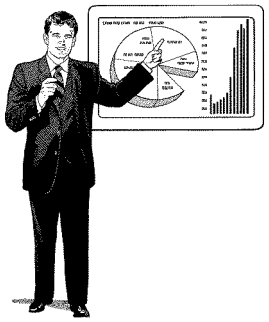
АНАЛИТИЧЕСКИЕ ПРОГНОЗЫ И ПРЕДЛОЖЕНИЯ



Г. О. ЧИТАЯ, Д. Д. КАЗУНОВА

СКОРИНГОВАЯ МОДЕЛЬ ОЦЕНКИ КРЕДИТНЫХ РИСКОВ

В статье представлена модель оценки кредитного риска розничного заемщика в рамках аппликационного скоринга. Рассмотрено методическое обеспечение построения скоринговой карты по обоснованному набору показателей и определению оптимального уровня скоринговых баллов, используемых для отсеивания рисковой группы потенциальных кредитополучателей. Классификация заемщиков осуществлена на основе разработанной эконометрической модели логистической регрессии и применения методов машинного обучения. Численная реализация математической модели проведена на основе данных выборочной совокупности заемщиков одного из минских банков с использованием программного продукта MatLab.

Ключевые слова: кредитный риск; заемщики; скоринг; вероятность дефолта; машинное обучение; логистическая регрессия; скоринговая карта; показатели кредитоспособности.

УДК 336.717:005.334 (476)

Введение. На кредитоспособность заемщика влияет многообразие факторов, что приводит к необходимости разработки и применения гибких систем управлениями рисками в банковской сфере. Этому способствует конкуренция между коммерческими банками за получение надежных и платежеспособных клиентов. В этой связи одним из наиболее приоритетных направлений выступает управление кредитным риском в рамках аппликационного скоринга, предполагающего создание системы быстрого рассмотрения кредитной заявки и оценки кредитоспособности широкого круга заемщиков. Модель логистической регрессии позволяет оценить вероятность того, что клиент может получить кредит или ему будет в этом отказано. Если вероятностная величина окажется меньше 0,5, то рекомендуется ему не выдавать кредит, так как он для банка может оказаться фактором дефолта. Подобная схема оценки может привести к потере клиентов для банка и уменьшать его прибыль, поэтому переменные (показатели) кредитоспособности заемщика, значимость которых устанавливается с помощью логистической регрессии, в рамках системы скоринга переводятся в категориальные и измеряются

Гигла Отарович ЧИТАЯ (chitaya_g@bseu.by), *доктор экономических наук, профессор, зав. кафедрой математических методов в экономике Белорусского государственного экономического университета (г. Минск, Беларусь);*

Дарья Денисовна КАЗУНОВА (Kazyn810@gmail.com), *магистрант Белорусского государственного экономического университета (г. Минск, Беларусь).*

по балльной шкале от 0 до 1 000. Кумулятивный балл по выбранным показателям улучшает процедуру классификации клиентов на дефолтные и недефолтные. Однако на практике возникает задача не только принятия решения в отказе или выдаче кредита конкретному заемщику на основе набранного количества баллов, но и задача определения оптимального или минимального их количества для выдачи кредита. Вторая задача решается на основе анализа распределения баллов «надежных» и «ненадежных» заемщиков на основе разработанной скоринговой карты и тесно связана с анализом соотношения риска и доходности во всем кредитном портфеле банка. В связи с этим цель настоящей работы состоит в разработке модели оценки кредитного риска розничного заемщика на основе аппликационного скоринга, с дальнейшим построением скоринговой карты и определением оптимального уровня отсечения баллов.

Показатели кредитоспособности и установление их значимости для выборочной совокупности заемщиков. С точки зрения применения экономико-математических методов задача определения кредитоспособности потенциального заемщика является типичной среди гораздо более широкого класса задач, связанных с распределением набора имеющихся объектов, каждый из которых характеризуется вектором определенных показателей (переменных) по нескольким заранее установленным категориям. Оценка кредитоспособности заемщиков с применением методов прикладного статистического анализа содержится в работе [1, с. *463—475*], оптимизация структуры просроченной задолженности на среднесрочную перспективу в сочетании с применяемой стратегией взыскания долга осуществлена в [2, с. *130—136*].

