретный набор предикторов для финальной модели по оптимуму метрики эффективности на тестовой выборке. В нашем случае, когда мы перейдем к сопоставлению уже построенных моделей, к которым не будут уже применяться какие-либо

URL: https://doi.org/10.24891/fc.25.9.2040

дополнительные настройки параметров, справедливо это сопоставление произвести именно на отложенной выборке. Отложенная выборка здесь сыграет роль новых поколений клиентов. Именно такой вариант, например, используется при определении победителей в соревнованиях на известной платформе Kaggle .

Стоит отметить, что имеют место два подхода к выбору модели-чемпиона. Во-первых, можно выбирать наиболее устойчивую модель. Для этого необходимо обращать внимание на показатель отсутствия пере- или недообучения классификатора — величину падения метрики эффективности модели от выборки построения к отложенной выборке (почему именно к отложенной, а не тестовой — мы обсудили). Незначительное падение будет свидетельствовать о достаточной обобщающей способности модели. Такая модель будет с близким к заявленному (на построении) уровнем точности классифицировать объекты в будущем. Во-вторых, можно оптимизировать эффективность модели с помощью наиболее сильной модели на отложенной выборке. Мы остановимся на втором варианте.

В литературе встречаются различные подходы к выбору метрики эффективности моделей для оценки эффекта сегментации и выбора оптимальной сегментации. Если выбранная метрика набора скоринговых моделей оказывается выше метрики базовой модели, то можно говорить о положительном эффекте сегментации, то есть о приросте эффективности скоринговой системы.

Учкеный К. Бижак и др. отмечают, что модели в наборе скоринговых карт сопоставимы между собой с точки зрения масштаба прогнозов [7]. Это позволяет объединить сегменты для расчета единой статистики AUROC, характеризующей разделяющую способность (способность ранжировать клиентов по кредитоспособности) набора скоринговых моделей. Авторы замечают, что эффект сегментации состоит из двух компонентов: разделяющая способность отдельных логистических регрессий, построенных на сегментах, а также разделяющая способность самих деревьев решений (какие клиенты относятся к какому сегменту). Для оценки первого компонента К. Бижак и др. предлагают оценивать среднее значение AUROC, взвешенное по величине клиентов в данном сегменте. Для оценки второго компонента авторы вычитают из AUROC набора скоринговых моделей первый компонент. В нашей статье мы воспользуемся подходом авторов по вычислению единой статистики AUROC. Однако подход по выделению компонентов эффекта сегментации считаем спорным, в силу того что оценивание AUROC набора скоринговых моделей через средневзвешенное значение видится довольно неточным методом.

Автор Ю. Ким и др. использовал функцию правдоподобия для оценки эффективности построенных моделей [8]. Соответственно, значение функции правдоподобия являлось критерием для выбора оптимальной сегментации.

Как упоминалось, в нашей статье метрикой выступит площадь под ROC-кривой (AUROC). Это одна из основных метрик для случая бинарной классификации

[9]. Мы последуем подходу К. Бижак и др. при оценке разделяющей способности набора скоринговых моделей, построенных на сегментах. Иными словами, учитывая, что модели в наборе скоринговых карт сопоставимы между собой с точки зрения масштаба прогнозов, мы объединим сегменты для расчета единой статистики AUROC, характеризующей разделяющую способность набора скоринговых моделей.

Очевидно, для сопоставления двух моделей на отложенной выборке нас не устроит сравнение двух чисел AUROC. Необходимо учесть особенности распределения данных величин. М. Пепе и др. рассмотрели возможные параметрические и непараметрические (независящие от распределения случайных величин) методы сопоставления ROC-кривых [10]. К. Робин и др. оформили наиболее популярные из данных методов в виде пакета для двух языков программирования [11]. Один из непараметрических методов был предложен И. Делонг и др. [12]. К. Сан предложил быструю реализацию данного метода [13]. Другой классический вариант — бутсрэп (bootstrap), в ходе которого производится многократное воспроизведение новых выборок на основе исходной случайным отбором с возвращением. Оценка метрики на каждой сгенерированной выборке позволяет оценить ее выборочное распределение. Бустрэп может отталкиваться от исходного множества клиентов, либо учитывать выделенные сегменты. Также анализ может обеспечивать стратификацию, то есть сохранять постоянной долю классов в генерируемых выборках [14]. При использовании метода бутстрэп возможно два варианта тестовой статистики для проверки гипотезы о равенстве AUROC. Во-первых, это классический тест Стьюдента (У. Госсет) о равенстве средних для зависимых выборок. Во-вторых, статистика, предложенная К. Робином [11], вычисляемая как отношение разности AUROC на оригинальной выборке к стандартному отклонению разностей AUROC на сгенерированных выборках и имеющая нормальное распределение.

