Опыт моделирования вероятности кредитного дефолта клиентов микрофинансовых организаций (на примере одной МФО)

Поляков К.Л., Жукова Л.В.

Микрофинансовые организации получили большое распространение в кризисные годы, выдавая микрокредиты (до 100000 рублей) под большие проценты практически без документов. Сегодня ЦБ РФ активно регулирует этот рынок, все больше и больше ужесточая требования, ограничивая ставки и пени на выданные кредиты. Это вызывает необходимость формирования новой стратегии оценки риска невозврата выданного кредита или займа, построенной на предотвращении просрочек со стороны клиентов. Для этого, во-первых, необходимо получать более информативные данные о клиентах, не осложняя взаимоотношения с ними. Во-вторых, необходимо иметь хорошее представление о возможностях тех или иных методов классификации при решении различных задач оценки потенциальных клиентов. Авторы данного исследования анализируют важность для качества классификации клиентов тех показателей, что уже традиционно собираются МФО, а также некоторых новых показателей, основанных на данных из социальных сетей. В данном случае важность показателей интерпретируется в контексте конкретных алгоритмов (методов) классификации.

Для моделирования кредитного дефолта (просрочки более 30 дней) авторы используют несколько алгоритмов построения деревьев классификации – алгоритмы CART и C 4.5, логистическую регрессию и алгоритм Random Forest (случайный лес). Моделирование осуществляется на основе выборки из анкет клиентов реальной МФО. Получены неоднозначные результаты. В зависимости от постановки задачи классификации клиентов преимуществом обладают различные алгоритмы дескриптивной аналитики (CART, C4.5, Logit). В тоже время, как и следовало ожидать, наилучшее качество прогнозов дает неинтерпретируемый прогностический алгоритм Random Forest. По результатам анализа было выявлено, что для классификации клиентов МФО во всех рассмотренных методах большой важностью обладает кредитная ис-

Поляков Константин Львович – к.т.н., доцент департамента прикладной экономики Национального исследовательского университета «Высшая школа экономики». E-mail: polyakov.kl@hse.ru
Жукова Людмила Вячеславовна – старший преподаватель департамента прикладной экономики Национального исследовательского университета «Высшая школа экономики». E-mail: lvzhukova@hse.ru

Статья поступила: 10.09.2019/Статья принята: 20.11.2019.

тория заемщика, а также его возраст. Гендерный фактор не оказал влияния на результаты классификации. Также во всех случаях оказались неважными для классификации данные из социальных сетей.

***Ключевые слова*:** микрофинансовые организации; дефолт; дерево классификации; логистическая регрессия; случайный лес.

DOI: 10.17323/1813-8691-2019-23-4-497-523

Для цитирования: Поляков К.Л., Жукова Л.В. Опыт моделирования вероятности кредитного дефолта клиентов микрофинансовых организаций (на примере одной МФО). *Экономический журнал ВШЭ*. 2019; 23(4): 497–523.

For citation: Polyakov K., Zhukova L. Modeling the Probability of Credit Default of Clients of Microfinance Organizations: The Case of One MFI. *HSE Economic Journal.* 2019; 23(4): 497–523. (In Russ.)

1. Введение

Управление рисками традиционно является одной из важнейших задач участников финансового рынка. Клиентская аналитика, которая позволяет оценить вероятность дефолта потенциального заемщика (наличие просроченного платежа с задержкой более 30 дней), всегда находилась в центре внимания руководства кредитных организаций. Отсутствие общей теории в этой области знаний повышает значимость эмпирических исследований и практического опыта отдельных организаций. В то же время это обстоятельство приводит к накоплению больших массивов данных, далеко не все из которых могут быть полезны в решении указанной выше задачи и собираются «на всякий случай», «по традиции», «исходя из опыта участников рынка» и т.д. Количество учитываемых показателей растет драматически с развитием ИТ. При этом очень велика доля качественных данных.

Основной целью данного исследования является оценка важности факторов, учитываемых микрофинансовыми организациями (МФО) при заключении договоров кредитования для прогнозирования дефолта потенциального заемщика. Результаты подобных исследований позволяют повысить эффективность клиентской аналитики. В рамках решения поставленной задачи выполняется сравнительный анализ качества классификации некоторых методов дескриптивной и прогностической аналитики [Berk, 2008] в ситуации, когда большая часть независимых переменных (или все переменные) являются качественными (номинальными или порядковыми) с большим числом уровней. В этом случае некоторые методы классификации или их популярные реализации требуют перекодирования качественных данных в системы фиктивных переменных. Если уровней достаточно много и они хорошо представлены в тренировочном множестве, т.е. укрупнение уровней не имеет смысла или неадекватно решаемой задаче, это приводит к драматическому росту ее размерности. В результате исследователь сталкивается с эффектом «проклятия размерности» (*the curse of dimensionality*) [Hastie et al., 2001]. Существует мно-

жество примеров проявлений этого эффекта. Например, плотность выборки объема N в p–мерном пространстве пропорциональна $N^{\frac{1}{p}}$. Таким образом, чтобы обеспечить в десятимерном пространстве ту же плотность, которую имеет в одномерном пространстве выборка объема N = 100, потребуется порядка 100^{100} измерений. Аналогично, чтобы отобрать 10% выборки для вычисления, например, локального среднего, потребуется использовать около 80% размаха каждого фактора [Hastie et al., 2001]. Таким образом, с ростом размерности задачи требуется адекватное увеличение числа измерений для сохранения точности оценки силы влияния каждого фактора. В то же время далеко не всегда можно обеспечить требуемый рост объема тренировочного множества. В ряде случаев он ограничен за счет специфики изучаемого объекта (системы). Если возникают подобные ситуации, крайне важно знать, насколько те или иные методы прогнозирования/классификации чувствительны к «проклятию размерности».

Авторы данного исследования остановились на четырех методах классификации, которые давно и заслуженно занимают первые строки в списках популярных инструментов бизнес-аналитиков. К ним относятся:

- два метода построения деревьев классификации – на основе алгоритмов CART [Breiman et al., 1984] и C4.5 [Quinlan, 1993];
- логистическая регрессия [Hosmer et al., 2001];
- метод классификации на основе алгоритма «случайного леса» (*random forest*) [Hastie et al., 2001].

Первые три относятся к дескриптивным методам, они позволяют получить интерпретируемые (человеко-читаемые) модели, четвертый относится к методам прогностической аналитики. Выбор методов не случаен. Задачи дескриптивной аналитики крайне важны для планирования деятельности организации, когда необходимо получить ответ на вопросы категории «Что будет, если …?» В частности, есть необходимость дать описание целевому множеству клиентов для организации маркетинговых коммуникаций. Вместе с тем можно предположить, что поддержка интерпретируемости модели влечет за собой снижение качества прогнозирования по сравнению с методами прогностической аналитики. В связи с этим для сравнения с дескриптивными методами моделирования в анализ введен алгоритм из области прогностической аналитики – «случайный лес».

2. Обзор литературы

Теме управления рисками в кредитных организациях и, в частности, в микрофинансовых организациях, посвящено немало публикаций. Остановимся на некоторых из них, тесно связанных с темой настоящего исследования.

Монография [Криворучко и др., 2013] посвящена специфике этого сегмента рынка финансовых услуг в Российской Федерации. В главе «Риски микрофинансирования и их регулирование» авторы выделяют четыре основных типа рисков для МФО (кредитный, рыночный, стратегический риски и риск ликвидности) и дают описание подходов к их регулированию. Так, стратегический риск, по мнению авторов, связан с возможностью потери репутации или неправильным выбором стратегии. Для управления этим видом риска отмечается важность знаний, умений и навыков высшего руководства МФО и его

наблюдательного совета. Отмечено, что управление операционным риском в МФО осложняется возросшей зависимостью от информационных технологий, слабым внутренним контролем, отсутствием корпоративной этики, низким профессионализмом персонала. Кредитный риск, как признают многие участники рынка, является самым существенным. Авторы монографии уделяют большое внимание особенностям этого риска, его подразделам – риску отдельной трансакции и портфельному риску. В работе дается обоснование причин возникновения этого риска, связанных как с недостаточно высоким уровнем профессионализма сотрудников, так и с особенностями работы МФО – краткие сроки кредитов, большой объем клиентской базы, отсутствие оперативного контроля заемщиков после выдачи кредита. Среди возможных решений задачи управления этим видом риска авторы работы предлагают типизировать заемщиков по желанию и способности погасить кредит, что должно позволить повысить качество управления риском. Для этого предлагается использовать набор индикаторов для классификации клиентов, а также вести тактическое отслеживание заемщиков после выдачи кредита.

