УДК 33.061.067.21

DOI: 10.21779/2542-0321-2020-35-2-18–26

П.М. Симонов, Е.В. Збоев

Скоринговая оценка кредитоспособности физических лиц

Пермский государственный национальный исследовательский университет; Россия, 614990, г. Пермь, ул. Букирева, 15; simpm@mail.ru, egor7796@yandex.ru

Эффективность применения скоринговых систем в банковской сфере напрямую зависит от возможности внесения изменений и перенастройки используемой модели, а также от периодических проверок качества работы системы. Изменение закрытых западных систем, используемых во многих российских банках, требует больших финансовых и временных затрат. По этой причине использование российских аналогов должно сократить издержки на внедрение систем оценки кредитоспособности. Однако на сегодняшний день решения, представленные разработчиками из стран СНГ, уступают западным аналогам. Данный факт обуславливает актуальность реализации моделей оценки кредитоспособности на основе алгоритмов случайного леса и градиентного бустинга, которые по своей эффективности превосходят распространенные логит-модели.

Ключевые слова: *скоринг, случайный лес, градиентный бустинг, логит-модель.*

Введение

Оценка кредитоспособности заемщика является одним из важнейших процессов при принятии решений по управлению кредитами в банках, т. к. минимизация кредитного риска зачастую является основным фактором, определяющим эффективность работы банковской организации. Под кредитоспособностью физического лица понимается оценка вероятности своевременной выплаты долга заемщиком, основанная на изучении кредитной истории, а также на анализе текущего финансового положения клиента. Кредитоспособность может быть представлена формализованным показателем, таким, как кредитный рейтинг или скоринговая оценка.

Помимо оценки эксперта процесс оценки кредитоспособности включает в себя множество статистических и математических методов, позволяющих определить вероятность своевременного погашения кредита заемщиком, он облегчает анализ и классификацию необходимой информации, определяет факторы и показатели, отражающие кредитоспособность физического лица.

В последнее время в России можно наблюдать рост рынка услуг кредитования, который включает в себя кредитование физических лиц. Увеличение этого рынка ведет к повышению уровня кредитных рисков не только для отдельных организаций, но и для всей банковской системы страны в целом.

Так как при выдаче кредита физическому лицу банк в первую очередь заинтересован в оценке кредитоспособности заемщика, существует острая необходимость автоматизации и оптимизации традиционного процесса кредитования, при котором реше-

ние об одобрении кредита или отказе в нем принимается кредитным экспертом. Практика применения банками систем оценки кредитоспособности показывает, что при сравнительно небольших суммах займов скоринговым системам отводятся значительно больше полномочий в процессе принятия решения, чем при более высоких, где скоринговая оценка используется больше как фактор «поддержки», который учитывается кредитным экспертом.

Особое внимание уделяется вопросу оценки кредитоспособности физического лица и в научной литературе. Так, анализу кредитоспособности на основе различных факторов посвящены работы следующих авторов: Лаврушин О.И. с соавторами [1–3], Ильясов С.М. [4], Пещанская И.В. [5], Абалакин А.А., Соболева Е.С. и Османова А.Э. [6], Ворошилова И.В. и Сурина И.В. [7], Рыкова И.Н. [8], Ленская Н.В. и Чернышева Т.Ю. [9]. Также теоретические аспекты кредитного скоринга рассмотрены в работах Алёшина В.А. и Рудаевой О.О. [10], Самойловой С.С. и Курочка М.А. [11], Яковлевой А.Ю. [12], Данилович В.Ю. и Курганской Г.С. [13].

На текущий момент разработано большое количество методик оценки кредитоспособности потенциального заемщика, среди которых авторские методики построения скоринговых систем. В работах данных авторов применяются такие методы, как логистическая регрессия, нейронные сети, кластерный анализ и т. д. Однако в большинстве публикаций преобладают логит-модели, построение которых выполнено в таких программных продуктах, как Matlab, Statistica Scorecard и Deductor Academic. В перечисленных программах реализована только малая часть статистических методов, вследствие чего построенные модели не всегда обладают высокой предсказательной мощностью, что в конечном итоге может привести к повышению банковского кредитного риска.

Таким образом, помимо анализа и сравнения эффективности распространенных алгоритмов бинарной классификации, используемых при создании моделей оценки кредитоспособности, стоит также рассмотреть алгоритмы, не получившие широкого распространения в скоринговых системах. Указанное обстоятельство обуславливает актуальность выбранной темы статьи и определяет объект, предмет исследования, его основную цель и задачи.

Объектом исследования являются модели оценки кредитоспособности физических лиц.

Предметом исследования является совокупность методических, теоретических и практических аспектов, связанных с моделированием оценки кредитоспособности физических лиц.

В соответствии с вышеуказанной целью были сформулированы следующие задачи:

1) на основе распространенных алгоритмов машинного обучения реализовать модели оценки кредитоспособности на языке программирования Python;

2) проверить гипотезу о высокой эффективности ансамблевых методов в задаче предсказания дефолта заемщика.