Для выбора одного из методов оценки значимости отличия двух AUROC мы применили каждый вариант для одного из сопоставлений (*табл. 3*). Не обнаружив различий в результатах, мы отобрали три метода: Делонг, нестратифицированный бутстрэп с тестом Стьюдента и статистикой из pRoc. Применив эти методы на всем наборе возможных сопоставлений моделей, мы поняли, что тест Стьюдента за редким исключением не обнаруживает значимого отличия двух AUROC, когда два других метода их обнаруживают. Также заключаем, что нестратифицированный бутстрэп со статистикой из pRoc является наиболее чувствительным к наличию значимых различий из рассмотренных методов, поэтому мы остановились на нем.

Были рассмотрены 8 методов эвристической сегментации клиентов розничного банка из банковской практики, 8 методов из литературы и 17 авторских методов (*табл. 5*). Сегменты на основе категориальной переменной формируются группировкой значений в бакеты на основе уровня дефолтности. Числовые предикторы подверглись дискретизации с дальнейшим объединением бакетов также по уровню дефолтности для обеспечения достаточного количества

наблюдений в каждом сегменте. Стоит отметить, что дискретизация производилась отдельно для кредитных карт и ссуд наличными.

Из банковской практики не использована сегментация по уровню дохода (mass retail, mass affluent, HNWI, UHNWI), так как почти все данные отражают кредитование клиентов, с зарплатой до 5 тыс. долл. США.

Результаты

Были сопоставлены наборы скоринговых карт, построенных на основе сегментов, выделенных 33 эвристическими методами сегментации. Базовым вариантом выступает набор скоринговых карт двух кредитных продуктов.

Согласно результатам по величине метрики эффективности классификатора (AUROC) наиболее предпочтительным оказался вариант сегментации по величине выданной ссуды – amount_credit (*рис. 4*). Более того, AUROC данного метода статистически значимо на 10% уровне отличается от AUROC остальных вариантов сегментации (белая заливка p-уровня значимости в матрице). Если более подробно, то только вариант сегментации на основе переменной annuity_to_all_appr (отношение размера аннуитета текущего займа к среднему аннуитету всех ранее выданных кредитов в данном банке) превышает p-значение в 0,05. Наборы скоринговых карт на основе остальных эвристических методов сегментации значимо менее эффективны на 1% уровне (p-значение менее 0,01).

Границы сегментов по величине выданной ссуды и результаты моделирования представлены в *табл. 4*. Можно видеть, что для одного из сегментов в рамках кредитных карт получена переобученная модель. Подобные результаты характерны и для других вариантов сегментации. При этом на оставшихся данных получаются довольно устойчивые модели. Вероятно, причина кроется в качестве данных для этого небольшого количества выдач кредитных карт.

Также полученные результаты свидетельствуют о возможности положительного эффекта от сегментации клиентов на эффективность скоринговой системы. Такие методы сегментации, как is_BKI (наличие кредитной истории в БКИ), is_mortgage (наличие ипотечного займа), amount_credit (величина выданной ссуды) и annuity_to_all_appr (отношение размера аннуитета текущего займа к среднему аннуитету всех ранее выданных кредитов в данном банке) на 10% уровне значимости дают более эффективный скоринг, чем базовый вариант.

Выводы

В данной работе, как и в большинстве подобных исследований, был доказан положительный эффект от сегментации клиентов на эффективность скоринговой системы. Предложенный авторами метод сегментации превзошел по выбранной метрике эффективность остальных эвристических методов сегментации.

В работе была устранена слабость многих подобных исследований: были созданы специфичные для сегментов переменные. Из слабых сторон исследования стоит отметить, что были сопоставлены лишь эвристические методы сегментации, которые тем не менее составляют доминирующую долю в российской банковской практике.

Возможным продолжением данной работы может стать рассмотрение статистических методов сегментации. Также возможно подробное изучение причин низкой эффективности некоторых из рассмотренных вариантов сегментации. Возможно, причиной мог стать дисбаланс классов в каких-то сегментах. Для преодоления чего возможно прибегнуть к балансированию классов. Причиной могло стать переобучение моделей на узких сегментах клиентов, что возможно исправить более «грубой» сегментацией.

Таблица 1
Исследования эвристических методов сегментации клиентов розничного банка

Table 1
Studies on heuristic methods for segmentation of retail bank customers

Исследование	Критерии сегментации	Основные выводы
Г. Чендлер и др. [15]	Пол клиента	Исключение гендера клиента контринтуитивно приводит к меньшему одобрению заявок от женщин
Н. Банасик и др. [16]	Брачный статус, наличие детей, наличие дохода супруга, уход на пенсию, срок проживания по адресу, наличие банковского счета, количество лет с банком, наличие кредитной карты, расходы на обслуживание ипотеки/рассрочки, возраст, жилой статус (собственник/арендатор), наличие ипотеки	Лишь брачный статус и уход на пенсию из рассмотренных критериев сегментации привели к улучшению эффективности скоринга
У. Фирдаус и др. [17]	Баланс клиента как сумма доходов банка от ссуды	Баланс клиента позволяет выделить маркетинговые сегменты клиентов
Г. Лис и др.[18]	Возраст, пол и социально-экономический статус	В разрезе рассмотренных характеристик не нарушается равномерность распределения клиентов по разным продуктам банка. Это говорит о том, что банки нацелены на работу с массовым сегментом рынка, не создавая продуктов под конкретные социально-демографические группы
Ш. Шашидхан и др.[19]	Соотношение просроченной суммы кредита и величины залога	Предложенный эвристический метод позволяет избежать проблемы «выбросов», характерной для кластерного анализа