В приводимой статье объектами выступают заемщики, переменными — показатели их кредитоспособности, классами — разбиваемые на две классификационные группы заемщики, отвечающие состояниям отсутствия дефолта и наступления дефолта соответственно (в дальнейшем дефолт/недефолт). Для большого количества заемщиков, составляющих несколько десятков тысяч, задачу их бинарной классификации целесообразно решать методами машинного обучения. Использование процедуры обучения «с учителем» предполагает разбиение выборочной совокупности заемщиков на обучающую и тестовую, по которым скоринговая модель сортирует заемщиков, соотнося их с заданными двумя классами. Методами машинного обучения можно автоматизировать процесс кредитного скоринга, так как это позволяет:

сократить время, требуемое для кредитного скоринга;

уменьшить вероятность ошибки принятого решения;

работать с большим объемом данных о заемщиках.

Скоринговая модель строится путем определения зависимой (целевой, классифицирующей) и независимых переменных. Зависимая переменная представляется дефолтом клиента и определяется по общепринятому правилу: максимальное количество дней просроченной задолженности заемщика превышает 90 за 12 месяцев с момента заключения кредитного договора.

В данной статье скоринг заемщиков осуществляется не только в теоретическом, но и в прикладном плане, что потребовало создания выборочной совокупности заемщиков по определенному набору показателей (переменных) их кредитоспособности. Показатели характеризуют социально-демографический статус заемщика, содержат информацию из банковской кредитной истории (БКИ), из фонда социальной защиты населения (ФСЗН) и министерства внутренних дел (МВД).

Социально-демографические показатели: пол; возраст; место жительства; семейное положение.

Информация из БКИ, ФСЗН и МВД: количество запросов на кредитование (БКИ); доход клиента по основному месту работы и стаж (ФСЗН); привлечение к административной ответственности и утерянный паспорт (МВД).

Прочие данные о клиенте: получение заработной платы на счета в банке.

Выборка заемщиков создана для набора перечисленных показателей по одному из минских банков и включает клиентов с датой сделки в период с 01.11.2017 по 31.12.2019. В результате выбрано 58 184 договоров, из них 963 (1,66 %) оказались неплатежеспособными. Процент «плохих» клиентов (*bad rate*) в выбранный период был стабильным и репрезентативным. Выборка для калибровки модели включает клиентов с датой сделки в период с 25.01.2021 по 24.01.2022. В результате выбрано 16 898 договоров, из них 580 (3,43 %) оказались неплатежеспособными. Здесь так же *bad rate* в выбранный период был стабильным и репрезентативным (табл. 1).

Таблица 1. **Сводное описание выборок**

Показатель	Обучающая выборка	Выборка для калибровки
Временной период для формирования выборки, дата заявки	01.11.2017 – 31.12.2019	25.01.2021 – 24.01.2022
Определение дефолта	Среднесрочный таргет – выход на просрочку 90 дней первые 12 месяцев с даты выдачи кредита	
Определение недефолта	Отсутствие выхода на просрочку 90 дней в первые 12 месяцев с даты выдачи кредита	
Общее количество сделок в выборке	58 184	16 898
Количество дефолтов	963	580
Количество недефолтов	57 221	16 318
Уровень дефолтов	1,66 %	3,43 %

Для включения в модель логистической регрессии количественных переменных необходимо проанализировать их на наличие мультиколлинеарности. Первоначальный анализ может быть произведен на основе матрицы парных и частных корреляций между независимыми переменными. Однако коэффициенты корреляции не всегда могут показать наличие мультиколлинеарности. Поэтому правомерно использовать параметр, называемый коэффициентом или фактором «вздутия» дисперсии (*VIF – Variance Inflation Factor*)

$$VIF = \frac{1}{1 - R_i^2}, \qquad (1)$$

где R_i^2 – квадрат множественного коэффициента корреляции [3, с. *158*].

