Современная практика разработки скоринговых карт для розничных клиентов в белорусских банках[1]

Олег ГИЧАН

Белорусский государственный университет, аспирант, Республика Беларусь, г. Минск, e-mail: ogichan@bk.ru

Екатерина ГОСПОДАРИК

Белорусский государственный университет, заведующий кафедрой аналитической экономики и эконометрики, кандидат экономических наук, доцент, Республика Беларусь, г. Минск, e-mail: gospodarik@bsu.by

УДК 330.43, 519.86

Ключевые слова:
кредитный риск; кредитный скоринг; логистическая регрессия.

Растущий спрос на кредитные продукты банков породил такую действенную методику автоматизированной оценки кредитоспособности заемщиков, как кредитный скоринг (от *англ.* credit scoring). Представляя собой математическую или статистическую модели, или ансамбль моделей, кредитный скоринг позволяет банковским организациям объективно и быстро оценивать кредитные риски. В отличие от других методик оценки кредитного риска (например, система экспертных оценок), кредитный скоринг использует быстроразвивающийся инструментарий новых технологий и машинного обучения, что позволяет таким банкам «быть в тренде», развивать клиентскую лояльность и максимизировать прибыль от текущей деятельности.

Кредитный скоринг применяется, как правило, в отношении физических лиц и субъектов малого бизнеса. Поскольку кредитные операции с данной группой заемщиков занимают существенное место в деятельности коммерческих банков, целью исследования является изучение такого инструмента риск-менеджмента, как кредитный скоринг.

Несмотря на долгую историю кредитования, история кредитного скоринга совсем незначительна. Кредитный скоринг как метод оценки кредитоспособности заемщиков используется не более 60 лет. «Пионерами» в практическом применении такого метода были американские банки, которые выдавали кредитные карты.

Кредитный скоринг представляет собой систему присвоения баллов кредитополучателю на основании его способности и потенциала выполнить взятые на себя обязательства. Баллы рассчитываются на основании имеющегося объема информации о клиенте с помощью статистической модели или математического алгоритма. Другими словами, кредитной организации необходимо преобразовать имеющуюся информацию о клиенте в количественные показатели, которые позволят принять объективное решение и оценить кредитоспособность клиента.

Применение кредитного скоринга имеет также некоторые ограничения: прямая зависимость от данных и исторической информации. Кроме того, такая методика зависит от качества и количества располагаемых данных о клиенте.

Кредитный скоринг применяется не только в ситуациях, когда заемщик подает заявку на получение кредита. В современном банковском деле выделяют следующие виды кредитного скоринга: аппликационный скоринг (оценка клиента при подаче заявки на кредит), поведенческий скоринг (используется для контроля за действующим клиентом), fraud-скоринг (выявление мошенников), скоринг для целей маркетинга и т. д.

При оценке клиента для построения скоринговой карты бан-

[1] *Статья подготовлена на основании работы, занявшей 2-е место на конкурсе на лучшую работу по экономической тематике среди студентов, магистрантов и аспирантов белорусских вузов, проводимом Национальным банком Республики Беларусь в 2019 г.*

ки пользуются как внутренними базами данных, так и внешними. Внутренние информационные базы банка включают анкетные данные (на момент подачи кредитной заявки) и внутреннюю кредитную историю банка (например, количество и состояние текущих счетов клиента, общая сумма всех кредитов, время погашения последнего кредита и др.). Внешними источниками информации являются бюро кредитных историй по всей банковской системе, страховые организации и социальные учреждения, а также компании – наниматели кредитополучателей.

В Республике Беларусь система сбора и распространения информации об исполнении кредитных обязательств функционирует с 2007 г., и передача данных является обязательной для всех банков[2]. С 01.01.2009 собирается информация по всем кредитным сделкам без ограничения по сумме договора, а с 21.08.2009 – по всем договорам обеспечения. Все это позволяет сформировать единый «пул данных» для всех банков и разрабатывать модели кредитного скоринга на основе общих данных физических лиц, имеющихся в белорусских банках. Такой единой для банков базой данных в Беларуси является Кредитный регистр Национального банка Республики Беларусь (далее – КР, Кредитный регистр). Если сравнивать ситуацию в Республике Беларусь с ситуацией в странах Таможенного союза, то можно отметить, что по количественным показателям кредитной информации (охват взрослого населения и вероятность нахождения кредитной истории) наша страна опережает партнеров по Таможенному союзу.

Скоринговые модели разрабатываются в разных странах в соответствии с разными условиями, но алгоритм разработки такой модели не отличается:

1) сбор и подготовка входных данных: скоринговая система оценки кредитоспособности заемщиков – это, прежде всего, статистическая модель, поэтому необходимо иметь достаточную по объему и качеству базу данных;

2) анализ и корректировка скоринговых переменных: поиск состоятельности переменных и возможных ошибок, проверка их статистической значимости;

3) построение и оценка скоринговой модели: выделяется ряд самых сильных и качественных характеристик, строится первоначальная модель с помощью различных статистических и математических моделей, затем она корректируется;

4) масштабирование и расчет скоринговых баллов: формирование т. н. скоринговой карты (набор характеристик и соответствующих им баллов по определенной шкале) [1].