Значимости рынка микрокредитования для российского общества и отечественной экономики посвящена работа [Уксусова, 2018]. Автор полагает, что главной задачей микрофинансовой организации является финансовая поддержка малоимущих граждан. В работе приведено описание истории развития рынка микрокредитов в России, изменение законодательства в этой сфере, отмечено ужесточение требований к микрофинансовым (МФК) и микрокредитным компаниям (МКК). Автор характеризует основные отличия этих компаний – по лимитам на сумму выдаваемых кредитов и по требованиям обязательного раскрытия финансовой отчетности. Отмечается также большой рост рынка микрокредитов за последние годы, высокие темпы прироста числа выданных кредитов и займов, рост конкуренции. Среди основных проблем МФО автор указывает рост числа неплательщиков, а также рост совокупной задолженности перед МФО. В работе отмечено, что важной тенденцией в области законодательства, регулирующего деятельность МФО, является введение ограничений на размер просрочки и на процент просрочки в день. Это обстоятельство существенно повышает актуальность исследования оценки платежеспособности заемщика.

В связи с необходимостью развития клиентской аналитики большое количество публикаций посвящено теме моделирования вероятности дефолта клиентов кредитной организации (значимой просрочки погашения кредита) или их классификации по отношению к возможному возникновению дефолта.

В статье [Снегова, 2013] автор предлагает новую скоринговую модель, основанную на методе логистической регрессии. Новым подходом, в отличие от классических моделей, является учет не только кредитных, но и операционных рисков (риски персонала), построенных на основе логистической регрессии. Даны практические рекомендации по методике построения подобной скоринговой системы и оценено качество прогнозирования. Для сравнения использованы ROC-кривые, доказана значимость фактора группы продаж на точке, а также возможность мошенничества со стороны партнеров.

Статья [Сорокин, 2014] посвящена оценке кредитоспособности заемщика – физического лица. В статье рассматривается методика эконометрического моделирования вероятности дефолта по кредитам на основе модели логистической регрессии. Акцентируется внимание на методологических аспектах построения модели. Большое внимание уделено выбору зависимой переменной. Автор использует для моделирования логисти-

ческую регрессию. Оценивание модели выполнялось на выборке из 650 клиентов. Также в работе рассматривалась методика построения скоринговых карт на основе модели логистической регрессии. Были изложены методические подходы к формированию и исследованию характеристик заемщика, исследована прогностическая возможность модели и обоснована ее высокая практическая значимость.

В работе [Софронова, 2016] отмечается рост просроченной кредитной задолженности для юридических лиц. Автор полагает, что необходимо улучшать способы управления кредитным риском как основным фактором роста просрочки по займам. Отмечается, что в системе количественной оценки кредитного риска заемщика слабым местом пока остается оценка вероятности дефолта. В практической части статьи построена модель оценки вероятности дефолта компании. По совокупности критериев логистическая регрессия была выбрана как более адекватная поставленной задаче. Важным значимым фактором оказался коэффициент платежеспособности, а также финансовое положение в исследуемом году. Итогом является предложение по совершенствованию управления кредитным риском, автор предлагает расширить понятие «дефолт заемщика», добавив и нежелание погасить кредитную задолженность в установленный договором срок.

3. Актуальные проблемы МФО

Клиентские базы многих микрофинансовых организаций включают в себя значительную долю клиентов с высокой долей риска дефолта, согласно оценкам банковских экспертов или скоринговых систем. Клиенты МФО – это, прежде всего, люди с плохой или отсутствующей кредитной историей, для которых обычные банковские продукты не доступны уже по ряду «стоп-факторов» в анкете (например, молодые предприниматели). Нестабильное финансовое положение, отсутствие необходимых документов – признаки клиентов МФО.

В результате за последний год МФО столкнулись с высоким уровнем невозврата выданных кредитов. Как отмечается в статье «Долги россиян перед МФО растут рекордными темпами»[1], граждане России просрочили возврат свыше 40% своих займов в МФО. Причем большая их часть имеет задержку возврата более 90 дней – так называемые «неработающие долги». Как правило, это небольшие суммы, например, займы «до зарплаты».

Влияние на рассматриваемый сегмент финансовых услуг этой просроченной задолженности велико. Как показывают результаты круглого стола «Микрофинансирование в России», проводимого экспертным агентством Эксперт РА[2], рост рынка микрокредитов в 2018 г. во многом связан с ростом просроченной невозвратной задолженности. По многочисленным исследованиям МФО, половина заемщиков действительно не в состоянии погасить долги и проценты по ним. Поэтому люди вынуждены обращаться в другие МФО за средствами для погашения уже сделанных займов и процентам по ним, все глубже опускаясь в долговую яму.

МФО пытаются своими силами оценить склонность клиента к дефолту. И здесь интересен опыт каждой организации. В частности, проанализировав портфель просрочен-

[1] См.: https://www.infox.ru/news/283/economy/finance/210653-dolgi-rossian-pered-mfo-rastut-rekordnymi-tempami

[2] См.: https://microcredit-rf.ru/servisi/expert-ra-kolichestvo-mfo-mozet-soktaycia.html

ных задолженностей, компания «Домашние деньги» пришла к выводу, что в большинстве случаев злостный неплательщик МФО – это заемщик, который уже на этапе оформления займа принимает решение не платить по нему. Компания, в настоящий момент находящаяся в стадии банкротства, одной из первых в стране занялась микрокредитованием и начала осваивать новый рынок еще в 2008 г. «Домашние деньги» до недавнего времени были на нем крупнейшим игроком, в апреле 2018 г. организация допустила технический дефолт по облигациям, в последний год столкнулась с угрожающей проблемой больших невозвратов и занималась анализом его причин.

Авторами было выдвинуто предположение, что эту склонность можно выявить путем сопоставления представленных заемщиком сведений с данными, полученными из социальных сетей. Результаты сравнения позволят оценить склонность клиента к искажению фактов. Например, можно предположить, что чем больше информации о человеке в его аккаунте, чем чаще он там появляется, тем более он склонен к публичности и тем меньше у него склонность к утаиванию данных о себе. Об открытости человека, социальной стабильности также может свидетельствовать количество его подписчиков («друзей») в социальной сети. Сравнение представленных данных (анкет) с данными из социальных сетей уже давно активно используют специалисты по набору персонала («Смотреть в профиль: как работодатели проверяют соцсети кандидатов»[3] на сайте подбора вакансий hh.ru). Анализируются содержание аккаунтов, круг друзей и настройки профиля.

Еще одним направлением использования социальных сетей является оценка финансового положения потенциального заемщика. В его анкетных данных, как правило, есть информация о заработной плате и дополнительном доходе, но чаще всего суммы завышены. Точную информацию о его доходах получить крайне сложно, если возможно вообще. Однако оценить корректность представленных данных можно по дополнительным сведениям из социальных сетей, например, из анкеты в аккаунте и в результате анализа выложенных фотографий. Сопоставление информации, косвенно свидетельствующей о расходах (отдых за границей, покупка машины, дома, крупные брендовые приобретения, свидетельства поездок, связанных с отдыхом, посиделки в ресторане), данные о платных медицинских процедурах, платном обучении, могут позволить оценить вероятность достоверности предоставленных сведений о заемщике.

3.1. Специфика микрокредитования

Спецификация системы классификации в любой предметной области требует хорошего понимания ее реалий. В связи с этим мы считаем необходимым остановиться на некоторых подробностях бизнеса МФО, необходимых для нашего исследования, в частности, для обоснования его актуальности.

Под микрокредитованием обычно понимают предоставление займов относительно небольшого объема (микрозаймов) клиентам, которые не удовлетворяют требованиям банков:

- недостаточен уровень дохода;
- нет собственности для предоставления залога;
- отсутствует устойчивая занятость;

[3] См.: https://hh.ru/article/301106

- нет верифицируемой положительной кредитной истории.