Расчет коэффициента вздутия дисперсии по основным количественным переменным проведен в Matlab R2022b (см. ниже).

Коэффициент вздутия дисперсии количественных переменных

Название характеристики	*VIF-значение*
Возраст	1,0254
Балл БКИ	1,0113
Доход	1,0675
Привлечение к административной ответственности за последние 12 месяцев	1,0556
Количество обращений в банки	1,0616

Так как *VIF*-значение для количественных переменных практически равно 1, мультиколлинеарность отсутствует.

Далее необходимо выполнить проверку статистической значимости всех характеристик по заемщику, т. е. определить наличие и силу связи между одной зависимой и независимыми переменными. Проверку удобно проводить с

помощью показателя IV (информационное значение), воспользовавшись при этом следующей формулой:

$$IV = \sum_{j=1}^{m} \left(d_j^{(1)} - d_j^{(2)} \right) \cdot WOE, \tag{2}$$

где $d_j^{(1)}$ и $d_j^{(2)}$ — относительные частоты «плохих» и «хороших» клиентов.

В свою очередь параметры WOE (веса) для каждой градации (категории) показателя рассчитываются по формуле

$$WOE_j = \ln\left(\frac{d_j^{(1)}}{d_j^{(2)}} \right). \tag{3}$$

Веса градаций показателя помогают найти «границы чувствительности» к появлению моделируемого события риска и провести оптимальным образом категоризацию количественных переменных [4, с. *10*].

Предварительный анализ взаимосвязи скоринговых переменных с вероятностью дефолта по кредиту позволяет ограничить количество рассматриваемых для построения модели логистической регрессии переменных. Значимость характеристик по расчетным величинам представлена ниже.

Информационная значимость характеристик

Название характеристики	*IV-значение*
Возраст	0,34992
Балл БКИ	0,27221
Доход	0,23704
Семейное положение	0,12886
Стаж	0,10571
Получение заработной платы на карту банка	0,05401
Привлечение к административной ответственности за последние 12 месяцев	0,05302
Количество обращений в банки	0,05291
Утерянный паспорт	0,01860
Пол	0,00355
Место проживания	0,00061

Следует определить, какое оптимальное количество из имеющихся характеристик является достаточным для построения скоринговой модели с помощью визуальной интерпретации их прохождения порогового значения показателя IV (рис. 1).

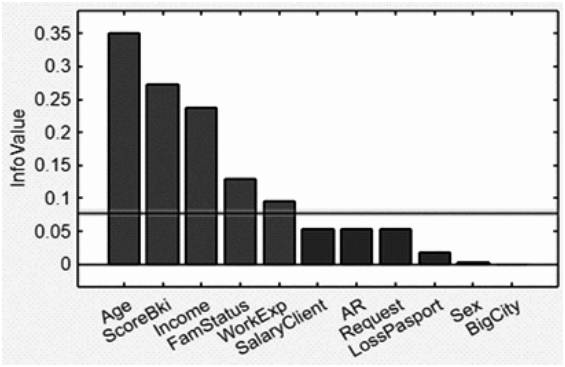


Рис.1. Градация характеристик по информационной значимости

Наиболее «информативной» характеристикой клиента, тесно коррелирующей с показателем дефолта, является возраст. Балл банковской кредитной истории является достаточно информативным показателем, занимающим вторую позицию среди наиболее статистически значимых характеристик. Следующая по значимости переменная — доход клиента, подтвержденный по базе ФСЗН. Четвертой значимой характеристикой является семейное положение, которое клиент выбирает из предлагаемого списка: женат/замужем, холост/не замужем, в разводе или вдова/вдовец. Стаж клиента также играет немаловажную роль в предсказании дефолтности заемщиков. Традиционно с увеличением стажа на одном месте работы возрастает и вероятность того, что клиент окажется платежеспособным.

В первоначальной выборке по каждому договору присутствовало 11 характеристик. После оценки значимости было принято решение оставить пять характеристик. При этом возраст, балл БКИ и доход клиента являются характеристиками с высокой прогностической способностью, а семейное положение и стаж работы относятся к характеристикам со средней прогностической способностью.