Сбор и подготовка исходных данных

Начальный этап разработки кредитной скоринговой модели представляется наиболее трудоемким и включает в себя сбор и представление информации о клиенте в логически структурированном и приемлемом для дальнейших расчетов виде.

На практике этап сбора данных начинается с определения временного интервала. Период моделирования подбирается таким образом, чтобы *«Bad Rate»* (соотношение «плохих клиентов» ко всем клиентам в группе) был наиболее стабильным.

Для примера используем традиционный набор обезличенных данных как с внутренних источников одного из белорусских банков, так и с внешних источников (МВД, МНС, ФСЗН, КР) за 2015 г. Нами было выбрано 15 337 договоров, из них 786 (5,1249%) попали в группу неплатежеспособных («дефолты»).

«Bad Rate» в выбранный период был стабильным и репрезентативным. Размер выборки играет большую роль при построении модели. В работах Sven F. Crone и Steven Finlay [5] эмпирически доказано, что размер выборки должен быть максимально расширенным, то есть вся доступная информация должна быть использована для построения модели.

Отметим также, что для стабильности и точности модели банк также устанавливает срок жизни кредитной заявки. Общепринятая практика заключается в выделении заявок, договоры по которым действовали в течение 9 и более месяцев. Такая процедура позволяет «отсечь» высоковолатильные кредиты со сверхкороткими сроками и получить более стабильные скоринговые модели.

Непосредственно перед анализом подготовленных скоринговых переменных необходимо иметь точно сформулированное и зафиксированное в соответствующих документах определение «платежеспособного» и «неплатежеспособного» клиента (или дефолта). Базельский комитет по банковскому надзору предлагает следующее определение дефолта: «существование просроченной задолженности на счету свыше 90 дней когда-либо на протяжении всей кредитной истории» [4, с. 33].

После того как понятие «дефолт» сформулировано, всех клиентов, попадающих в эту категорию, специалисты банка относят к «неплатежеспособным», а остальную часть рассматривают как «платежеспособных» и тем самым формируют зависимую переменную. Это значение и будет прогнозироваться на основе имеющихся данных о клиенте.

Определение «дефолта» сопровождается также установлением критериев, по которым клиент считается благонадежным. Формально можно применить простую логику: «плохой клиент» – если DPD[3] > 90 дней, «хороший клиент» – если DPD < = 90 дней. Такой подход часто используется, но вероятность того, что клиент с DPD = 90 перейдет в просрочку DPD > 90, крайне велика. По вышеприведенному критерию этот клиент соответствует определению «хорошего клиента». Однако мы не можем с уверенностью сказать, платежеспособен клиент или нет. Поэтому целесообразно выделить по целевой переменной

[2] В соответствии с постановлением Правления Национального банка Республики Беларусь от 28 ноября 2006 г. № 196 «Об утверждении Инструкции о порядке получения, формирования, обработки, хранения и предоставления Национальным банком Республики Беларусь сведений о кредитных договорах».

[3] DPD (от англ. Days Past Due) – количество дней, в течение которых наблюдается просроченная задолженность.

«серую зону» (неопределенный тип). При определении целевой переменной в работе использовались следующие правила:

- статус = 1 («плохой»), если DPD > 90;
- статус = 0 («хороший»), если DPD < = 30;
- статус = 2 («неопределенный»), если DPD > 30 и DPD < = 90.

Клиенты со статусом «неопределенный» исключаются из обучающей выборки, однако используются в тестировании скоринговой модели (в основном по распределению: «плохие» клиенты должны в среднем иметь более низкий скоринговый балл и располагаться в левой части; «хорошие» клиенты должны в среднем иметь более высокий балл и располагаться правее; «серая зона» должна быть между «плохими» и «хорошими» клиентами).

Определение независимых переменных представляет собой следующий этап, он заключается в первоначальном экспертном отборе характеристик, которые не имеют предсказательную силу для прогноза дефолтов. Например, по таким признакам, как возраст, тип образования, доходы клиента, можно с определенной точностью разделить сформированную базу клиентов на «платежеспособных» и «неплатежеспособных»; в противоположность объективным связям, например номер дома, где проживает клиент, никак не может помочь нам с классификацией клиентов, поэтому такие характеристики необходимо исключить.

В *Приложении* приведены все использующиеся в модели характеристики. В целом все имеющиеся переменные можно объединить в несколько блоков:

Социально-демографическая информация:

- пол: мужской / женский;
- возраст;
- место жительства: область, район, город, улица, номер дома и квартиры;
- дата регистрации по месту прописки;
- образование: среднее / среднее специальное / высшее;
- семейное положение: женат (замужем) / холост (не замужем) / вдовец (вдова) / разведен (разведена) / совместное проживание;
- количество иждивенцев;
- наличие автомобиля, приобретенного за последние 5 лет.

Информация из КР, ФСЗН:

- кредитный рейтинг Кредитного регистра: A1 / A2 / A3 / B1 / B2 / B3 / C1 / C2 / C3 / D1 / D2 / D3 / E1 / E2 / E3 / F [2];
- количество действующих кредитов;
- ежемесячная долговая нагрузка клиента в бел. руб.;
- ФСЗН: клиент работает / не работает.

Информация о финансах:

- вид дохода: зарплата / пенсия;
- доход клиента по основному месту работы в бел. руб.;
- должность клиента: ИП / не руководящий работник и т. д.;
- количество сотрудников в организации;
- сфера деятельности организации по классификации ОКЭД[4];
- стаж на последнем месте работы.