Российское законодательство[4] устанавливает предельную сумму микрозайма физическому лицу (основного долга физического лица перед МФО по договорам микрозайма) для МФО в виде микрофинансовой компании (МФК) в размере 1 млн руб., для микрокредитной компании (МКК) в размере 500 тыс. руб.

В нашем исследовании мы не будем делать различий между этими формами финансовых организаций и будем использовать следующее определение: под микрофинансовыми организациями (МФО) понимаются юридические лица, обладающие двумя признаками: они осуществляют микрофинансовую деятельность и зарегистрированы в государственном реестре микрофинансовых организаций. На рынке финансовых услуг такие компании занимают нишу коротких займов и предлагают особые кредиты. Они не предоставляют более широкий спектр традиционных банковских операций, таких как обслуживание пластиковых карт, ведение расчетных счетов, предоставление ипотечных кредитов или обслуживание корпоративных клиентов. Из сказанного видно, что деятельность МФО лежит в области повышенного риска невозврата кредитов, который компенсируется более высокими процентами по займам.

Можно выделить несколько основных типов микрокредитов.

«Переходные» – финансирование восходящих социальных движений, особенно в критические моменты перехода от одной группы к другой. К этим займам прибегают, например, индивидуальные предприниматели, развивающие свой бизнес, но имеющие недостаточно устойчивый доход для обычного банка. В этом случае клиент рассчитывает выплатить кредит с новых доходов от деятельности. Однако велик риск неудачи, который приводит к дефолту и потерям МФО.

«Стартапы» – финансирование новых идей и предпринимательства, когда руководитель проекта может получить кредит как физическое лицо под собственный автомобиль, не имея подтверждения дохода или вообще справки с места работы. В этом случае кредит может превратиться в просрочку при неуспешной реализации идеи, и МФО будет вынуждена искать должника для получения выданных средств.

«Финансирование последней надежды» – финансирование сопротивления угрозе нищеты, особенно когда она вызвана «внешними» обстоятельствами (такими как болезнь, неудача в бизнесе, потеря работы и т. д.). В этом случае заемщик, как правило, обращается к более дешевому банковскому финансированию в течение некоторого времени после наступления неблагоприятного события. Банк обычно не сразу получает информацию об этом, например, о потере работы клиентом. В результате ответные меры, например, отзыв кредитной карты или уменьшение ее кредитного лимита, принимаются с запаздыванием. Только после того, как более дешевые источники исчерпаны (дефолт по банковскому кредиту без возможности реструктуризации), человек переходит на более дорогостоящее микрокредитование. К сожалению, шансы на погашение кредита при таких обстоятельствах относительно низки, и большинство случаев невыплат по кредитам возникают именно в сделках такого рода. Одним из наиболее часто встречающихся микрозаймов этого типа является микрозайм «до зарплаты». Под ними понимают микрозаймы, выданные физическим лицам в размере не более 30 тыс. руб. на срок до одного месяца.

[4] Федеральный закон от 02.07.2010 г. № 151-ФЗ (ред. от 27.12.2018 г.) «О микрофинансовой деятельности и микрофинансовых организациях» (с изм. и доп., вступ. в силу с 28.01.2019 г.).

Следует отметить, что размер рынка подобных услуг достаточно велик, чтобы оказывать влияние на экономическую ситуацию в стране. Так, согласно данным РБК и отчету рейтингового агентства Эксперт РА[5], портфель микрозаймов по итогам 2018 г. превысил 150 млрд руб. На текущий момент число участников этого рынка равно примерно 2300 организаций.

Микрофинансовые организации осуществляют деятельность в тех сегментах рынка финансовых услуг, которые по тем или иным причинам не привлекают внимание банков, например, в населенных пунктах с небольшой численностью населения (до 50 тыс. человек), в которых, по мнению некоторых экспертов, сложно обеспечить высокую рентабельность классических банков.

Основные проблемы этого вида финансового бизнеса можно свести в три группы.

- Источники финансирования. Для МФО закон ограничивает возможность привлечения депозитов. Более 50% пассивов МФО составляют банковские кредиты, прочих инвесторов привлечь очень сложно.
- Нормативная база. Начиная с 1 сентября 2013 г. МФО находятся в ведении ЦБ РФ, однако нормативно-правовая база регулирования этого вида деятельности нуждается в развитии.
- Клиентская база. Уровень финансовой грамотности и, в частности, финансовой дисциплины рядовых клиентов МФО невелик.

Все это снижает возможности МФО технического и организационного развития, предъявляет повышенные требования к работе с потенциальными заемщиками и стимулирует использование дополнительных данных из бесплатных и/или низко затратных источников для решения аналитических проблем.

3.2. Использование социальных сетей для построения скоринговых моделей

Как правило, в скоринговых моделях используются характеристики источников дохода и социальной стабильности клиента. В условиях МФО данные такого рода можно получить только из анкеты, заполняемой потенциальным заемщиком. На их основательную проверку у организации, как правило, нет достаточно средств и времени – принятие решения о выдаче займа должно осуществляться быстро, это привлекает клиентов. В то же время оценка потенциального заемщика по неподтвержденным данным из анкеты может быть не надежной. В этом случае можно воспользоваться открытыми данными, характеризующими возможного клиента. К ним, в частности, относится его профиль в социальной сети.

Понятие социальной сети как группы связанных и коммуницирующих друг с другом людей возникло задолго до появления Интернета. Однако именно специализированные интернет-ресурсы, относительно недавно созданные для формирования социальных сетей, позволяют в одночасье собрать огромную базу социально-демографических данных, полезных для решения различных аналитических задач, – в маркетинговых исследованиях, при подборе персонала, в криминалистике и т.д. В частности, из профиля участ-

[5] Рынок микрофинансирования по итогам 2018 г.: адаптивная стратегия (https://marketing.rbc.ru/research/39330/).

ника социальной сети может быть получена как прямая, так и косвенная информация о нем. К прямой относятся пол, место и год рождения или возраст, место проживания, семейное положение, уровень образования и т.д. Косвенная помогает опосредованно оценить личность: закрытость профиля, количество выложенных постов и фотографий, количество друзей и подписчиков (социальный капитал). Даже само наличие профиля в социальной сети является характерным признаком публичности личности и уровня ее социализации.

4. Описание данных

Для анализа были взяты данные о клиентах реальной МФО, получивших одобрение на заем. В выборке присутствуют как клиенты, вернувшие его вовремя, так и вернувшие с просрочкой и не вернувшие его вообще. Из анкеты, заполненной при обращении в МФО, а также из ее учетной информационной системы были получены следующие данные о клиентах.

1. Договор:
– шифр клиента;
– имя;
– дата рождения;
– серия паспорта;
– кем выдан паспорт;
– адрес места жительства (по паспорту);
– валюта ссуды;
– дата планируемого возврата;
– размер ссуды;
– дата выдачи.

2. Данные анкеты:
– зарплата после НДФЛ;
– место работы (в свободной форме);
– количество иждивенцев.

3. Факты просрочки по договору:
– номер договора;
– дата погашения просрочки (null – если просрочка не погашена);
– дата переноса займа на просрочку.

Всего было получено 2332 наблюдения. Далее была проведена проверка корректности полученных значений для отсева сомнительных или недостоверных по следующим правилам (признаки достоверных данных):

- сумма займа – не менее 1000 руб.;
- зарплата не более 100 тыс. руб., удаление значений, превышающих этот уровень;
- расчетная дата раньше даты взятия ссуды, строка исключается из-за ошибки данных;
- срок ссуды меньше 5 дней – такие данные исключаются как недостоверные.

На основании полученных данных были рассчитаны следующие показатели:

- возраст – на основании поля «дата рождения»;

- количество просрочек, срок просрочек, средние показатели просрочки – на основании полей «дата погашения просрочки (null – если просрочка не погашена)», «дата переноса займа на просрочку»;
- показатель просрочки: 0 – просрочек не было, 1 – была хотя бы одна просрочка менее трех дней, 2 – есть хотя бы одна просрочка более 30 дней.

Далее к полученной базе добавляются данные из социальных сетей «ВКонтакте» и «Одноклассники».