Категоризация переменных и построение модели логистической регрессии. Следующим этапом построения скоринговой модели является категоризация, или биннинг, количественных переменных. Для этого было использовано *WOE*-значение, рассчитанное по формуле (3).

Процедура биннинга количественных переменных проведена на примере возраста заемщика. После первоначального биннинга возраста заемщиков получили категории, изображенные на рис. 2.

Bin	Good	Bad	Odds	WOE	InfoValue
{'[-Inf,26)'}	5199	1325	3.9238	-0.49998	0.03346
{'[26,28)' }	3460	798	4.3358	-0.40012	0.013516
{'[28,31)' }	5422	1072	5.0578	-0.24609	0.0073877
{'[31,34)' }	5099	839	6.0775	-0.062442	0.00040709
{'[34,37)' }	4637	733	6.3261	-0.022354	4.6498e-05
{'[37,41)' }	5623	856	6.5689	0.015319	2.5985e-05
{'[41,45)' }	5070	705	7.1915	0.10587	0.00107
{'[45,50)' }	4879	672	7.2604	0.11541	0.0012178
{'[50,56)' }	5699	736	7.7432	0.17978	0.0033456
{'[56,Inf]' }	5306	54	98.259	2.7206	0.26759
{'Totals' }	50394	7790	6.4691	NaN	0.32807

Рис. 2. Первоначальный биннинг возраста заемщиков

Можно заметить, что у 1—2, 4—5 и 7—8 категории *WOE*-значения достаточно близкие, следовательно, эти категории могут быть объединены. Полученные категории вместе с их *WOE*-значениями и соотношениями «плохих» и «хороших» заемщиков отображены на рис. 3.

Далее для всех без исключения градаций категориальных переменных необходимо рассчитать показатели *WOE* и заменить ими фактические значения независимых переменных. По значениям *WOE* и будет строиться модель логистической регрессии.

Значения показателя *WOE* для семейного положения представлены на рис. 4.

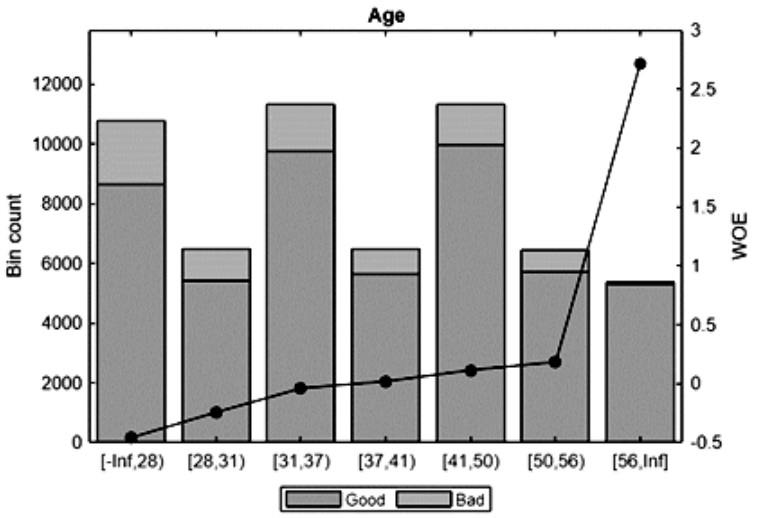


Рис. 3. Итоговые категории возраста

Bin	Good	Bad	Odds	WOE
{'1' }	33151	3857	8.595	0.28415
{'2' }	7237	1308	5.5329	-0.15632
{'3' }	2725	552	4.9366	-0.27036
{'4' }	7281	2073	3.5123	-0.61076
{'Totals'}	50394	7790	6.4691	NaN

Рис. 4. Показатели *WOE* семейного положения заемщика

В данном случае первая категория соответствует статусу женат/замужем, вторая — холост/ не замужем, третья — вдова/ вдовец и четвертая — в разводе. Подобным образом представлены показатели *WOE* для стажа заемщика (рис. 5).