Источник: авторская разработка

Source: Authoring

Таблица 2
Исследования влияния сегментации на эффективность кредитного скоринга

Table 2
Studies on the impact of segmentation on credit scoring effectiveness

Исследование	Выводы
В. Мэкух [20]	Как правило, сегментация повышает эффективность скоринга на 5—10% относительно скоринговой модели, построенной на всем множестве клиентов. Несмотря на то, что основанная на опыте сегментация может повысить эффективность скоринговых моделей для отдельных сегментов, нет гарантий на подобное повышение для всего множества клиентов
К. Бижак и Л. Томас [7]	Авторы констатируют, что разработчиками скоринговых моделей обычно утверждается, что набор скоринговых карт (отдельная модель для каждого сегмента) позволяет лучше оценить кредитный риск, нежели общая модель для всех клиентов. В своем исследовании авторы обнаружили, что ни один из наборов скоринговых карт не показал значимо более высокую эффективность, чем у единой модели. Единственным преимуществом набора моделей служит более удобная настройка порогов отсечения, так как это легче сделать по отдельным ROC-кривым. Авторы отмечают, что разделяющая способность сегментации может превышать эффективность отдельных скоринговых моделей. И это может быть результатом того, что сегментация не оставляет для скоринговых моделей пространства для дальнейшей дискриминации заемщиков
Н. Банасик и др.[16]	Авторы установили пороги отсечения по величине скорингового балла для одной общей модели и для моделей, построенных на отдельных сегментах, и измерили эффективность по доле ошибок на отложенной выборке. В результате авторы пришли к выводу, что сегментирование не всегда приводит к повышению эффективности
С. Скитовски и др. [21]	Авторы предполагают, что отдельное моделирование для каждого кластера позволит повысить качество скоринговых моделей на 5—10%

Источник: авторская разработка

Source: Authoring

Таблица 3
Приложение доступных тестов проверки равенства двух AUROC

Table 3
Application of available tests for checking the equality of two AUROCs

Метод	Тестовая статистика	Вариант	Стратификация	Модель № 1	Модель № 2	p-value
Бутстрэп	Из пакета pRoc	Отбор с возвращением	Нет	0,691	0,71	0,00871
			Да	0,691	0,71	0,00503
		Отбор с возвращением отдельно по сегментам	Нет	0,691	0,71	0,00573
			Да	0,691	0,71	0,00756
	Из теста Стьюдента	Отбор с возвращением	Нет	0,6906	0,7099	0
			Да	0,6906	0,7095	0
		Отбор с возвращением отдельно по сегментам	Нет	0,6902	0,7094	0
			Да	0,6908	0,7098	0
Делонг	—	—	—	—	—	0

Источник: авторская разработка

Source: Authoring

Таблица 4
Моделирование с сегментацией по величине выданной ссуды

Table 4
Modeling with segmentation by disbursed loan amount

Показатель	Сегмент	Уровень дефолтов, %	Количество	Джини, %
Кредит наличными «>=2170$»	oot	6,62	11 732	48,39
	test	6,78	9 525	48,69
	train	6,86	38 261	48,86
Кредит наличными «<2170$»	oot	9,33	18 122	54,61
	test	9,35	14 586	55,58
	train	9,41	57 774	54,98
Кредитные карты «>=690$»	oot	4,83	3 623	46,41
	test	4,21	2 801	46,88
	train	4,15	11 263	58,56
Кредитные карты «<690$»	oot	6,72	2 218	44,02
	test	8,54	1 850	44,45
	train	7,18	7 245	47,35