Данные из сети «ВКонтакте»:

- давность посещения – дата последнего визита не менее полугода;
- заполненность профиля, расчетный показатель.

Данные из сети «Одноклассники»: давность посещения – дата последнего визита не менее полугода.

Список показателей и их основные статистические характеристики приведены в Приложении. Во всех построенных моделях в качестве зависимой переменной использовалась переменная «overdue», остальные переменные выступали в качестве независимых переменных.

5. Экспериментальная проверка влияния «проклятия размерности» на качество прогнозирования

Как было отмечено выше, в задачах кредитного скоринга МФО значительная часть факторов, характеризующих потенциального заемщика, измеряется по качественным шкалам – номинальным или порядковым. В контексте данного исследования к характеристикам такого типа относятся пол (два уровня), место работы (8 уровней) и давность посещения социальной сети «ВКонтакте» (два уровня). При этом, наиболее сложный фактор – место работы, предположительно может существенно повысить качество классификации потенциального клиента, и упрощать его (укрупнять уровни) нежелательно. Таким образом, при бинаризации этих факторов, т.е. превращении их в системы фиктивных переменных, мы должны будем добавить в спецификацию модели девять переменных вместо трех (с учетом предотвращения мультиколлинеарности). К сожалению, для большинства реализаций логистической регрессии это единственно возможный вариант. В то же время алгоритмы CART и C4.5 позволяют использовать как качественные, так и количественные переменные.

В рамках данного исследования было проведено две группы экспериментов. В рамках первой группы экспериментов мы сопоставили прогностические возможности выбранных методов классификации для двух вариантов спецификации – без бинаризации и с ее использованием, за исключением логистической регрессии. Для сравнения использовались ROC-кривые и площади под ними. В рамках второй группы экспериментов мы сопоставили результаты использования лучших вариантов спецификаций моделей дескриптивной аналитики с результатами использования алгоритма «случайного леса» с использованием алгоритма кросс-валидации (*cross-validation*) и ROC-кривых. Остановимся на результатах подробнее.

5.1. Краткое описание инструментов анализа качества классификации

Кратко остановимся на определении понятий ROC-кривая и таблица результатов классификации.

В результате обучения алгоритму классификации на обучающем множестве возникает классификация его объектов выбранным алгоритмом. Она может совпадать или не совпадать с истинным классом объектов в тренировочном множестве.

Таблица 1.

Классификация результатов

Истинный класс	Прогноз	
	дефолт был	дефолта не было
Дефолт был	TP	FN
Дефолта не было	FP	TN

На основе табл. 1 определяются следующие характеристики качества классификатора.

- Общая погрешность классификации (*Error rate*)

$$ER = \frac{FP + FN}{\Sigma}, \Sigma = TP + FN + FP + TN.$$

- Чувствительность (*Sensitivity*) $Sen = \frac{TP}{TP + FN}$ – способность классификатора правильно классифицировать наличие потенциального кредитного дефолта (*Positive*).

- Специфичность (Specificity) $Spe = \frac{TN}{TN + FP}$ – способность классификатора правильно классифицировать отсутствие потенциального кредитного дефолта (*Negative)*. Может рассматриваться как «чувствительность» для данного класса. В связи с этим иногда используется общий термин Recall.

- Точность (*Precision*) $\Pr ecision(yes) = \frac{TP}{TP + FP}$ – доля правильной классификации при отнесении клиента к дефолтным.

В связи со специфичностью иногда рассматривают *1-Spe* – склонность классификатора принимать *Negative* за *Positive*. Можно сказать, что это аналог вероятности ошибки второго рода в теории проверки статистических гипотез. Полезной характеристикой классификатора является соотношение между *Sen* и *1-Spe* . Очевидно, что первая величина должна превышать вторую. В случае, если они равны, классификация напоминает процесс подбрасывания идеальной монеты, т.е. просто угадывание. Сам по себе процесс бинарной классификации, как правило, основан на сопоставлении доли класса *Positive* в некотором множестве объектов (деревья принятия решений) или его вероятности для конкретного объекта (логистическая регрессия) с некоторым пороговым значением. При превышении его все объекты данного множества или отдельный объект классифициру-

ются как *Positive*. Величина порога – важный параметр классификатора. Его варьирование приводит к появлению множества вариантов классификатора с разными прогностическими способностями. График зависимости чувствительности от *1-Spe* носит название ROC-кривой. Очевидно, что площадь под ней может рассматриваться как характеристика прогностических способностей семейства классификаторов, которые отличаются величиной порога для доминирующего класса [Hosmer et al., 2001].

Техника кросс-валидации возникла как реакция на необходимость использования в исследовании одного множества объектов без возможности получить новые данные о них в новых условиях или данные о новых объектах. Эта ситуация нередко складывается в тех отраслях знаний, где затруднено или невозможно проведение планируемых экспериментов, например в эконометрике. В этом случае эффективной стратегий проверки качества алгоритма классификации является разбиение случайным образом исходного множества данных на несколько одинаковых по величине подмножеств (карманов) и выделение одного их них в качестве тестового множества. Остальные карманы включаются в тренировочное множество. Процедура повторяется перебором карманов в качестве тестовых множеств. Таким образом, например, при десяти карманах мы получим десять оценок качества классификации, что даст нам возможность подсчитать среднюю оценку качества и оценить его разброс. Эту процедуру можно повторять многократно, каждый раз по-новому выполняя разделение на карманы.

5.2. Проверка качества классификации с использованием ROC-кривых

Прежде всего отметим, что некоторые из выбранных алгоритмов классификации имеют настраиваемые параметры, значения которых потенциально могут повлиять на качество их работы. В данном исследовании мы обратили внимание на два показателя, доступных в имеющейся у нас реализации данных алгоритмов. Во-первых, минимальное количество объектов в листе дерева. Этот показатель использовался как критерий останова процесса обучения классификатора, как для CART, так и для C4.5. Во-вторых, доля исходного множества, которая выделяется для реализации процедуры обрезки дерева в алгоритме CART. Мы сопоставили качество классификации для различных значений указанных параметров и пришли к выводу, что минимальное количество объектов в листе не влияет на качество классификатора. От этого показателя существенно зависит глубина дерева. В итоге мы остановились на величине 50 объектов. Варьирование второго показателя показало наличие зависимости качества от его значения. Приведем график нескольких ROC-кривых.

На рис. 1 мы наблюдаем нелинейную связь между значением данного показателя и качеством. Кривая Score 8 соответствует доли 33% (AUC=0,6192), а кривая Score 11 – 15% (AUC=0,6617). В то же время кривые Score 9 и Score 10 соответствуют долям 25% (AUC=0,7195) и 20% (AUC=0,7199). Видно, что большие и маленькие значения этого показателя приводят к снижению качества классификации, а средние значения около 20% дают малоразличимые удовлетворительные результаты. В результате мы остановились на величине 20%.

Результаты анализа влияния спецификации на качество классификации приведено на рис. 2.

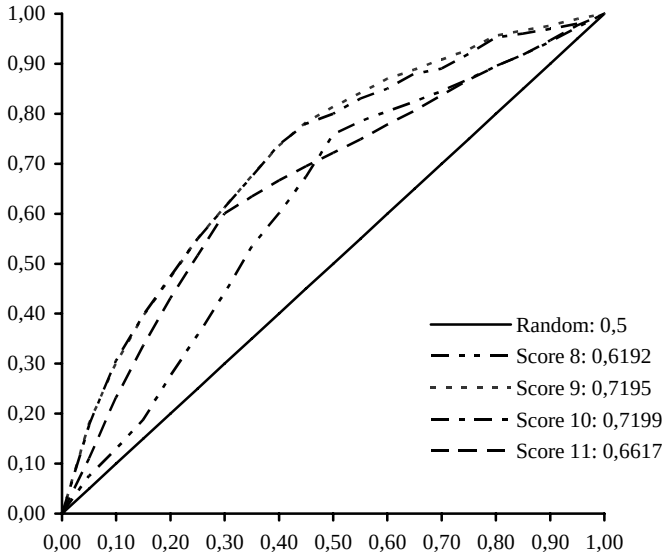

Рис. 1. Влияние значения параметра «prunning set size» (PSS) на качество работы CART. Random соответствует угадыванию