Реализация ансамблевых методов на языке программирования Python определяет новизну данного исследования, т. к. на текущий момент Python широко применяется в сфере data mining (добыча данных, глубинный анализ данных) и имеет ряд существенных преимуществ перед языком программирования *R*, который в большей степени

предназначен для статистической обработки данных и работы с графикой, в то время как *Python* позволяет значительно проще реализовать готовое программное решение.

Выбор спецификаций и обучение моделей

В качестве исходной выборки были использованы данные о заемщиках в обезличенном виде (без указания персональных данных, страны и валюты кредита), опубликованные банком «Тинькофф Банк» в качестве исходных данных для чемпионата, целью которого являлся поиск зависимости между анкетными показателями клиентов и фактом невыполнения обязательств по кредитному договору [14]. Размерность исходной выборки – 205296 записей.

Для построения моделей оценки кредитоспособности физических лиц был определен следующий перечень алгоритмов: алгоритм случайного леса; градиентный бустинг; логистическая регрессия; метод ближайших соседей; наивный байесовский классификатор.

Реализации всех выбранных алгоритмов представлены в библиотеке алгоритмов машинного обучения Scikit-learn для языка Python версии 2.7 и выше.

Большинство функций в библиотеке Scikit-learn не требует детальной настройки всех параметров, так как заданные значения по умолчанию обеспечивают оптимальные результаты работы алгоритмов для разных задач.

Функция RandomForestClassifier (), реализующая алгоритм случайного леса, принимает на вход следующие параметры: *n*_estimators – число решающих деревьев в ансамбле (165 для текущей модели); max_depth – максимальная глубина дерева (4 для текущей модели); criterion – критерий качества разбиения (индекс Джини для текущей модели).

Индекс Джини определяется как [15]:

$$Gini\ (T) = 1 - \sum_{i=1}^{n} p_i^2,$$

где p_i – вероятность (относительная частота) класса *i* в *T*.

Качество разбиения оценивается путем минимизации значения индекса Джини:

$$Gini_{split}(T) = \frac{N_1}{N} Gini(T_1) + \frac{N_2}{N} Gini(T_2) \rightarrow \min,$$

где T – набор данных; T_1 и T_2 – две части набора *T*; N – количество примеров в наборе *T*; N_1 и N_2 – количество примеров в наборах T_1 и T_2 соответственно.

После преобразования формула для конкретного дерева решений имеет следующий вид:

$$Gini_{split} = \frac{1}{L} \cdot \sum_{i=1}^{n} l_i^2 + \frac{1}{R} \cdot \sum_{i=1}^{n} r_i^2 \rightarrow \max,$$

где L, R – число примеров соответственно в левом и правом потомке; l_i и r_i – число экземпляров *i*-того класса в левом/правом потомке.

В конечном итоге лучшим будет то разбиение, для которого величина $Gini_{split}$ максимальна [14].

Воспользуемся функцией export_graphviz() для визуализации случайно выбранного дерева решений из ансамбля.