Источник: авторская разработка

Source: Authoring

Таблица 5
Рассмотренные варианты сегментации

Table 5
Considered options of segmentation

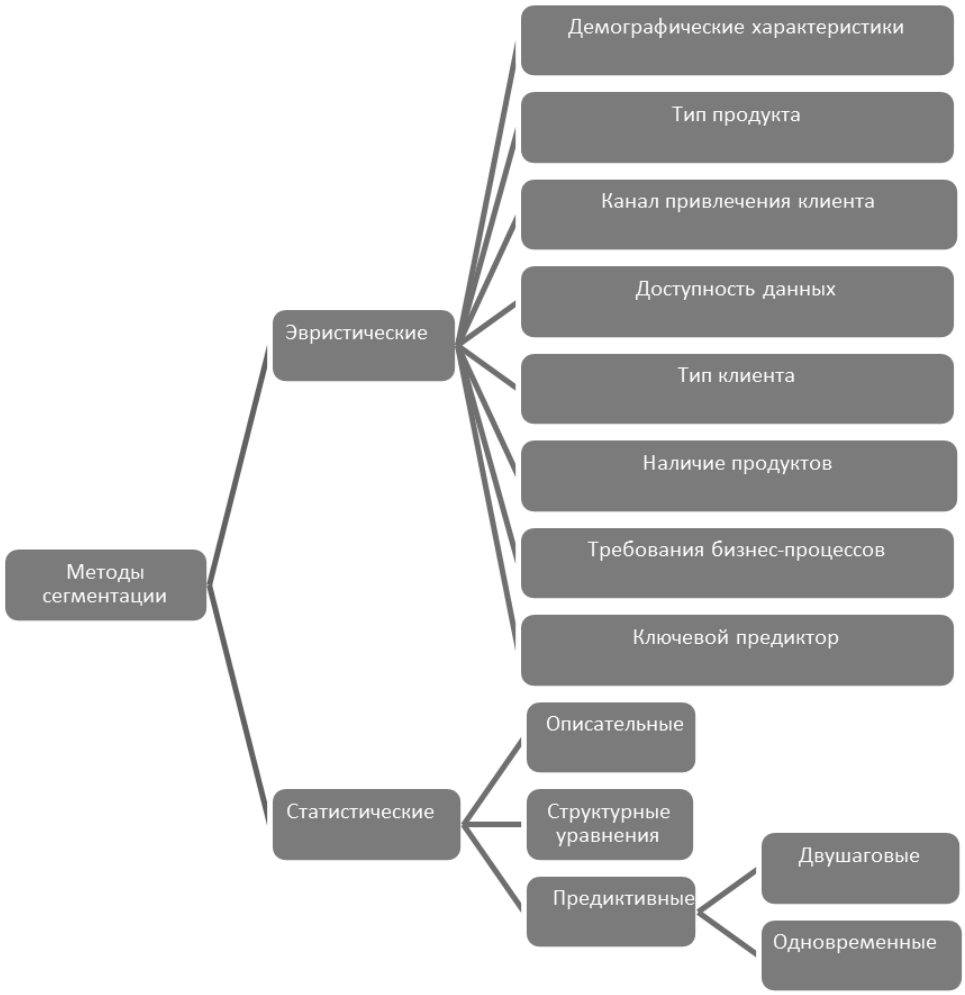
Название сегментации	Содержание переменной для выделения сегментов	Источник
contract_type	Тип текущего кредитного продукта (кредитная карта или ссуда наличными)	Базовый
available_data_sources	Наличие источников данных (КИ, БКИ)	Практика
gender_family_status	Пол клиента и брачный статус	Практика
is_BKI	Наличие кредитной истории в БКИ	Практика
is_CH	Наличие кредитной истории в данном банке (выданные кредиты)	Практика
molodezh_pensioneer	Флаг молодежи/пенсионера	Практика
pokolenie	Поколения теории Хоува—Штрауса	Практика
RBC	Сегментация Королевского банка Канады	Практика
region_population	Размер численности населения места проживания	Практика
avg_max_overdue_bki	Средняя максимальная просроченная задолженность по кредитам из БКИ	Литерат.
gender	Пол клиента	Литерат.
is_own_realty	Наличие собственной жилой недвижимости	Литерат.
is_mortgage	Наличие ипотечного займа	Литерат.
is_retired	Флаг пенсионера	Литерат.
family_status	Брачный статус	Литерат.
income_type	Тип источника доходов	Литерат.
yield_group	Группа доходности клиента для банка	Литерат.
annuity_all_appr	Средний аннуитет по выданным кредитам в истории данным банком	Автор.
annuity_last_same_appr	Аннуитет по последнему выданному банком аналогичному кредитному продукту	Автор.
annuity_to_all_appr	Отношение размера аннуитета текущего займа к среднему аннуитету всех ранее выданных кредитов	Автор.
annuity_to_last_appr	Отношение размера аннуитета текущего займа к аннуитету последнего выданного кредита	Автор.
amount_credit	Величина выданного кредита	Автор.
amount_price_all_appr	Средняя цена товаров, для покупки которых были выданы потребительские ссуды данным банком	Автор.
amount_goods_price	Цена товаров, для покупки которых выдана текущая потребительская ссуда	Автор.
amount_req_BKI_year	Количество обращений за ссудой в банки за последний год	Автор.
cnt_active_BKI	Количество активных кредитных договоров в БКИ	Автор.
cnt_appr_cash	Количество выданных невольверных кредитов (потребительские ссуды, ссуды наличными) данным банком	Автор.
cnt_closed_BKI	Количество закрытых кредитных договоров в БКИ	Автор.
cnt_refused_cash	Количество отклоненных клиентом невольверных кредитов (потребительские ссуды, ссуды наличными) в данном банке	Автор.
days_employed	Трудовой стаж на последнем рабочем месте	Автор.
is_work_phone	Предоставил ли клиент рабочий номер телефона	Автор.
is_own_car	Наличие собственного автомобиля	Автор.
is_refused_in_hist_all	Наличие отклоненных клиентом кредитов в данном банке	Автор.
is_refused_in_hist_cash	Наличие отклоненных клиентом невольверных кредитов (потребительские ссуды, ссуды наличными) в данном банке	Автор.