Другая информация:

– наличие непогашенной просроченной задолженности по действующим кредитным обязательствам, сумма кредитной заявки и др.;

– тип кредита: потребительский кредит, кредитная карта, рассрочка.

Что касается соотношения «плохой» / «хороший» заемщик, то большинство практических экспериментов и специалистов в данной области подтверждают, что для построения скоринговой карты на основе логистической регрессии рекомендуется использовать выборку, где доля «неплатежеспособных» заемщиков составляет не менее 5% от общего количества наблюдений.

Вместо пропущенных значений в выборке подставлялись средние значения (для количественных величин – медиана, для качественных переменных – мода). Однако отметим, что использование импутации пропущенных значений не корректно в случаях, когда по одной переменной доля пустых значений превышает 20%. По этой причине переменные с долей пропущенных значений, превышающих заданное пороговое значение, были исключены из выборки на первом этапе построения скоринговых карт.

В первоначальной («сырой») выборке по каждому договору присутствовало 97 характеристик. После проведения первичного анализа (исследование и решение проблемы пропущенных и некорректных значений, а также выделение характеристик, демонстрирующих стабильную и/или логическую тенденцию к *Bad Rate*) осталось 23 характеристики.

Следующий этап преобразования данных заключается в обработке категориальных и количественных данных, их кодировании. Эмпирически подтверждается, что представление данных в виде бинарных величин положительно влияет на качество построенной логистической регрессии. Бинаризация – перевод значений переменных к виду {0,1} – представляет собой промежуточный этап и для разных типов переменных (количественные и качественные) отличается. Для качественных переменных происходит разбивка по их уникальным значениям (например, одна переменная «пол» разделяется на две: «пол_женский» и «пол_мужской»). В основе алгоритма кодирования количественных переменных, как правило, лежит простая идея: переменная разбивается на 5% (или 10%) персентили. Есть и другие (более автоматизированные) подходы для бинаризации непрерывных переменных, например определение наиболее оптимальной длины интервалов при заданном их количестве методом перебора. Такой способ является одним из наилучших, однако требует больших затрат компьютерных ресурсов и времени. В работе применен первый метод (разбивка по 5% персентилям) с дополнительной ручной корректировкой.

На основе имеющихся данных необходимо построить статистическую модель, которая будет правильно классифицировать клиентов на «неплатежеспособных» и «платежеспособных».

[4] *Общегосударственный классификатор видов экономической деятельности.*

Приложение

Весовой коэффициент, прогностическая способность и балл переменных

№ переменной	Переменная	Значение переменной	№ приказа	Весовой коэффициент	Балл	WOE	Доля в выборке
1	Вид дохода	Пенсия	1	-0,33	18,80	0,23	1,30
2	Возраст	<= 28	2	0,22	-12,51	0,30	60,40
		32–53	3	0,09	-5,28	-0,07	19,00
		>= 54	4	-0,64	37,16	-0,32	10,90
3	Время с момента регистрации	20 и более лет	5	-0,15	8,81	0,20	19,10
4	Долговая нагрузка + Количество действующих кредитов	0 руб. 0 ед.	6	-0,19	11,24	0,11	19,60
		90–350 руб. 1 ед.	7	0,09	-5,33	-0,18	12,80
		90–350 руб. 2 ед.	8	-0,09	5,48	0,13	13,50
		>= 350 руб. 1 ед.	9	-0,40	23,27	0,62	2,30
		>= 350 руб. 2 ед.	10	-0,15	8,91	0,22	4,90
		>= 350 руб. >= 3 ед.	11	0,26	-15,20	-0,31	12,70
5	Должность	ИП, Владелец бизнеса	12	0,73	-42,16	-1,23	1,90
		Неруководящий работник	13	-0,13	7,70	0,00	71,30
		Руководство подразделения	14	-0,34	19,83	0,66	15,80
6	Доход клиента	700–900 бел. руб.	15	-0,09	5,16	0,05	17,10
		>= 900 бел. руб.	16	-0,17	9,91	0,11	29,50
7	Зарплатный клиент	Да	17	-1,09	62,68	1,01	43,70
8	Количество стоп-факторов	0	18	-0,60	34,56	0,18	87,50
9	Количество сотрудников в организации	< 10	19	0,13	-7,51	-0,92	6,10
		>100	20	-0,39	22,35	0,22	73,10
10	Кредитный рейтинг	A1–A3	21	-1,88	108,75	3,55	6,20
		B1	22	-1,13	65,43	1,70	15,30
		B2–B3	23	-0,93	53,84	0,97	26,50
		C1–D1, Не определен	24	-0,46	26,67	0,25	36,90
		E1–F	25	0,09	-5,23	-0,98	15,10
11	Наличие автомобиля в собственности	Имеется	26	-0,67	38,51	0,36	37,00