Bin	Good	Bad	Odds	WOE
{'1' }	7800	1879	4.1511	-0.44365
{'2' }	7814	1507	5.1851	-0.22124
{'3' }	21767	3159	6.8905	0.063108
{'4' }	13013	1245	10.452	0.47978
{'Totals'}	50394	7790	6.4691	NaN

Рис. 5. Показатели *WOE* стажа заемщика

Первой категории соответствует стаж до 3 месяцев на текущем основном месте работы, второй — от 3 месяцев до 1 года, третьей — от 1 года до 3 лет и заключающей, четвертой, — от 3 лет и более.

Для двух рассмотренных выше характеристик типично такое деление, поскольку информация о семейном положении и стаже указана в анкете для заемщика именно в такой форме и других значений иметь не может.

В результате программной реализации в Matlab R2022b были получены коэффициенты уравнения логистической регрессии, каждый из которых является статистически значимым (рис. 6).

	Estimate	SE	tStat	pValue
(Intercept)	1.8597	0.013319	139.63	0
Age	0.84687	0.031394	26.975	2.8757e-160
ScoreBki	0.93505	0.025561	36.581	5.7023e-293
Income	0.98766	0.045869	21.532	7.7852e-103
FamStatus	0.6655	0.039566	16.82	1.7417e-63
WorkExp	0.32279	0.047711	6.7656	1.3277e-11

Рис. 6. Коэффициенты уравнения логистической регрессии

Для интерпретации коэффициентов модели логистической регрессии целесообразно использовать экспоненциальную форму записи

$$p_i = \frac{1}{1+e^{-(1{,}8597+0{,}84687x_{i1}+0{,}93505x_{i2}+0{,}98766x_{i3}+0{,}6655x_{i4}+\mathbf{0},32279x_{i5}+\varepsilon)}},$$

$$i = 1, 2, \ldots, n.$$

Построение скоринговой карты и валидация скоринговой модели. После разработки скоринговой модели необходимо выполнить перевод коэффициентов логистической регрессии в скоринговые баллы путем их масштабирования, которое не изменяет прогностическую способность скоринговой карты, а лишь переводит скоринговые баллы в новую шкалу, удобную для использования. Для масштабирования требуется задавать диапазон числовой шкалы (например, от 0 до 1 000). На результат влияют два показателя: количество баллов, которое удваивает шансы стать «хорошим» заемщиком, и значение шкалы, в котором достигается заданное отношение шансов «хороших» к «плохим» [5, с. *766*]

Приведение коэффициента логистической регрессии в скоринговый балл в линейной шкале предполагает осуществить преобразование по формуле

$$\text{балл} = -\left(WOE_j \cdot b_i + \frac{b_0}{n}\right) \cdot R + \frac{A}{n}, \tag{4}$$

где b_i — коэффициенты логистической регрессии для i-ой переменной; b_0 — константа; n — количество независимых переменных в уравнении регрессии; R — множитель; A — смещение [2, с. *10*].

Множитель определяется по формуле

$$R = \frac{D}{\ln(2)}, \tag{5}$$

где D — количество баллов, удваивающее шансы.

Смещение устанавливается по формуле:

$$A = B - R \cdot \ln C, \tag{6}$$

где B — значение на шкале баллов, в которой соотношение шансов C:1.

В результате масштабирования получена скоринговая карта (табл. 2).