Источник: авторская разработка

Source: Authoring

Рисунок 1
Классификация методов сегментации клиентов розничного банка для целей аппликативного скоринга

Figure 1
Classification of methods for market segmentation of retail bank customers for the purpose of applicative scoring

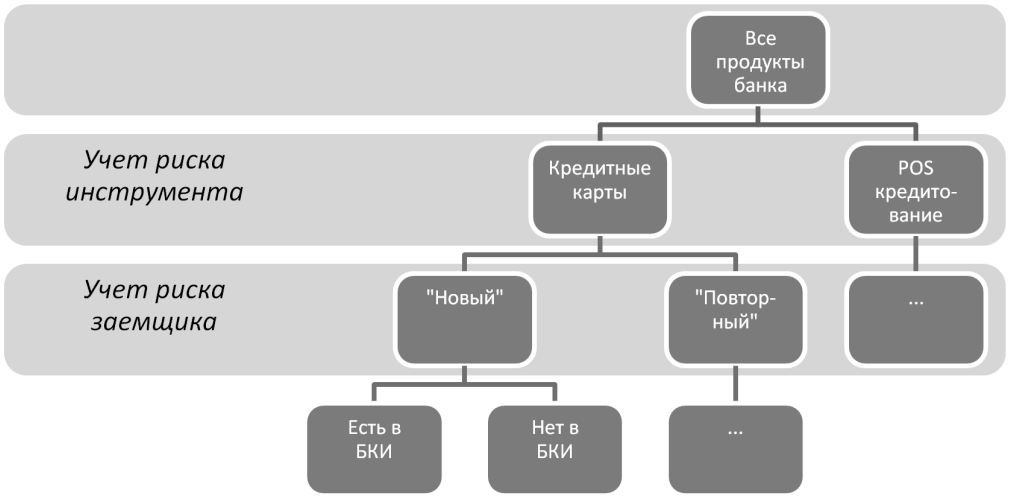


Источник: построено авторами на основе источников: [1—3]

Source: Authoring, based on sources: [1–3]