12	Образование	Среднее	27	0,25	-14,47	-0,45	9,90
		Высшее / неск. высших	28	-0,61	35,46	0,54	42,10
13	Пол	Женский	29	-0,41	23,76	0,16	53,40
14	Проживание	В собств. квартире	30	-0,16	9,23	0,20	51,50
15	Адрес проживания: область, обл. центры	БРЕСТ	31	-0,33	18,94	0,71	5,50
		ВИТЕБСК	32	0,18	-10,58	-0,42	3,70
		ГОМЕЛЬСКАЯ	33	0,12	-6,67	-0,15	8,50
		ГРОДНЕНСКАЯ	34	-0,16	8,95	0,11	2,80
		МИНСК	35	-0,17	10,05	0,29	28,50
		МОГИЛЕВ	36	0,24	-13,66	-0,30	11,00
16	Семейное положение Брак является первым Количество иждивенцев	Женат Да 0 чел.	37	-0,30	17,29	0,35	21,00
		Женат Да >=1 чел.	38	-0,10	5,53	0,08	27,80
		Холост Нет 0 чел.	39	-0,14	8,35	-0,14	20,50
		Холост Нет >=1 чел.	40	0,34	-19,86	-0,60	2,80
17	Стаж на последнем месте работы	до 1 года	41	0,37	-21,10	-0,52	14,30
		3 и более лет	42	-0,24	13,62	0,24	55,90
18	Сфера деятельности организации	Вооруженные силы/ органы внутренних дел/	43	-0,47	27,41	0,33	2,60
		Другие отрасли	44	0,28	-16,41	-0,26	15,10
		Образование	45	-0,22	12,60	0,69	3,00
		Оптовая/ розничная торговля	46	0,43	-24,87	-0,16	24,60
		Промышленность и машиностроение	47	0,18	-10,67	0,11	19,20
		Строительство	48	0,31	-17,92	-0,53	6,40
		Услуги	49	0,20	-11,81	-0,28	7,80
		Финансы, банки страхования	50	-0,60	34,42	2,35	7,60
19	Тип кредита	Кредитные карты	51	-0,24	13,57	0,20	56,60
20	ФСЗН	Не работает	52	0,11	-6,06	-0,84	8,60
		Работает / силовые структуры	53	-0,46	26,65	0,10	91,40

Примечание. *Разработка авторов.*

Прежде чем переходить к анализу и отбору переменных, требуется убедиться, что модель будет хорошо работать на новой выборке. Такая проверка делается путем разбиения собранных размеченных данных на две части. На одной части выборки обучается модель (обучающая выборка), на другой тестируется ее прогнозная сила (тестовый набор данных). Пропорции разделения всей выборки на обучающую часть и тестовую составляют 80% и 20% соответственно.

В работе в качестве математической модели использована логистическая регрессия. Также для получения наиболее достоверных и несмещенных оценок классификатора использован такой метод регуляризации, как *L1* (метод *lasso*[5]). Такой шаг делается для отбора некоррелируемых переменных при большой размерности данных.

Стоит остановиться на важном моменте – в работе употребляются такие термины, как переменная (другое название – характеристика) и признак (или атрибут). Например, переменная «пол» имеет два признака («мужской», «женский»).

Анализ и корректировка скоринговых переменных

После сбора и преобразования данных к бинарному виду следует анализ и сжатие данных путем снижения размерности (группировка признаков характеристик). Для этого необходимо для всех переменных рассчитать *WOE* и *IV*-значение.

Мера прогностической способности атрибута характеристики – весомость:

$$WOE = ln\frac{p_i}{q_i},$$

где p_i – доля «хороших», имеющих атрибут i, во всех хороших; q_i – доля «плохих», имеющих атрибут i, во всех плохих.

Предиктивная сила атрибута (информационный критерий) рассчитывается как:

$$IV = (p_i - q_i) \times ln\frac{p_i}{q_i}.$$

Расчет величины *IV* позволяет на раннем этапе определить, насколько «информативна» данная характеристика; как правило, если этот показатель менее 2%, то такую характеристику исключают из имеющегося множества. *WOE*-значение, в свою очередь, дает четкое представление о «направлении» признаков рассматриваемой группы. Лучше всего эту логику проследить в представленном выше примере расчета для переменной «возраст». Вначале любая характеристика разбивается на оптимальные группы (способы описаны ранее). Затем для каждой группы рассчитывается количество «дефолтов» и «недефолтов» (например, согласно *таблице 1* в группе людей в возрасте до 28 лет 198 клиентов имеют просроченную задолженность более 90 дней, остальные 2 717 человек выполнили свои обязательства перед банком). Далее проводится несложный расчет доли таких клиентов и показателя *WOE* по указанным формулам. В итоге можно увидеть «антинаправление»[6] той или иной группы переменной. Другими словами, люди в возрасте до 28 лет в среднем оказываются менее благонадежными, нежели более возрастное поколение.

Также рассчитанные *WOE*-значения используются для снижения размерности данных. На *рисунке 1* представлен пример рассчитанных *WOE*-значений для характеристики «кредитный рейтинг». Оценка «кредитного рейтинга» выставляется Национальным банком Республики Беларусь по каждому клиенту на основании имеющейся кредитной информации о его поведении в прошлых периодах. Эта характеристика формируется с помощью собственной скоринговой модели банка, где каждому диапазону полученного Национальным банком скорингового балла присвоен класс кредитного рейтинга. Такие значения легче сравнивать, поскольку они более понятны как для пользователей кредитных историй, так и для их субъектов [2].