Таблица 2. **Скоринговая карта, основанная на анкетных данных и запросах**

Характеристика	Категория	Балл
Возраст	До 28	110,92
	От 28 до 31	121,44
	От 31 до 37	131,34
	От 37 до 41	134,21
	От 41 до 50	138,87
	От 50 до 56	142,25
	От 56	266,42
Балл БКИ	До 122	90,88
	От 122 до 146	107,18
	От 146 до 162	116,08
	От 162 до 177	126,96
	От 177 до 190	131,8
	От 190 до 203	142,84
	От 203 до 217	154,33
	От 217 до 232	161,16
	От 232 до 252	168,74
	От 252	187,8
	Нет данных	0
Доход	До 903,8	118,9
	От 903,8 до 1278,9	126,02
	От 1278,9 до 1496,4	130,09
	От 1496,4 до 1709,5	136,2
	От 1709,5 до 2152,9	143,18
	От 2152,9 до 2984,2	153,85
	От 2984,2	185,18
	Нет данных	0
Семейное положение	Женат/замужем	144,38
	Холост/не замужем	124,46
	Вдова/вдовец	123,08
	В разводе	110,2
	Нет данных	0
Стаж	До 3 месяцев	125,2
	От 3 месяцев до 1 года	129,34
	От 1 года до 3 лет	134,64
	Более 3 лет	142,4
	Нет данных	0

Как уже было отмечено ранее, для масштабирования требуется определить количество баллов, удваивающее шансы стать «хорошим» заемщиком, и значение шкалы, в которой и достигается заданное отношение шансов «платежеспособных» к «неплатежеспособным». В банковской практике принято использовать систему, в которой каждые 40 баллов удваивают шансы стать «платежеспособным» клиентам, а также в точке 600 баллов отношение шансов составляет 72:1 [6].

По всем оценкам выявлено хорошее качество модели как для обучающей, так и для тестовой выборки. На обучающей выборке значение площади под *ROC*-кривой составило 0,72, что говорит о хорошем качестве модели (рис. 7).

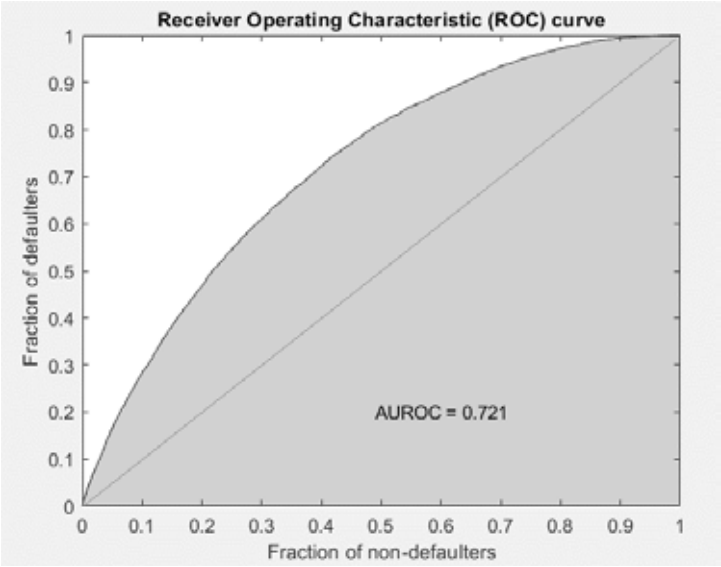


Рис. 7. ROC-кривая для обучающей выборки

В свою очередь для тестовой выборки данный показатель равен 0,695 (рис. 8). Это позволяет утверждать, что логит-регрессия корректно обучилась и относительно точно обобщает результаты. Схожие показатели, полученные на обеих выборках, — признак того, что в дальнейшем модель будет выдавать верные прогнозы (рис. 8.).