Рисунок 2
Стандартная карта моделей розничного банка

Figure 2
Standard map of retail bank models

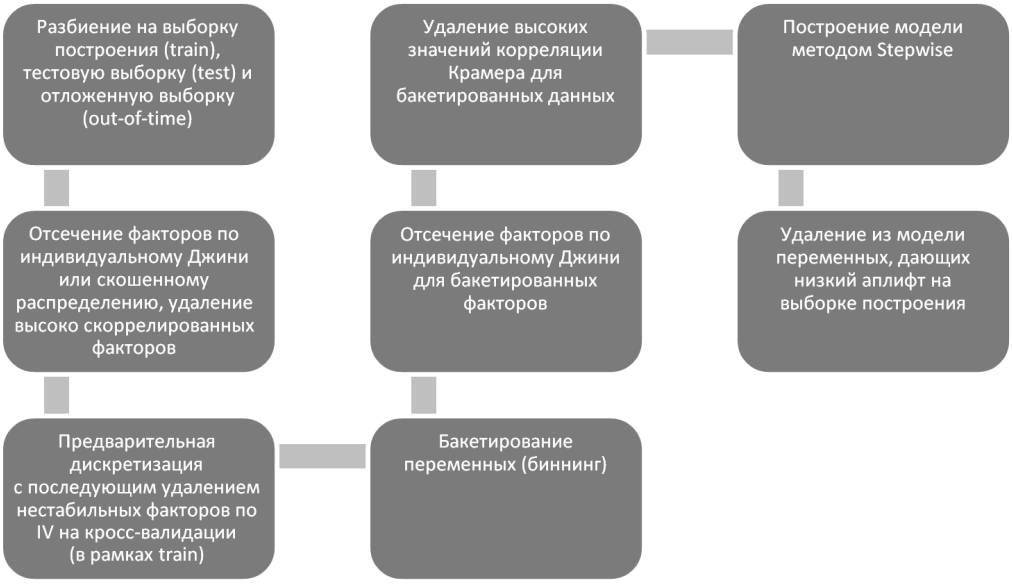


Источник: авторская разработка

Source: Authoring

Рисунок 3
Схема обучения классификатора

Figure 3
Classifier training scheme

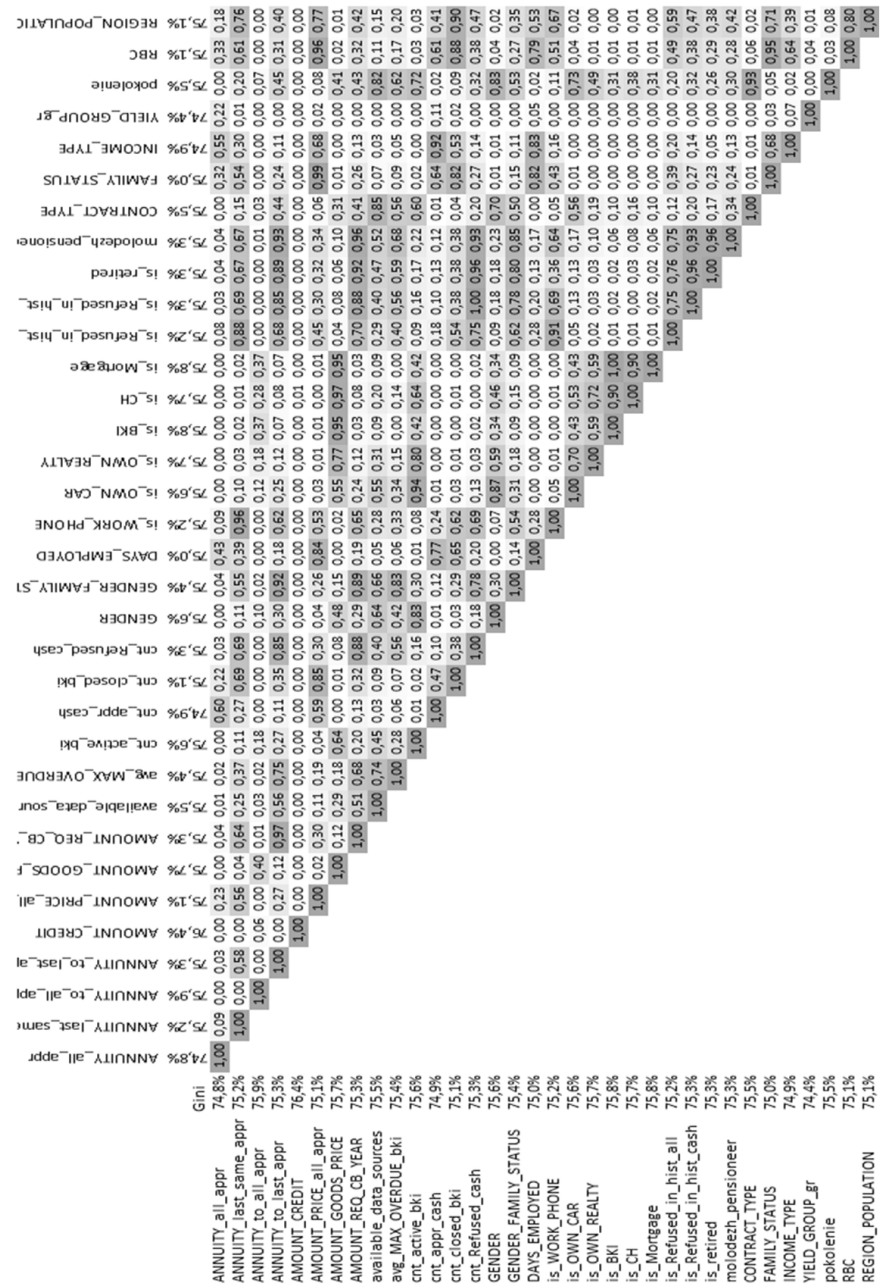


Источник: авторская разработка

Source: Authoring

Рисунок 4
Проверка значимости отличия auroc на отложенной выборке между наборами скоринговых карт

Figure 4
Checking the significance of AUROC differences, using the deferred sampling between the sets of scorecards



Примечание. Приведенные значения p-value округлены до второго знака.

Источник: авторская разработка

Source: Authoring

Список литературы

1. *Siddiqi N.* Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Hoboken, N.J., Wiley, 2005, 208 p.

2. *Aurifeille J.-M.* A bio-mimetic approach to marketing segmentation: Principles and comparative analysis. *European Journal of Economic and Social Systems*, 2000, vol. 14, pp. 93–108. URL: http://dx.doi.org/10.1051/ejess:2000111

3. *Thomas L.C.* Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press, 2009, 385 p.

4. *Бояринцев В.А., Бровкина Н.Е.* Современные подходы к сегментации клиентской базы розничного банка // Universum: экономика и юриспруденция. 2016. № 12. С. 12—16.
URL: https://7universum.com/ru/economy/archive/item/3936

5. *Галкина Н.А.* Потенциальные сегменты населения для расширения клиентской базы коммерческих банков в условиях старения населения // Вестник Московского университета. Серия 6: Экономика. 2015. № 1. С. 60—86. URL: https://cyberleninka.ru/article/n/potentsialnye-segmenty-naseleniya-dlya-rasshireniya-klientskoy-bazy-kommercheskih-bankov-v-usloviyah-stareniya-naseleniya

6. *Карпова С.В., Рожков И.В., Воронина В.С.* Критерии и признаки сегментации потребителей банковских услуг // Практический маркетинг. 2020. № 6. С. 3—9. URL: https://cyberleninka.ru/article/n/kriterii-i-priznaki-segmentatsii-potrebiteley-bankovskih-uslug

7. *Bijak K., Thomas L.C.* Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, 2012, vol. 39, iss. 3, pp. 2433–2442. URL: https://doi.org/10.1016/j.eswa.2011.08.093