Характеристика «кредитный рейтинг» делится на 18 отдельных классов. Эти классы можно объединить (в целях снижения размерности выборки и получения более стабильных оценок) по

Таблица 1

Расчет WOE-значений для переменной «возраст»

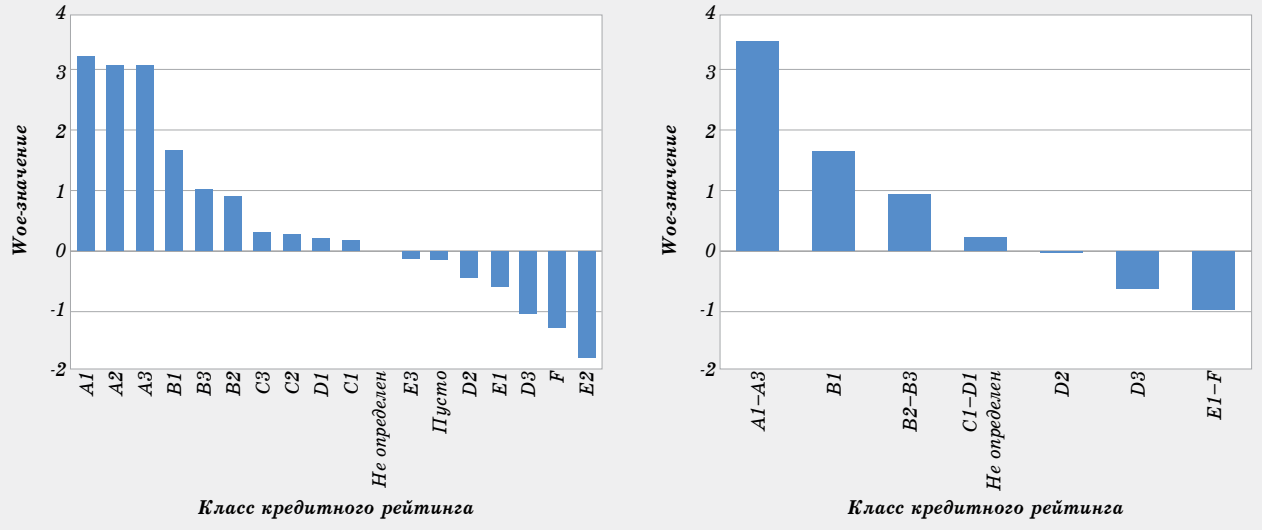
Группы переменной		Наличие дефолта по кредиту		Итого	Доля недефолтов	Доля дефолтов	WOE	IV
		Нет	Да					
Возраст	<= 28	2 717	198	2 915	0,1867	0,2519	0,2994	1,9519
	29–31	1 406	81	1 487	0,0966	0,1031	0,0644	0,0414
	32–53	8 814	444	9 258	0,6057	0,5649	-0,0698	0,2852
	>= 54	1 614	63	1 677	0,1109	0,0802	-0,3249	0,9996
Итого		14 551	786	15 337	1,0000	1,0000		3,2781

Примечание. *Разработка авторов.*

[5] *Регуляризация L1, или lasso (англ. least absolute shrinkage and selection operator, русс. оператор наименьшего абсолютного стягивания и отбора, представляет собой один из подходов к решению проблемы с переобучением).*

[6] *Для такого анализа требуется умножить WOE-значение на «–1».*

Сравнительный анализ WOE-значений для характеристики «кредитный рейтинг» до и после кластеризации



Примечание. *Разработка авторов.*

Рисунок 1

Таблица 2

Статистическая значимость переменных на основе IV-значения

№	Название переменной	IV-значение
1	Зарплатники	0,4085
2	Кредитный рейтинг	0,2771
3	Сфера деятельности организации	0,2707
4	Количество стоп-факторов	0,1378
5	Количество сотрудников в организации	0,1311

Примечание. *Разработка авторов.*

схожим *WOE*-значениям. Так, классы «А1», «А2», «А3» ведут себя схожим образом по отношению к целевой переменной «дефолт», поэтому эти признаки можем объединить в одну группу «А1–А3» *(рисунок 1)*.

Для проверки статистической значимости переменных использовался такой показатель, как *IV* (информационное значение). Наиболее значимые переменные по расчетным величинам представлены в *таблице 2*.

Наиболее «информативной» характеристикой клиента, тесно коррелирующей с показателем дефолта, является признак «зарплатники», принимающий бинарные значения («1» – незарплатник; «0» – зарплатник). Если клиент обслуживается в банке в рамках зарплатного проекта, то шанс просрочить платежи более чем на 90 дней, относительно невелик (предиктивная сила – 0,4058).

В реальных наборах данных, как правило, присутствует взаимосвязь между переменными в той или иной мере (т. н. мультиколлинеарность). Сильная связь между регрессорами может сделать оценки коэффициентов нестабильными, смещенными и несостоятельными [3, с. 13]. Существует множество способов устранить эту проблему, к примеру, самый простой – это установление регуляризации типа *L1* с помощью изучения корреляционной матрицы всех количественных регрессоров, а для качественных переменных – коэффициенты сопряженности Пирсона, которые позволяют вручную проанализировать связь всех переменных и выделить те из них, которые потенциально могут дестабилизировать прогнозные оценки модели.