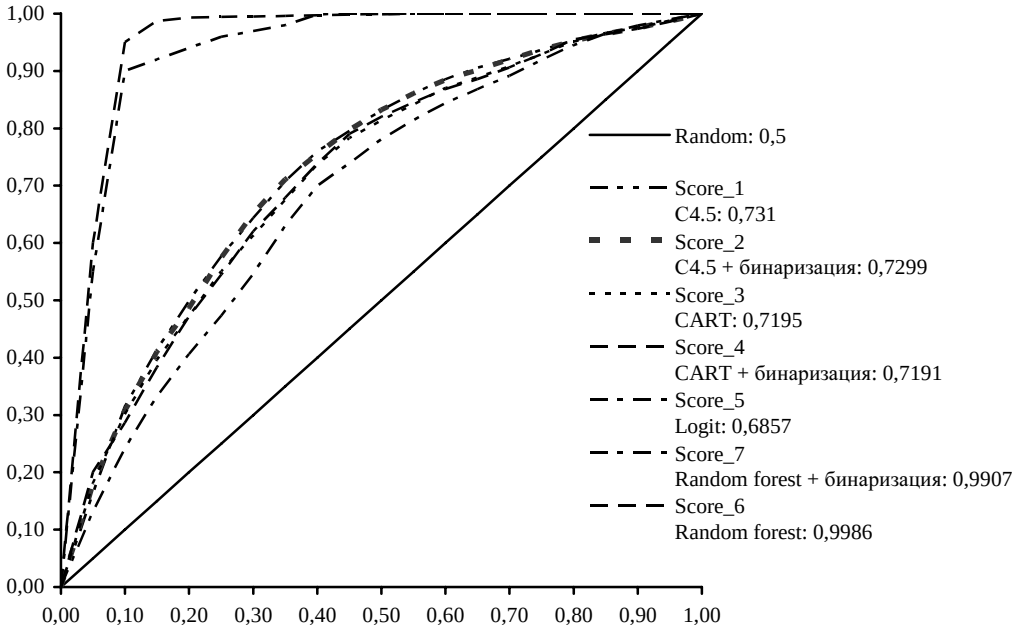

Рис. 2. Влияние спецификации модели на качество классификации. Random соответствует угадыванию

Сразу отметим, что наихудший результат (рис. 2) был получен для единственно возможной спецификации логистической регрессии – Score 5 (AUC=,6857). Кривые Score 1 (AUC=0,7310) и Score 2 (AUC=0,7299) соответствуют алгоритму C4.5 (без бинаризации и с бинаризацией) и они достаточно близки друг другу. Кривые Score 3 (AUC=0,7195) и Score 4 (AUC=0,7191) соответствуют алгоритму CART. Результаты также мало зависят от спецификации. Кривые Score 6 (AUC=0,9986) и Score 7 (AUC=0,9907) соответствуют алгоритму случайного леса. В итоге можно сделать следующие выводы.

1. Наиболее качественную классификацию обеспечивает, как того и следовало ожидать, прогностический алгоритм «случайного леса». На втором месте оказался алгоритм C4.5, несколько хуже результаты для CART и, наконец, наихудшее качество продемонстрировала логистическая регрессия.

2. Спецификация оказала очень незначительное влияние на качество классификации дескриптивных алгоритмов и оказалась существенной для прогностического алгоритма.

5.3. Оценка качества классификации методом кросс-валидации

Кросс-валидация позволяет проанализировать чувствительность классификатора к изменению (удалению и добавлению части) данных. В рамках этой части исследований использовались спецификации, которые обладали наилучшими характеристиками в предыдущей части исследования. Для проведения эксперимента была выбрана кросс-валидация с десятью карманами. Осуществлялось десять повторов этой процедуры для каждого алгоритма. Результатом работы данной процедуры была средняя величина показателя «Общая погрешность классификации» (*Error rate*) для каждой из попыток. Итог исследования представлен в табл. 2 и 3. Следует отметить, что метод кросс-валидации является развитием традиционного деления выборки на обучающее и тестовое множества. Он обладает существенно бóльшими возможностями для анализа качества функционирования тех или иных методов моделирования статистических взаимосвязей за счет накопления информации в ходе повторяющихся экспериментов с данными. В табл. 3 приведены средние значения показателей качества, полученные в результате множества повторяющихся на различных подвыборках процедур оценивания и классификации.

Таблица 2.

Значения показателя Error Rate в экспериментах кросс-валидации

	C4.5	CART	Логистическая регрессия	Случайный лес
Минимальное	0,3408	0,3524	0,3524	0,0000
Максимальное	0,3614	0,3708	0,3601	0,0004
Среднее	0,3493	0,3609	0,3559	$<10^{-4}$

Таблица 3.

Оценка качества классификации для каждого из алгоритмов

Оценка качества			Таблица ошибок			
значение	чувствительность	1-точность	истинный класс/ предсказанный класс	да	нет	сумма
Алгоритм C4.5						
Да	0,7810	0,3372	Да	10312	2892	13204
Нет	0,4804	0,3735	Нет	5246	4850	10096
			Сумма	15558	7742	23300
Алгоритм CART						
Да	0,6472	0,3050	Да	8546	4658	13204
Нет	0,6285	0,4233	Нет	3751	6345	10096
			Сумма	12297	11003	23300
Логистическая регрессия						
Да	0,8397	0,3578	Да	11087	2117	13204
Нет	0,3883	0,3507	Нет	6176	3920	10096
			Сумма	17263	6037	23300
Случайный лес						
Да	1,0000	0,0001	Да	13204	0	13204
Нет	0,9999	0,0000	Нет	1	10095	10096
			Сумма	13205	10095	23300

Из табл. 2 видно, что наивысшее качество и стабильность классификации обеспечивает алгоритм «случайного леса». Качество дескриптивных алгоритмов позволяет поставить на первое место алгоритм C4.5, на втором месте оказывается логистическая регрессия, а алгоритм CART занимает третье место. Таблица 3 позволяет сопоставить другие характеристики качества указанных алгоритмов:

$$\begin{cases} Sensitivity(yes) = \dfrac{TP}{TP+FN} - \text{доля правильной классификации дефолта среди дефолтных клиентов,} \\ 1 - \Pr ecision(yes) = \dfrac{TP}{TP+FP} - \text{доля ошибочной классификации при отнесении клиентов к дефолтным.} \end{cases}$$

Для отсутствия дефолта («нет») аналогично. Показатель Recall характеризует предсказуемость соответствующего класса, а показатель 1-Precision – надежность классификации.

Для логистической регрессии и C4.4 наблюдается некоторая асимметрия – дефолт классифицируется лучше, чем его отсутствие, однако надежность классификации примерно одинаковая. Для алгоритма CART оба уровня прогнозируются в равной степени хорошо, но наличие дефолта прогнозируется более надежно. В итоге с точки зрения надежности прогноза дефолта этот алгоритм предпочтительнее.

Таким образом, если исследователя интересует общий уровень ошибочной классификации среди дескриптивных алгоритмов, следует использовать C4.5, но если надежность прогноза дефолта (т.е. отнесения клиента к дефолтным) более важна, следует отдать предпочтение CART. Следует отметить, что для практики, возможно, более важна надежность отнесения клиента к недефолтным. Ошибка здесь состоит в том, что на самом деле он склонен к дефолту. Потери в этом случае могут быть высокие. Наилучшее значение этого показателя дает логистическая регрессия.

6. Описание потенциально дефолтных клиентов на основе дескриптивных алгоритмов

Одной из задач данного исследования является оценка важности характеристик потенциального заемщика для его классификации в рамках принятия решения о выдаче микрокредита. Термин «важность переменной» не имеет однозначного определения и, как отмечено в работе [Berk, 2008], его следует интерпретировать в контексте конкретного метода оценивания важности, в данном случае – в контексте конкретного метода классификации. В частности, для алгоритмов CART, C4.5 и Random Forest к важным факторам в рамках данного исследования относились факторы, отобранные алгоритмами для осуществления классификации. Числовые характеристики степени важности не учитывались в силу различий методов их расчетов. Для логистической регрессии к важным факторам относились статистически значимые факторы по результатам оценивания на всей выборке.