8. *Hand D.J., Sohn S.Y., Kim Y.* Optimal bipartite scorecards. *Expert Systems with Applications*, 2005, vol. 29, iss. 3, pp. 684–690.
URL: https://doi.org/10.1016/j.eswa.2005.04.032

9. *Hanley J.A., McNeil B.J.* The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 1982, vol. 43, no. 1, pp. 29–36.
URL: https://doi.org/10.1148/radiology.143.1.7063747

10. *Pepe M., Longton G., Janes H.* Estimation and Comparison of Receiver Operating Characteristic Curves. *Stata Journal*, 2009, vol. 9, no. 1, pp. 1–16.
URL: https://doi.org/10.1177/1536867X0900900101

11. *Robin X. et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 2011, vol. 12. URL: https://doi.org/10.1186/1471-2105-12-77

12. *DeLong E.R., DeLong D.M., Clarke-Pearson D.L.* Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 1988, vol. 44, no. 3, pp. 837–845. URL: https://doi.org/10.2307/2531595

13. *Sun X., Xu W.* Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters*, 2014, vol. 21, iss. 11, pp. 1389–1393. URL: https://ieeexplore.ieee.org/document/6851192/

14. *Pons O.* Bootstrap of means under stratified sampling. *Electronic Journal of Statistics*, 2007, vol. 1, pp. 381–391. URL: https://doi.org/10.1214/07-EJS033

15. *Chandler G.G., Ewert D.C.* Discrimination on the basis of sex under the equal credit opportunity act. Krannert Graduate School of Management, Purdue University, 1976, 20 p.

16. *Banasik J.L., Crook J.N., Thomas L.C.* Does scoring a subpopulation make a difference? *The International Review of Retail, Distribution and Consumer Research*, 1996, vol. 6, iss. 2, pp. 180–195. URL: https://doi.org/10.1080/09593969600000019

17. *Firdaus U., Utama D.N.* Balance as one of the attributes in the customer segmentation analysis method: Systematic literature review. *A Bimonthly Peer-Review Journal*, 2020, vol. 5, iss. 3, pp. 334–339. URL: http://dx.doi.org/10.25046/aj050343

18. *Lees G., Winchester M., De Silva S.* Demographic product segmentation in financial services products in Australia and New Zealand. *Journal of Financial Services Marketing*, 2016, vol. 21, pp. 240–250. URL: https://doi.org/10.1057/s41264-016-0004-3

19. *Shashidhar H.G., Subramanian V.* Customer Segmentation of Bank based on Data Mining – Security Value based Heuristic Approach as a Replacement to *k*-means Segmentation. *International Journal of Computer Applications*, 2011, vol. 19, no. 8, pp. 13–18. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.735.214&rep=rep1&type=pdf

20. *Mays E.* Handbook of Credit Scoring. AMACOM, 2001, 370 p.

21. *Scitovski S., Sarlija N.* Cluster analysis in retail segmentation for credit scoring. *Croatian Operational Research Review*, 2014, vol. 5, no. 2, pp. 235–245. URL: https://doi.org/10.17535/crorr.2014.0010

Информация о конфликте интересов

Мы, авторы данной статьи, со всей ответственностью заявляем 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.

pISSN 2071-4688
eISSN 2311-8709

Banking

SEGMENTATION OF RETAIL BANK CUSTOMERS FOR THE PURPOSES OF MODELING THE LOAN CLAIM DEFAULT

Elena V. PAVLOVA [a,•], Vladislav V. ROSKOSHENKO [b]

[a] Lomonosov Moscow State University,
Moscow, Russian Federation
lena.pavlova@gmail.com
ORCID: not available

[b] Lomonosov Moscow State University,
Moscow, Russian Federation
roskoshenkoeco@mail.ru
ORCID: not available

[•] Corresponding author

Article history:
Article No. 581/2020
Received 15 Oct 2020
Received in revised form
29 October 2020
Accepted 12 Nov 2020
Available online
27 November 2020

JEL classification: G21

Keywords: credit scoring, logistic regression, heuristic segmentation, ROC curve

Abstract

Subject. In the banking practice, approaches to separate modeling of loan claim default (for new and repeat customers, for customers having and not having a history in the Credit Bureau, etc.) are widespread. Such a segmentation of retail bank customers may increase the efficiency of applied scoring system. The practical problem of choosing the optimal heuristic method of segmentation for the scoring remains unresolved.

Objectives. The purpose of this work is to determine the optimal heuristic method of segmentation from those that are known in the literature and the industry.

Methods. The study employs statistical analysis and content analysis of information sources.

Results. We compared over thirty heuristic methods for segmentation of retail bank customers. The comparison showed that according to the classifier of the efficiency metric (AUROC), our proposed segmentation by the disbursed loan size turned out to be optimal. The method consists in the 'disbursed loan' variable discretization under the TreeR method.

Conclusions and Relevance. The findings may be helpful in loan scoring and in any statistical modeling, using the logit regression.