Так как качественные характеристики «брак является первым» и «количество иждивенцев» разбиваются на признаки (одна характеристика раскладывается на несколько признаков в бинарной форме), то для выяснения силы взаимосвязи характеристик следует построить таблицу сопряженности и рассчитать коэффициент сопряженности Пирсона *(таблица 3)*.

В *таблице 3* представлены фактические и ожидаемые (в квадратных скобках) частоты для рассматриваемой зависимости между признаками характеристик «брак является первым» и «количество иждивенцев». Согласно данной таблице значение статистики хи-квадрат равняется 10 113,98. Критическое значение при уровне значимости 5% и 4 степенях свободы будет меньше рассчитанного значения. Дальнейший расчет коэффициента сопряженности Пирсона (равный 0,6304) также подтверждает наличие взаимосвязи между характеристиками «брак является первым» и «количество иждивенцев». С рассматриваемыми характеристиками коррелирует также

Таблица 3

Значения частот для расчета критерия хи-квадрат по переменным «брак является первым» и «количество иждивенцев»

Брак является первым?	Семейное положение					Итог
	вдовец/ вдова	женат/ замужем	холост/ не замужем	разведен/ разведена	Совместное проживание	
ДА	0 [227]	7 489 [4 428]	0 [1 747]	0 [968]	1 [119]	7 490
НЕТ	465 [238]	1 579 [4 640]	3 578 [1 831]	1 983 [1 015]	242 [124]	7 847
Итог	465	9 068	3 578	1 983	243	15 337
Значение статистики хи-квадрат		10 113,98	Коэффициент взаимной сопряженности Пирсона			0,6304

Примечание. *Разработка авторов.*

Таблица 4

Список признаков новой переменной «семейное положение + брак является первым + количество иждивенцев»

№ признака	Семейное положение	Брак является первым?	Количество иждивенцев	Доля в выборке, %
1	Женат/замужем	Да	0	21,0
2	Женат/замужем	Да	1 и более	27,8
3	Холост/не замужем	Нет	0	20,5
4	Холост/не замужем	Нет	1 и более	2,8

Примечание. *Разработка авторов.*

и характеристика «количество иждивенцев» (коэффициент взаимной сопряженности Пирсона равен 0,4328). Для того чтобы сохранить как можно больше данных для моделирования, эти три переменные были преобразованы в одну, уникальные значения которой представлены в *таблице 4*.

Также в имеющемся наборе данных присутствует логическая взаимосвязь некоторых переменных, например: «количество действующих кредитов» тесно коррелирует с характеристикой «долговая нагрузка клиента». До бинаризации признаков рассчитан парный коэффициент корреляции Пирсона, равный 0,5080. По этой причине данные характеристики были объединены в одну, признаки которой представлены в *таблице 5*.

В результате проведенных мероприятий по устранению коррелируемых признаков количество характеристик сократилось с 23 до 20. Однако такое сокращение размерности положительно отражается на качестве разрабатываемой модели по следующим причинам: исключается «шум» в данных из-за наличия корреляции в независимых переменных, а также сохраняется ценная информация о клиентах.

Таблица 5

Список признаков новой переменной «долговая нагрузка + количество действующих кредитов»

№ признака	Долговая нагрузка, бел. руб.	Количество действующих кредитов, шт.	Доля в выборке, %
1	0	0	19,6
2	90–350	1	12,8
3	90–350	2	13,5
4	>= 350	1	2,3
5	>= 350	2	4,9
6	>= 350	>= 3	12,7

Примечание. *Разработка авторов.*

Построение и оценка скоринговой модели

Взяв на вооружение модель логистической регрессии, выборку в виде категориальных переменных, а также *L1*-регуляризацию, мы построили модель следующего вида:

$$\hat{y} = 0{,}049 - 0{,}33 \times x_1 - 0{,}22 \times \times x_2 + 0{,}09 \times x_3 - 0{,}64 \times x_4 - \ldots - - 0{,}46 \times x_{53},$$

где $\hat{y}$ – натуральный логарифм прогнозной вероятности дефолта заемщика;

x – независимые переменные (регрессоры);

i – порядковый номер признака ($i \in [1,53]$, см. *Приложение*).

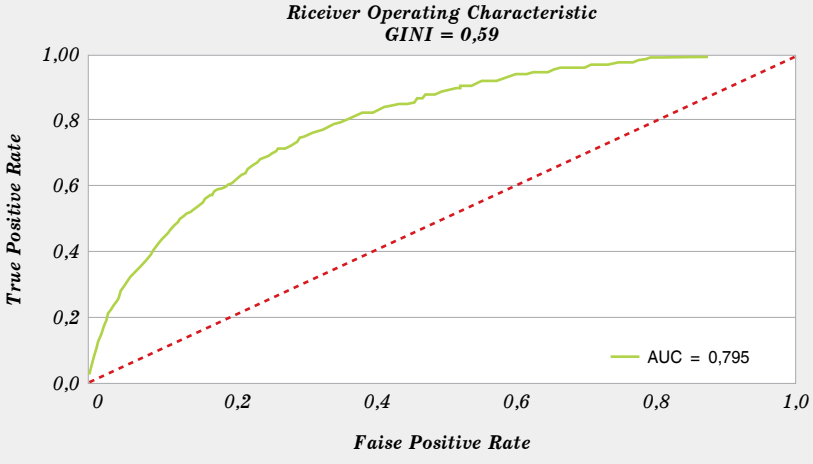
После построения модели специалисты первым делом анали-

зируют статистическую значимость переменных с помощью расчета *p*-значения[7]. В рассматриваемом примере все оценки коэффициентов признаны статистически значимыми (их *p*-значение менее 5%).