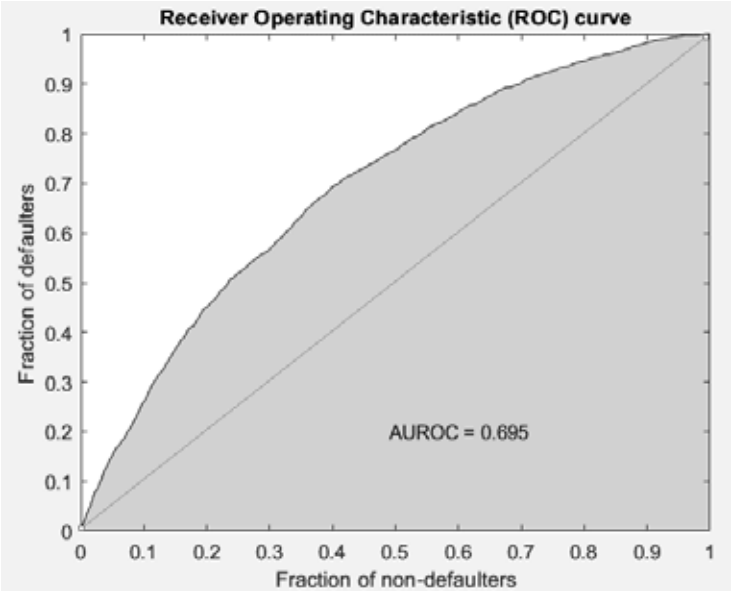


Рис. 8. ROC-кривая для тестовой выборки

Заключительным является процесс определения уровня отсечения, который состоит в установлении приемлемого для банка баланса между недополученной прибылью (отказы неверно классифицируемым «хорошим» клиентам) и прогнозируемыми потерями (верно определенные «плохие» клиенты). Критерий согласия Колмогорова — Смирнова демонстрирует, что наибольшая удаленность между функциями распределения «хороших» и «плохих» клиентов достигается в точке 671,9 балла — это и есть рекомендуемый порог отсечения (рис. 9).

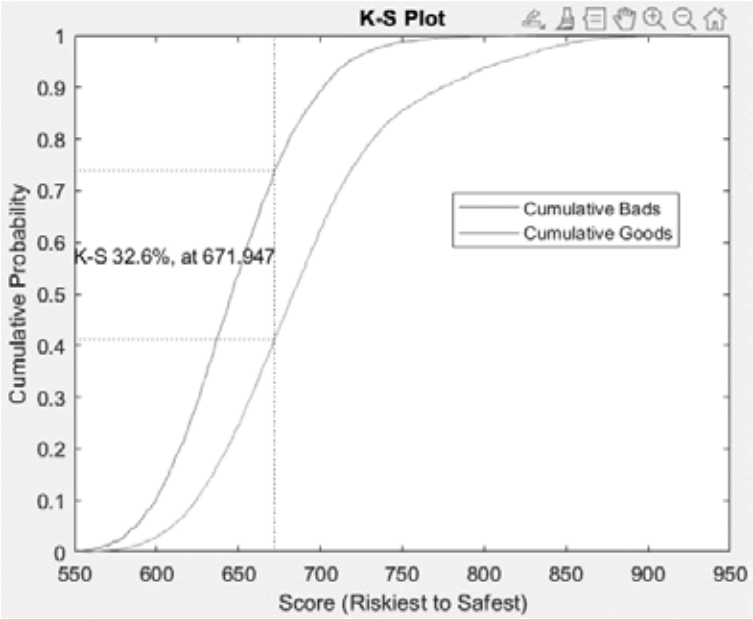


Рис. 9. Графическая интерпретация критерия согласия Колмогорова — Смирнова

Стоит отметить, что на практике определение оптимального уровня отсечения является очень трудоемким процессом, основанным на исторических данных о доходности, дефолтности и уровне одобрения заходящих заявок. В данной статье определен оптимальный уровень отсечения, предполагающий допустимую дефолтность, а именно 6,64 % в случае неверно идентифицированных неплатежеспособных клиентов.

Заключение. В скоринговой модели оценке подлежат заемщики, переменными служат показатели их кредитоспособности, классами являются разбиваемые на две классификационные группы заемщики, отвечающие состояниям отсутствия дефолта и наступления дефолта соответственно. Для большого количества заемщиков, составляющих несколько десятков тысяч, задачу их бинарной классификации целесообразно решать методами машинного обучения.

При построении уравнения логистической регрессии значимыми переменными выбраны возраст заемщика, его банковская кредитная история, доход, семейное положение и стаж работы. Отсутствие мультиколлинеарности экзогенных переменных подтверждается близкими к единице значениями коэффициента «вздутия» дисперсии.

Построение скоринговой карты заемщиков предполагает перевод количественных переменных в категориальные с помощью формул преобразования коэффициентов уравнения логистической регрессии в скоринговые баллы по выбранной шкале от 0 до 1 000 баллов.