Please cite this article as: Pavlova E.V., Roskoshenko V.V. Segmentation of Retail Bank Customers for the Purposes of Modeling the Loan Claim Default. *Finance and Credit*, 2020, vol. 26, iss. 11, pp. 2594–2616.
https://doi.org/10.24891/fc.26.11.2594

References

1. Siddiqi N. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Hoboken, N.J., Wiley, 2005, 208 p.

2. Aurifeille J.-M. A bio-mimetic approach to marketing segmentation: Principles and comparative analysis. *European Journal of Economic and Social Systems*, 2000, vol. 14, no. 1, pp. 93–108. URL: http://dx.doi.org/10.1051/ejess:2000111

3. Thomas L.C. Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press, 2009, 385 p.

4. Boyarintsev V.A., Brovkina N.E. [Modern processes to the customer base segmentation of the retail bank]. *Universum: ekonomika i yurisprudentsiya = Universum: Economics and Law*, 2016, no. 12, pp. 12–16. URL: https://7universum.com/ru/economy/archive/item/3936 (In Russ.)

5. Galkina N.A. [Potential population segments to expand the customer base of commercial banks in the context of population ageing]. *Vestnik Moskovskogo universiteta. Seriya 6: Ekonomika = Bulletin of Moscow University. Series 6: Economics*, 2015, no. 1, pp. 60–86. URL: https://cyberleninka.ru/article/n/potentsialnye-segmenty-naseleniya-dlya-rasshireniya-klientskoy-bazy-kommercheskih-bankov-v-usloviyah-stareniya-naseleniya (In Russ.)

6. Karpova S.V., Rozhkov I.V., Voronina V.S. [Criteria and traits of segmentation of banking services consumer]. *Prakticheskii marketing = Practical Marketing*, 2020, no. 6, pp. 3–9. URL: https://cyberleninka.ru/article/n/kriterii-i-priznaki-segmentatsii-potrebiteley-bankovskih-uslug (In Russ.)

7. Bijak K., Thomas L.C. Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, 2012, vol. 39, iss. 3, pp. 2433–2442. URL: https://doi.org/10.1016/j.eswa.2011.08.093

8. Hand D.J., Sohn S.Y., Kim Y. Optimal bipartite scorecards. *Expert Systems with Applications*, 2005, vol. 29, iss. 3, pp. 684–690. URL: https://doi.org/10.1016/j.eswa.2005.04.032

9. Hanley J.A., McNeil B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 1982, vol. 43, no. 1, pp. 29–36. URL: https://doi.org/10.1148/radiology.143.1.7063747

10. Pepe M., Longton G., Janes H. Estimation and Comparison of Receiver Operating Characteristic Curves. *Stata Journal*, 2009, vol. 9, no. 1, pp. 1–16. URL: https://doi.org/10.1177/1536867X0900900101

11. Robin X. et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 2011, vol. 12. URL: https://doi.org/10.1186/1471-2105-12-77

12. DeLong E.R., DeLong D.M., Clarke-Pearson D.L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 1988, vol. 44, no. 3, pp. 837–845. URL: https://doi.org/10.2307/2531595

13. Sun X., Xu W. Fast Implementation of DeLong's Algorithm for Comparing the Areas under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters*, 2014, vol. 21, iss. 11, pp. 1389–1393. URL: https://ieeexplore.ieee.org/document/6851192/

14. Pons O. Bootstrap of means under stratified sampling. *Electronic Journal of Statistics*, 2007, vol. 1, pp. 381–391. URL: https://doi.org/10.1214/07-EJS033

15. Chandler G.G., Ewert D.C. Discrimination on the basis of sex under the equal credit opportunity act. Krannert Graduate School of Management, Purdue University, 1976, 20 p.

16. Banasik J.L., Crook J.N., Thomas L.C. Does scoring a subpopulation make a difference? *The International Review of Retail, Distribution and Consumer Research*, 1996, vol. 6, iss. 2, pp. 180–195. URL: https://doi.org/10.1080/09593969600000019

17. Firdaus U., Utama D.N. Balance as one of the attributes in the customer segmentation analysis method: Systematic literature review. *A Bimonthly Peer-Review Journal*, 2020, vol. 5, iss. 3, pp. 334–339. URL: http://dx.doi.org/10.25046/aj050343

18. Lees G., Winchester M., De Silva S. Demographic product segmentation in financial services products in Australia and New Zealand. *Journal of Financial Services Marketing*, 2016, vol. 21, pp. 240–250. URL: https://doi.org/10.1057/s41264-016-0004-3

19. Shashidhar H.G., Subramanian V. Customer Segmentation of Bank based on Data Mining – Security Value based Heuristic Approach as a Replacement to *k*-means Segmentation. *International Journal of Computer Applications*, 2011, vol. 19, no. 8, pp. 13–18. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.735.214&rep=rep1&type=pdf

20. Mays E. Handbook of Credit Scoring. AMACOM, 2001, 370 p.

21. Scitovski S., Sarlija N. Cluster analysis in retail segmentation for credit scoring. *Croatian Operational Research Review*, 2014, vol. 5, no. 2, pp. 235–245. URL: https://doi.org/10.17535/crorr.2014.0010

Conflict-of-interest notification

We, the authors 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.