Следующий шаг заключается в количественном определении, насколько точно работает полученная модель. На сформированной до моделирования тестовой выборке проверяется способность модели отличить «платежеспособного» клиента от «неплатежеспособного». Для оценки качества построенной модели используют *ROC*-кривую (англ. *Receiver Operator Characteristic, кривая ошибок*) вместе с последующим расчетом коэффициента GINI (Джини). Эта оценка модели показывает зависимость количества верно классифицированных положительных исходов от количества неверно классифицируемых отрицательных исходов.

На обучающем наборе данных получена высокая оценка качества модели, коэффициент GINI, равный 60,2%. Полученная оценка качества построенного бинарного классификатора на тестовом наборе (*рисунок 2*), а именно показатель GINI, равный 59%, позволяет утверждать, что логит-регрессия корректно обучилась на обучающей выборке и относительно точно обобщает результаты на тестовой выборке. При относительно низком пороге отсечения, например, точка (0,2; 0,6), модель правильно классифицирует «плохих» клиентов (отсекает 60% таких заемщиков), однако и неверно определяет «хороших» клиентов (граница отсечения на уровне 20%). После этапа моделирования следует процесс определения уровня отсечения («cut-off»), который заключается в установлении приемлемого для банка баланса между недополученной прибылью (отказы неверно классифицируемым «хорошим» клиентам) и прогнозируемыми потерями (верно определенные «плохие» клиенты).

ROC-кривая и коэффициент GINI тестовой выборки



Примечание. *Разработка авторов.*

Рисунок 2

Масштабирование и расчет скоринговых баллов

По формуле логистической регрессии в результате мы получаем прогноз в шкале натуральных логарифмов, что довольно трудно интерпретировать и применять. Необходимо привести весовые коэффициенты в линейную шкалу. Как правило, применяются методы масштабирования.

Для масштабирования требуется, во-первых, определить желаемый диапазон распределения баллов (минимальное и максимальное значения), а также два показателя: количество баллов, удваивающее шансы стать «хорошим» заемщиком, и значение шкалы, в котором и достигается заданное отношение шансов «платежеспособных» к «неплатежеспособным». Наиболее часто используют систему, в которой каждые 40 баллов удваивают шансы стать «платежеспособным» клиентом, а также в точке 600 баллов отношение шансов составляет 72:1. При дальнейших расчетах будет использоваться эта методика.

Для того чтобы привести коэффициенты логистической регрессии в линейную шкалу, применяют следующее преобразование:

$$\text{Балл} = A + R \times \sum_{i=1}^{n} b_i ,$$

где A – смещение;

R – множитель;

b_i – весовой коэффициент при регрессоре i (n – их количество).

$$R = D/\ln 2,$$

где D – количество баллов, удваивающее шансы наступления дефолта.

$$A = B - R \times \ln C,$$

где C – константа, B – значение по шкале баллов, в которой соотношение шансов составляет C:1 [3].

Приведем пример расчета скоринговых баллов – кредитный рейтинг заемщика, который Национальный банк Республики Беларусь предоставляет в виде классов (от A1 до F). Предполагая, что каждые 40 баллов удваивают шансы наступления дефолта по кредиту, а также в точке 600 баллов отношение шансов составляет 72:1, в *таблице 6* приведены результаты расчетов по вышеприведенной формуле: множитель будет равен 57,71, а смещение составит 413,2. Поскольку класс A1 является наилучшим и отдаление от него в сторону класса F предполагает увеличение вероятности дефолта заемщика, то множитель нужно брать с отрицательным знаком.

[7] *Р-значение (уровень значимости) – вероятность того, что оценка коэффициента сформировалась случайным образом. Коэффициент статистически значим, если его р-значение не превышает заданную границу (как правило, 5%).*

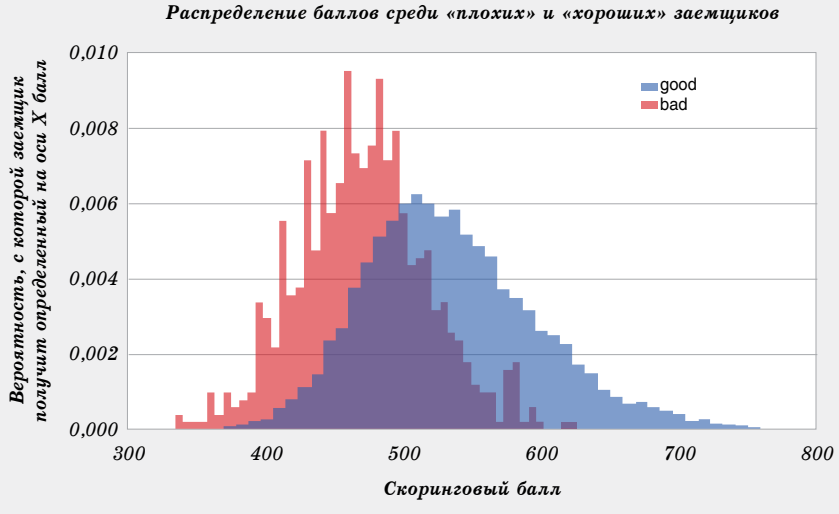
Таблица 6

Расчет скоринговых баллов для переменной «кредитный рейтинг»