Существенным преимуществом дескриптивных алгоритмов перед прогностическими является возможность получить описания классов в терминах учтенных в модели факторов. При этом такие алгоритмы как C4.5 и CART самостоятельно отбирают наиболее важные для классификации показатели. Не исключено, что при тщательной работе со спецификацией логистической регрессии можно было бы добиться нехудших результатов, но это требует много времени и большого практического опыта. В данном разделе мы сопоставим описания классов, которое дают алгоритмы C4.5 и CART, и проверим гипотезы о признаках потенциально дефолтных клиентов, сформулированных выше.

Гипотеза о важности данных из социальных сетей для прогнозирования склонности к дефолту не подтвердилась ни для одного метода классификации, что может быть связано с небольшим разбросом в собранных значениях – 91% всех имеющих профиль клиентов недавно посещали свою страницу в ВК, а профиль в среднем заполнен лишь на 14%. Количество друзей в социальных сетях также не является важным фактором в указанном смысле, так как за исключением нескольких выбросов эти сведения имеют небольшой разброс.

Наиболее важными признаками наличия высоких рисков являются характеристики кредитной истории – общее количество заявок на ссуду, срок и размер ссуды, а также возраст заемщика.

Показатели зарплаты, количество иждивенцев, пол не вошли в признаки классификации. Часто указанные самими заемщиками данные по зарплатам завышены и неадекватны, их нельзя использовать для оценки их финансового состояния.

Всего в выборке было представлено 47%. Пол не оказывает существенного влияния на результаты просрочки, так как примерно равное количество мужчин и женщин берут микрозайм и примерно одинаковое их количество оказывается дефолтным. И хотя причины просрочек у мужчин и женщин различны, склонность к дефолту (на основании анализа выборки) просрочки больше зависит от кредитной истории.

Распределение наличия дефолтных просрочек по полу также подтверждает отсутствие значимой связи просрочки с полом.

Таблица 4.

Распределение наличия дефолтных просрочек по полу, %

	Дефолтной просрочки не было	Дефолтная просрочка была
Мужчины	66	34
Женщины	63	37

Также дополнительно проверена и не отвергнута гипотеза об отсутствии статистической связи между полом и наличием кредитного дефолта – она не отвергается на пятипроцентном уровне значимости.

6.1. Алгоритм C4.5

Результаты классификации по всей выборке с использованием бинаризации представлены в табл. 5.

Таблица 5.

Оценка качества классификации алгоритма C4.5

Доля ошибок			0,3100			
Оценка качества			Таблица ошибок (confusion matrix)			
значение	чувствительность	1-точность	истинный класс/ предсказанный класс	да	нет	сумма
Да	0,7926	0,3001	Да	1047	274	1321
Нет	0,5559	0,3278	Нет	449	562	1011
			Сумма	1496	836	2332

- time < 30,5000
  - req_num < 1,5000
    - time < 8,5000 then overdue = **no** (57,69 % of 52 examples)
    - time >= 8,5000 then overdue = **yes** (71,18 % of 628 examples)
  - req_num >= 1,5000
    - req_num < 5,5000
      - loan < 7500,0000
        - age < 32,5000
          - loan < 4500,0000 then overdue = **no** (60,00 % of 50 examples)
          - loan >= 4500,0000
            - age < 26,5000 then overdue = **no** (53,23 % of 62 examples)
            - age >= 26,5000 then overdue = **yes** (62,79 % of 86 examples)
        - age >= 32,5000
          - time < 29,5000
            - time < 14,5000 then overdue = **no** (71,83 % of 71 examples)
            - time >= 14,5000 then overdue = **yes** (56,25 % of 80 examples)
          - time >= 29,5000 then overdue = **no** (70,67 % of 225 examples)
      - loan >= 7500,0000
        - work in [OOO]
          - age < 33,5000 then overdue = **yes** (71,70 % of 53 examples)
          - age >= 33,5000
            - experience < 9,5000 then overdue = **no** (53,85 % of 52 examples)
            - experience >= 9,5000 then overdue = **yes** (61,40 % of 114 examples)
        - work in [Other] then overdue = **no** (51,28 % of 39 examples)
        - work in [Pensioner] then overdue = **yes** (57,14 % of 14 examples)
        - work in [Ind] then overdue = **no** (60,71 % of 56 examples)
        - work in [Absent] then overdue = **yes** (55,56 % of 9 examples)
        - work in [Budget] then overdue = **yes** (53,49 % of 86 examples)
        - work in [MVD/Med] then overdue = **yes** (52,08 % of 48 examples)
        - work in [Bank] then overdue = **no** (87,50 % of 8 examples)
    - req_num >= 5,5000 then overdue = **no** (76,92 % of 221 examples)
- time >= 30,5000 then overdue = **yes** (81,75 % of 378 examples)

Согласно алгоритму C4.5 выявлено 10 классов дефолтных клиентов. 30% всех дефолтных клиентов – это взявшие ссуду на срок от 8 до 30 дней при первой заявке. Это подтверждает гипотезу о важности кредитной истории и необходимости создания регулятором общего бюро кредитных историй. Среди людей, обратившихся за ссудой до 30 дней, уже ранее получавших одобрение по заявкам, и взявших ссуду более 7500 руб., большой процент дефолтных клиентов, работающих в ООО, а также в бюджетных организациях, в сфере МВД, медицины и других социальных секторах экономики. Еще примерно 20% всех дефолтных клиентов – взявшие ссуду более 30 дней назад. Отметим классы надежных заемщиков – 20% надежных – это заемщики, взявшие ссуду до 30 дней, имеющие более пяти одобренных заявок (77% всех в этом классе не имеют просрочки). Также к надежным относятся клиенты, уже имеющие как минимум одну одобренную заявку ранее, возраста более 32 лет, взявшие на 30 дней, что является косвенным признаком цели «кредита до зарплаты».

6.2. Алгоритм CART

Результаты классификации по всей выборке с использованием бинаризации.

Таблица 6.

Оценка качества классификации алгоритма CART

Доля ошибок			0,3169			
Оценка качества			Таблица ошибок (confusion matrix)			
значение	чувствительность	1-точность	истинный класс/ предсказанный класс	да	нет	сумма
Да	0,7653	0,2979	Да	1011	310	1321
Нет	0,5757	0,3475	Нет	429	582	1011
			Сумма	1440	892	2332

- req_num < 2,5000
  - time < 30,5000
    - req_num < 1,5000 then overdue = **yes** (69,63 % of 540 examples)
    - req_num >= 1,5000
      - loan < 7500,0000
        - age < 41,5000
          - experience < 14,0000 then overdue = **yes** (62,71 % of 59 examples)
          - experience >= 14,0000 then overdue = **no** (53,76 % of 93 examples)
        - age >= 41,5000 then overdue = **no** (66,22 % of 74 examples)
      - loan >= 7500,0000

        - work in [OOO,Other,Pensioner,Ind,Budget,MVD/Med] then overdue = **yes** (64,93 % of 66 examples)
        - work in [Absent,Bank] then overdue = **no** (100,00 % of 3 examples)
    - time >= 30,5000 then overdue = **yes** (85,96 % of 235 examples)
- req_num >= 2,5000
    - time < 30,5000
        - req_num < 5,5000
            - loan < 7500,0000 then overdue = **no** (65,15 % of 241 examples)
            - loan >= 7500,0000
                - work in [OOO,Absent] then overdue = **yes** (64,04 % of 109 examples)
                - work in [Other,Pensioner,Ind,Budget,MVD/Med,Bank] then overdue = **no** (60,90 % of 17 examples)
        - req_num >= 5,5000 then overdue = **no** (75,47 % of 159 examples)
    - time >= 30,5000 then overdue = **yes** (71,05 % of 76 examples)

В алгоритме CART выделилось 5 классов дефолтных клиентов. Так же как и в алгоритме C4.5, важное значение имеют ранее одобренные заявки – почти половина всех дефолтных заемщиков (45%) имеют не более одной ранее одобренной заявки. Высокая доля (70%) дефолтных клиентов среди имеющих не более одной одобренной заявки и взявших ссуду до 30 дней. Очень высока доля дефолтных заемщиков среди имеющих одну одобренную ранее заявку и взявших кредит на срок более 30 дней – 86%. К надежным могут быть отнесены заемщики, имеющие более пяти одобренных ранее заявок (что согласуется с данными предыдущего алгоритма). Также надежными являются заемщики с одной или двумя ранее одобренными ссудами и суммой до 7500 руб. Данная сумма также соответствует сумме, обнаруженной в алгоритме C4.5.