Оценка кредитоспособности заемщиков проведена по выборочной совокупности заемщиков с численностью 16 898 человека для одного из минских банков. Установлен оптимальный уровень отсечения заемщиков по прогностическому состоянию наступления дефолта ниже суммарных 672 скоринговых баллов по выбранным пяти показателям, что отвечает допустимому уровню дефолта 6,64 % потенциальных кредитополучателей.

Литература и электронные публикации в Интернете

1. *Читая, Г. О.* Оценка кредитоспособности заемщиков коммерческого банка методами прикладного статистического анализа / Г. О. Читая // Науч. тр. Белорус. гос. экон. ун-та. — Минск, 2019. — Вып. 12. — С. 463—475.

Chitaja, G. O. Ocenka kreditosposobnosti zaemshhikov kommercheskogo banka metodami prikladnogo statisticheskogo analiza [Assessment of The Creditworthiness of Commercial Bank Borrowers Using of Applied Statistics] / G. O. Chitaja // Nauch. tr. Belorus. gos. jekon. un-ta. — Minsk, 2019. — Vyp. 12. — P. 463—475.

2. *Читая, Г. О.* Оптимизация структуры задолженности заемщиков банка в среднесрочной перспективе / Г. О. Читая // Бизнес. Инновации. Экономика : сб. науч. ст. — Минск : Ин-т бизнеса БГУ, 2023. — Вып. 7. — С. 130—137.

Chitaja, G. O. Optimizacija struktury zadolzhennosti zaemshhikov banka v srednesrochnoj perspektive [Optimization of Debt Structure of The Bank's Borrowers In The Medium-Term Dynamics] / G. O. Chitaja // Biznes. Innovacii. Jekonomika : sb. nauch. st. — Minsk : In-t biznesa BGU, 2023. — Vyp. 7. — P. 130—137.

3. *Altman, E.* Default Recovery Rates in Credit Risk Modeling: A Re-Уview of the Literature and Empirical Evidence/ E. Altman, A. Resti, A. Sironi // Economic Notes by Banca Monte dei Paschi di Siena SpA. — 2004. — Vol. 33, N 2. — 183 p.

4. *Сорокин, А. С.* Построение скоринговых карт с использованием модели логистической регрессии / А. С. Сорокин // Науковедение. — 2014. — № 25. — 13 с.

Sorokin, A. S. Postroenie skoringovyh kart s ispol'zovaniem modeli logisticheskoj regressii [Construction of Scoring Cards Using a Logistic Regression Model] / A. S. Sorokin // Naukovedenie. — 2014. — N 25. — 13 p.

5. *Carol, A.* The Professional Risk Managers' Handbook / A. Carol, E. Sheedy. — PRM, 2004. — 1360 p.

6. Инструкция по применению Поведенческого скоринга Кредитного регистра Национального банка Республики Беларусь [Электронный ресурс]. — Режим доступа: https://nbrb.by/today/creditregistry. — Дата доступа 10.12.2023.

GIGLA CHITAYA,
DARYA KAZUNOVA

SCORING MODEL FOR ASSESSING CREDIT RISKS

Authors affiliation. *Gigla CHITAYA* (chitaya_g@bseu.by), *Belarus State Economic University (Minsk, Belarus); Darya KAZUNOVA* (Kazyn810@gmail.com), *Belarus State Economic University (Minsk, Belarus).*

Abstract. The article presents a model for assessing the credit risk of a retail borrower within the framework of application scoring. The methodological support for constructing a scoring card based on a reasonable set of indicators and determining the optimal level of scoring points used to screen out a risky group of potential borrowers is considered. Borrowers are classified based on the developed econometric logistic regression model and the use of machine learning methods. The numerical implementation of the mathematical model was carried out on data from a sample population of borrowers from one of the Minsk banks using the MatLab software product.

Keywords: credit risk; borrowers; scoring; probability of default; machine learning; logistic regression; scoring card; creditworthiness indicators.

UDC 336.717:005.334 (476)

Статья поступила
в редакцию 12. 12. 2023 г.