№	Признаки переменной	Весовой коэффициент	Скоринговый балл по линейной шкале bi×R
1	A1 / A2 / A3	-1,88	108,75
2	B1	-1,13	65,43
3	B2 / B3	-0,93	53,84
4	C1 / C2 / C3 / D1	-0,46	26,67
5	E1 / E2 / E3 / F	0,09	-5,23

Примечание. *Разработка авторов.*

Распределение платежеспособных и неплатежеспособных клиентов по скоринговым баллам



Примечание. *Разработка авторов.*

Рисунок 3

В *Приложении* приведены результаты расчетов скоринговых баллов для всех характеристик и признаков. Для получения общего скорингового балла необходимо сложить баллы по каждой независимой переменной и прибавить рассчитанное значение смещения. Для иллюстрации распределения клиентов по скоринговым баллам построена гистограмма *(рисунок 3)*.

После построения скоринговой карты на заключительном этапе руководство конкретного подразделения (в данном случае директор департамента управления розничными кредитными рисками) утверждает и защищает перед комитетом новую скоринговую карту, ее качество и стабильность работы, границу отсечения («cut-off»), расчетную величину прогноза отношения «плохих» клиентов к кредитному портфелю и запускает в работу.

В банковской сфере оценка платежеспособности заемщиков представляет собой первостепенную задачу в управлении кредитным риском. И точность таких оценок оказывает прямое влияние на качество кредитного портфеля банка, его надежность и эффективность деятельности.

Кредитный скоринг зарекомендовал себя как наиболее стабильный и действенный метод оценки вероятности невыполнения кредитных обязательств со стороны заемщика. В основе работы скоринговой системы лежит автоматическое присвоение определенных баллов в зависимости от параметров запрашиваемого кредита, кредитной истории, социально-демографических характеристик заемщика, а также размера его доходов. В свою очередь, скоринговые баллы, соответствующие каждой характеристике клиента, вычисляются на основе ретроспективного эконометрического анализа поведения клиентов. Использующийся математический инструментарий состоит из (но не ограничивается) построения корреляционных матриц для решения проблемы взаимосвязанных характеристик, расчета показателя информационного значения для отбора наиболее значимых переменных, вычисления, настройки параметров и проверки значимости весовых коэффициентов логистической регрессии, а также приведения полученных коэффициентов в баллы путем масштабирования.

В работе приведена необходимая теоретическая база, показан практический пример применения наиболее современных математических методов и моделей для создания успешных скоринговых систем.

Для поддержания устойчивости и увеличения эффективности банка требуется проведение регулярного мониторинга построенных скоринговых систем, изменение клиентского потока в банке и во всей банковской системе, учет экономической и политической ситуации в стране, а также повышение навыков и знаний разработчиков скоринговых систем.

* * *

Материал поступил 27.11.2019.

Библиографический список:

1. Ковалев, М. Методика построения банковской скоринговой модели для оценки кредитоспособности физических лиц [Электронный ресурс] / М. Ковалев, В. Корженевская // Белорусский государственный университет. – Режим доступа: https://www.bsu.by/Cache/pdf/49623.pdf. – Дата доступа: 23.02.2019.

2. Рейтинг Кредитного регистра Национального банка [Электронный ресурс] // Национальный банк Республики Беларусь. – Режим доступа: https://www.nbrb.by/today/CreditRegistry/Instructions/docs/Rating_CR_NBRB.pdf. – Дата доступа: 11.07.2019.

3. Сорокин, А.С. Построение скоринговых карт с использованием модели логистической регрессии / А.С. Сорокин // Науковедение. – 2014. – № 2. – С. 1–29.

4. Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework [Electronic resource] / Bank Of International Settlements // Basel Committee on Banking Supervision. – Mode of access: https://www.bis.org/publ/bcbs118.pdf. – Date of access: 29.02.2019.

5. Sven, F. Crone. Instance sampling in credit scoring: An Empirical study of sample size and balancing / F. Crone Sven, F. Steven // International Journal of Forecasting. – 2012. – № 28. – P. 224–238.

Modern Practice Of Developing Credit Scoring Cards For Retail Clients In Belarusian Banks

Oleg GICHAN, *Ph.D. Student, Belarusian State University, Minsk, Republic of Belarus, e-mail: ogichan@bk.ru.*

Catherine GOSPODARIK, *Ph.D. in Economics, Associate Professor, Head of the Analytical Economy and Econometrics Department, Belarusian State University, Minsk, Republic of Belarus, e-mail: gospodarik@bsu.by.*

Abstract. *Scorecards are developed in the majority of Belarusian and foreign banks to assess the borrowers' credit worthiness. The banking organizations are recommended to reexamine the valid scorecards, create the new ones and choose, by means of comparing their forecasting power, the more effective and precise scorecards with a view to increasing competitiveness in the credit resources market. The authors created the scorecard for domestic banks on the basis of logistic regression with the use of certain possibilities of computer-aided learning and modern technological toolkit.*

Keywords: *credit risk; credit scoring; logistic regression.*