6.3. Логистическая регрессия

Таблица 7.

Оценка качества классификации логистической регрессии

Доля ошибок			0,3465			
Оценка качества			Таблица ошибок (confusion matrix)			
значение	чувствительность	1-точность	истинный класс/ предсказанный класс	да	нет	сумма
Да	0,8501	0,3520	Да	1123	198	1321
Нет	0,3966	0,3306	Нет	610	401	1011
			Сумма	1733	599	2332

Таблица 8.

Оценка модели логистической регрессии

Фактор	Оценка коэффициента	Стандартная ошибка оценки	Статистика Вальда	*p*-уровень
constant	0,987993	0,3517	7,8937	0,0050
age	–0,014748	0,0044	11,1150	0,0009
wage	0,000000	0,0000	0,0011	0,9733
time	0,011394	0,0034	11,0899	0,0009
loan	0,000045	0,0000	9,5933	0,0020
depend	–0,078119	0,0959	0,6635	0,4153
req_num	–0,325909	0,0265	151,1044	0,0000
friends_ok	0,000382	0,0008	0,2364	0,6268
followers	–0,001635	0,0033	0,2471	0,6191
profile_ok	0,052977	0,1421	0,1389	0,7093
profile_vk	0,222370	0,2623	0,7189	0,3965
experience	0,001883	0,0008	5,6929	0,0170
gender_m_2	0,010280	0,0932	0,0122	0,9121
work_OOO_2	0,129490	0,2615	0,2453	0,6204
work_Other_2	–0,160415	0,2866	0,3132	0,5757
work_Pensioner_2	0,045300	0,3430	0,0174	0,8949
work_Ind_2	–0,189014	0,2844	0,4418	0,5062
work_Absent_2	0,066746	0,3684	0,0328	0,8562
work_Budget_2	0,102991	0,2822	0,1332	0,7151
work_MVD/Med_2	–0,148982	0,2917	0,2609	0,6095
time_vk_	0,022093	0,1605	0,0190	0,8905

В соответствии с введенным выше определением важности факторов для классификации в рамках данного исследования можно отметить, что множество факторов, статистически значимо влияющих на вероятность кредитного дефолта, пересекаются со множеством важных для классификации факторов для предыдущих двух алгоритмов. Все значимые факторы выбраны хотя бы одним из алгоритмов для классификации. Направление влияния также согласуется с выявленными выше закономерностями. В частности рост параметра «срок ссуды» (time) согласуется с ростом доли числа дефолтных клиентов. Фактор «размер ссуды в руб.» положительно связан с вероятностью кредитного дефолта, что согласуется с CART и C4.5. Также отрицательное влияние на вероятность кредитного

дефолта фактора «Общее количество заявок на ссуду» соответствует результатам работы этих алгоритмов. Влияние фактора «стаж в месяцах» согласуется с результатами алгоритма C4.5.

7. Заключение

В результате проделанной в исследовании работы мы пришли к следующим практически значимым выводам.

Во-первых, бинаризация качественных показателей, приводящая к росту размерности задачи, оказывает незначительное влияние на качество дескриптивных алгоритмов C4.5 и CART. В то же время это влияние заметно для прогностического алгоритма «случайный лес». Бинаризация снижает его качество.

Во-вторых, с точки зрения соотношения доли правильных и ложных прогнозов при разных порогах отнесения клиента к дефолтным (показатель AUC) наиболее качественным оказывается алгоритм C4.5. Он же обладает наименьшей чувствительностью к изменению данных. Однако если интерес исследователя не сводится к минимизации общего уровня ошибок, то выбор алгоритма классификации не столь очевиден и в ряде случаев логистическая регрессия может оказаться предпочтительней.

В-третьих, представляют интерес результаты проверки гипотез о признаках дефолтных клиентов.

Для характеристик финансовой и социальной стабильности клиентов были привлечены данные из социальных сетей. Оказалось, что они не являются важными для классификации факторами. Возможно, следует пересмотреть формирование признаков давности посещения, учесть распределение числа друзей и подписчиков для включения в модель.

Маловажным признаком для классификации клиентов является место работы. Это может быть связано с ненадежностью указанных данных самими клиентами. Большую склонность к дефолту проявляют указавшие государственные, бюджетные организации и ООО. Работники ИП в меньшей степени склонны к дефолту. Возможно, руководству МФО стоит обратить внимание на разработку специальных условий и программ для самозанятых.

Возраст относительно важен для классификации. Люди до 32 лет (по алгоритму C4.5) или до 40 лет (по алгоритму CART) чаще склонны к невозвратам. Однако связь классификации с возрастом нелинейна – пенсионеры, не имеющие дохода, также попадают в число ненадежных.

Важным признаком является количество ранее одобренных заявок, клиенты с пятью и более ранее одобренными заявками редко становятся дефолтными клиентами (это подтвердилось обоими алгоритмами).

Безопасная сумма ссуды – до 7500 руб., большая часть надежных клиентов – взявшие до 7500 руб. на срок до 30 дней.

Данные по полу и зарплатам не вошли в признаки классификации, таким образом, гендерные модели скоринга в МФО не требуются, а финансовые данные, указываемые заемщиком без подтверждения соответствующими документами, нельзя использовать для оценки финансового состояния. Требуется привлекать другие методы оценки его дохода, например, наличие в собственности автомобиля или земельного участка.

Алгоритмы C4.5 и CART выявили одинаковые закономерности от количества одобренных ранее заявок (более пяти, более двух), от суммы (до и более 7500 руб.), от места работы (ООО). При этом более однородные узлы получены с помощью алгоритма C4.5.

В качестве дальнейшего развития стоит рассмотреть другую классификацию информации из соцсетей для обогащения данных, использовать косвенные признаки оценки финансового состояния вместо указанной в анкете зарплаты.

С практической точки зрения необходимо отметить, что полученные в данном исследовании результаты являются аргументом в пользу необходимости создания единой доступной базы кредитных историй. Это, как следует из результатов использования всех построенных моделей, наиболее результативный способ снизить риск невозврата, позволяющий предиктивно определить самых надежных и ненадежных заемщиков.

Выявленная чувствительность методов машинного обучения к бинаризации качественных показателей означает, что при разработке анкет для оценки потенциальных заемщиков стоит уделять особое внимание закрытым вопросам с небольшим числом закрытий (вариантов ответа). Также была выявлена необходимость более тщательной формулировки вопросов о цели займа и месте работы для повышения значимости этих факторов в моделях скоринга.

Приложение

Таблица П1.

Статистические характеристики количественных показателей

Показатель	Название в модели	Минимум	Максимум	Среднее	Стандартное отклонение	Вариация
Возраст (полных лет)	age	20	73	39,7757	11,6440	0,2927
Зарплата (0 – не указана или не работает)	wage	0	100000	47823	23563	0,4927
Срок ссуды в днях	time	5	214	27,6	15,4	0,5603
Размер ссуды в руб.	loan	2000	30000	7448,9708	3688,3021	0,4951
Количество иждивенцев	depend	0	1	0,3431	0,4748	1,3841
Общее количество заявок на ссуду	req_num	1	11	2,7140	2,0894	0,7699
Количество друзей в соц. сети «Одноклассники»	friends_ok	0	818	21,3259	66,2690	3,1074
Количество подписчиков в соц. сети «Одноклассники»	followers	0	456	2,9065	14,4250	4,9630

Окончание табл. П1.

Показатель	Название в модели	Минимум	Максимум	Среднее	Стандартное отклонение	Вариация
Заполненость профиля в соц. сети «Одноклассники» (1 – более половины, 0 – менее половины или нет)	profile_ok	0	1	0,1402	0,3473	2,4767
Заполненость профиля в соц. сети «ВКонтакте» (1 – более половины, 0 – менее половины или нет)	profile_vk	0	1	0,0334	0,1798	5,3768
Стаж в месяцах	experience	0	576	36,5926	60,1629	1,6441

Таблица П2.

Статистические характеристики качественных факторов

Показатель	Название	Джини	Значение	Количество	Доля	Распределение
Пол (M – мужской, F – женский)	gender	0,4993	M	1211	51,93%	
			F	1121	48,07%	
Место работы	work	0,7531	OOO	1020	43,74%	
			Other	244	10,46%	
			Pensioner	108	4,63%	
			Ind	258	11,06%	
			Absent	77	3,30%	
			Budget	314	13,46%	
			MVD/Med	231	9,91%	
			Bank	80	3,43%	
Давность посещения соц. сети «ВКонтакте»	time_vk	0,1568	<6 months	2132	91,42%	
			>6 months	200	8,58%	
Наличие просрочек	overdue	0,4912	Yes	1321	56,65%	
			No	1011	43,35%	

* *
*

СПИСОК ЛИТЕРАТУРЫ

Криворучко С.В., Абрамова М.А., Мамута М.В., Тенетник О.С., Шакер И.Е. Микрофинансирование в России. М.: Кнорус: ЦИПСиР, 2013.

Руководство по кредитному скорингу / под ред. Э. Мэйз; пер. с англ. И.М. Тикота; науч. ред. Д.И. Вороненко. Минск: Гревцов Букс, 2008.

Снегова Е.Г. Применение метода логистической регрессии для прогнозирования вероятности дефолта при экспресс-кредитовании // Национальные интересы: приоритеты и безопасность. 2013. Т. 9. Вып. 5.

Сорокин А.С. Построение скоринговых карт с использованием модели логистической регрессии // Интернет-журнал «Науковедение». 2014. № 2 (21). (http://naukovedenie.ru/PDF/180EVN214.pdf).

Софронова В.В. Оценка дефолта заемщика // Финансовая аналитика: проблемы и решения. 2016. Т. 9. Вып. 3. С. 39–48.

Уксусова М.С. Микрофинансирование: содержание, особенности, проблемы и перспективы развития // Экономический журнал. 2018. № 3(51). С. 50–66.

Цхададзе Н.В. Рынок микрофинансовых услуг как социально ориентированный бизнес: зарубежный опыт // Вестник экономической безопасности. 2016. № 4. С. 304–310.

Azen R., Budescu D.V. The Dominance Analysis Approach for Comparing Predictors in Multiple Regression // Psychological Methods. 2003. 8(2). P. 129–148. (doi: 10.1037/1082-989X.8.2.129).

Berk R. Statistical Learning from a Regression Perspective. New York, NY: Springer, 2008.

Breiman L., Friedman J., Olshen R.A., Stone C.J. Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.

Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer, 2001. (https://doi.org/10.1007/978-0-387-21606-5).

Hosmer D.W., Lemeshow S., Sturdivant R. Applied Logistic Regression. 3rd ed. Hoboken: John Wiley & Sons, 2013. (https://doi.org/10.1002/9781118548387).

Nathans L.L., Oswald F.L., Nimon K. Interpreting Multiple Linear Regression: A Guidebook of Variable Importance // Practical Assessment, Research and Evaluation. 2012. Vol. 17. № 9. P. 1–19.

Quinlan J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc., 1993.

A Risk Management Framework for Microfinance Institutions // Deutsche Gesellschaft fur Technische Zusammenarbeit (GTZ). 2000. (http://www.gtz.de).

Modeling the Probability of Credit Default of Clients of Microfinance Organizations: The Case of One MFI

Konstantin Polyakov[1], **Ludmila Zhukova**[2]

[1] National Research University Higher School of Economics,
20, Myasnitskaya ul., Moscow, 101000, Russian Federation.
E-mail: polyakov.kl@hse.ru

[2] National Research University Higher School of Economics,
20, Myasnitskaya ul., Moscow, 101000, Russian Federation.
E-mail: lvzhukova@hse.ru

Microfinance organizations have become widespread in the crisis years, issuing microloans (up to 100000 rubles) at high interest rates almost without documents. Today, the Central Bank of Russia actively regulates this market, more and more tightening requirements, limiting rates and pennies on loans. This necessitates the formation of a new strategy for assessing the risk of non-repayment of a loan or loan, based on the prevention of delinquency on the part of customers. To do this, first, it is necessary to obtain more informative data about customers, without complicating the relationship with them. Secondly, it is necessary to have a good understanding of the possibilities of certain methods of classification in solving various problems of evaluating potential customers. The authors of this study analyze the importance for the clients classification quality of those indicators that are traditionally collected by MFIs, as well as the importance of some new indicators based on data from social networks. In this case, the importance of indicators is interpreted in the context of specific classification algorithms (methods). To model credit default (delay of more than 30 days), the authors use several algorithms for constructing classification trees – CART and C 4.5 algorithms, logistic regression and Random Forest algorithm. Modeling is carried out based on a sample of customer profiles of real MFI. Ambiguous results were obtained. Depending on the formulation of the problem of classification of customers have advantage different algorithms descriptive analysis (CART, C4.5, Logit). At the same time, as you might expect, the non-interpreted predictive algorithm "Random Forest" provides the best quality of forecasts. According to the results of the analysis, it was revealed that the credit history of the borrower, as well as his age, is of great importance for the classification of MFI clients. Gender had no impact on the classification results. In addition, data from social networks turned out to be unimportant.

Key words: microfinance organization; default; classification tree; logistic regression; random forest.

JEL Classification: G21, C38.

* *
*

References

A Risk Management Framework for Microfinance Institutions (2000) *Deutsche Gesellschaft fur Technische Zusammenarbeit (GTZ)*. Available at: http://www.gtz.de

Azen R., Budescu D.V. (2003) The Dominance Analysis Approach for Comparing Predictors in Multiple Regression. *Psychological Methods*, 8(2), pp. 129–148. (doi: 10.1037/1082-989X.8.2.129).

Berk R. (2008) *Statistical Learning from a Regression Perspective*. New York, NY: Springer.

Breiman L., Friedman J., Olshen R.A., Stone C.J. (1984) *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Hastie T., Tibshirani R., Friedman J. (2001) *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer. Available at: https://doi.org/10.1007/978-0-387-21606-5

Hosmer D.W., Lemeshow S., Sturdivant R. (2013) *Applied Logistic Regression*. 3rd ed. Hoboken: John Wiley & Sons. Available at: https://doi.org/10.1002/9781118548387

Krivoruchko S.V., Abramova M.A., Mamuta M.V., Tenetnik O.S., Shaker I.E. (2013) *Mikrofinansirovanie v Rossii* [Microfinance in Russia]. Moscow: KnoRus: TsIPSiR Publ.

Nathans L.L., Oswald F.L., Nimon K. (2012) Interpreting Multiple Linear Regression: A Guidebook of Variable Importance. *Practical Assessment, Research and Evaluation*, 17, 9, pp. 1–19.

Quinlan J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc.

Rukovodstvo po kreditnomu skoringu (2008) [Handbook of Credit Scoring] (eds. E. Mays). Minsk: Grevtsov Books.

Snegova E.G. (2013) Primenenie metoda logisticheskoj regressii dlya prognozirovaniya veroyatnosti defolta pri ekspress-kreditovanii [Application of Logistic Regression Method for Forecasting of Probability of Default at Express Crediting]. *National Interests: Priorities and Security*, 9, iss. 5.

Sofronova V.V. (2016) Ocenka defolta zaemshchika [Assessment of the Borrower's Default]. *Financial Analytics: Science and Experience*, 9, iss. 3, pp. 39–48.

Sorokin A.S. (2014) Postroenie skoringovyh kart s ispol'zovaniem modeli logisticheskoj regressii [Building a Scorecard Using a Logistic Regression Model]. *Internet-Journal «Science of Science»*, 2(21). Available at: http://naukovedenie.ru/PDF/180EVN214.pdf

Tskhadadze N.V. (2016) Rynok mikrofinansovyh uslug kak social'no orientirovannyj biznes: zarubezhnyj opyt [The Market for Microfinance Services As a Socially Oriented Business: Foreign Experience]. *Vestnik of Economic Security*, 4, pp. 304–310.

Uksusova M.S. (2018) Mikrofinansirovanie: soderzhanie, osobennosti, problemy i perspektivy razvitiya [Microfinance: Content, Features, Problems and Prospects of Development]. *Economic Journal*, 3(51), pp. 50–66.