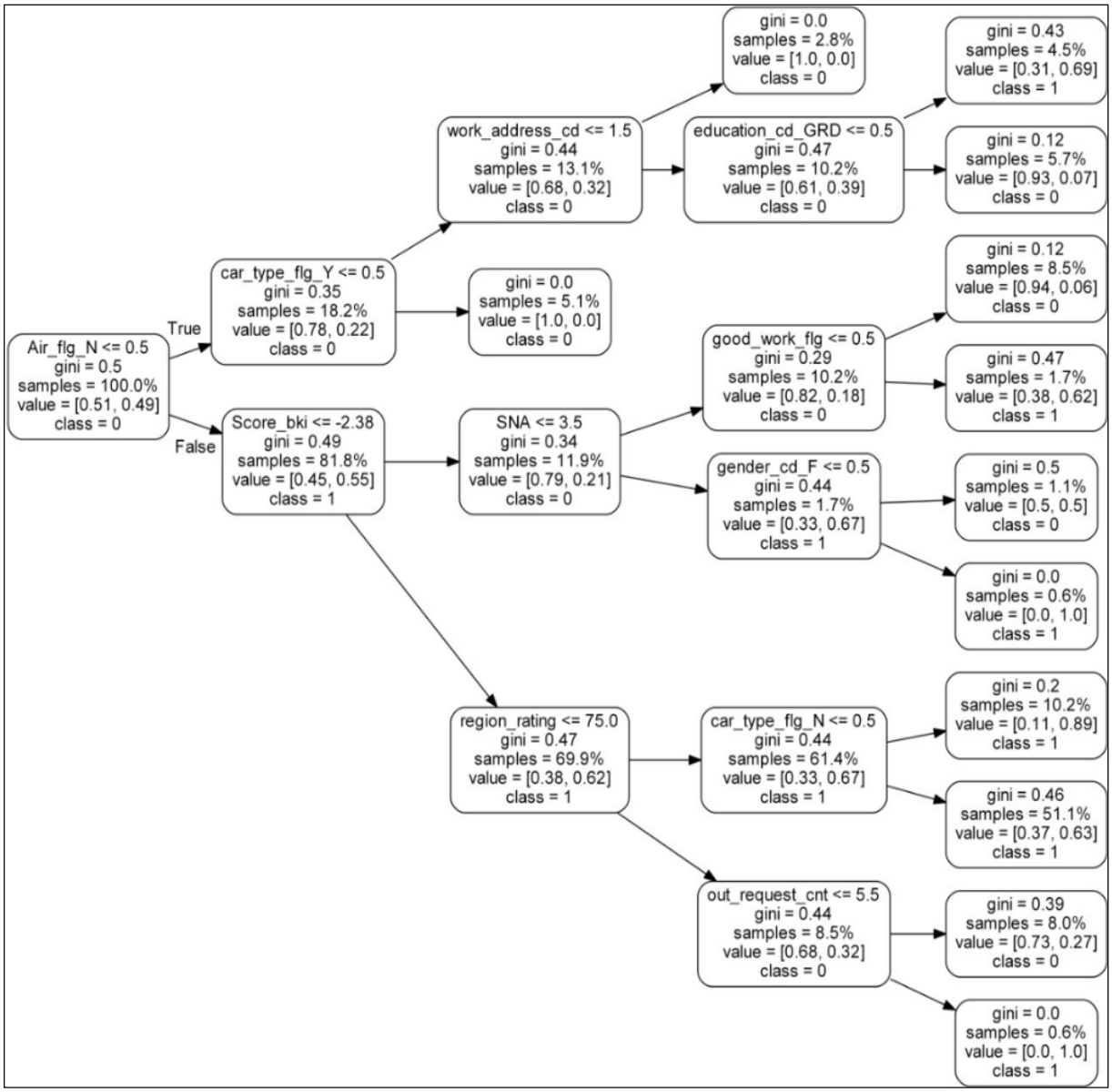

Рис. 1. Случайное дерево решений из ансамбля

Полученное дерево (рис. 1) состоит из 10 терминальных узлов (листьев).

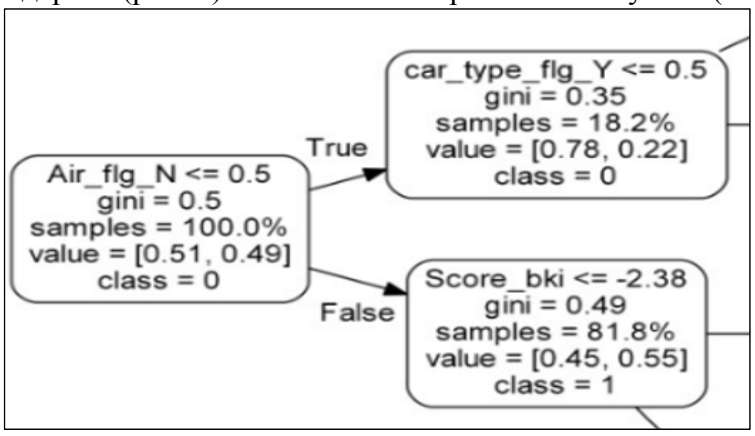

Рис. 2. Графическое представление узлов дерева

Графическое представление дерева описывается набором узлов (рис. 2), в которых отображаются: условие разбиения, коэффициент Джини, процент исходной выборки, доля наблюдений каждого из классов и прогнозируемый класс.

Алгоритм градиентного бустинга реализован в функции GradientBoostingClassifier(): GradientBoostingClassifier(max_depth=4).

Как и у алгоритма случайного леса, значение параметра max_depth, указывающего максимальную глубину дерева, равно 4. Остальным параметрам функции присвоены значения по умолчанию.

У функций LogisticRegression() и GaussianNB(), реализующих метод логистической регрессии и модель наивного байесовского классификатора, параметрами являются значения по умолчанию.

После настройки функций производится обучение полученных моделей на тренировочных выборках. Для этого используется функция model.fit(), принимающая в качестве параметров выборку, содержащую значения независимых переменных, и выборку, содержащую значения зависимой переменной: model.fit(TRNtrain, TARtrain).

Сравнительный анализ работы алгоритмов и описание результатов

Для представления результатов бинарной классификации была использована мера AUC.

AUC (area under ROC curve) является количественным показателем площади под ROC-кривой. Теоретически AUC может принимать значения от 0 до 1. Однако обычно подразумевается изменение значений в пределах от 0,5 («бесполезный» классификатор, т. е. полная неразличимость двух классов) до 1 («идеальная» модель), так как модель всегда характеризуется кривой, расположенной выше прямой $y = x$. ROC-кривая отображает соотношение долей верных положительных классификаций от общего числа положительных значений (*TPR*, true positive rate) и долей ошибочных положительных классификаций от общего числа отрицательных значений (*FPR*, false positive rate) [16].

Рассмотрим два понятия:

1) специфичность (Specificity) – доля истинноотрицательных случаев:

$$Specificity = \frac{TN}{TN+FP},$$

где *TN* – количество истинноотрицательных случаев; *FP* – количество ложноположительных случаев;

2) чувствительность (Sensitivity) – доля истинноположительных случаев:

$$Sensitivity = \frac{TP}{TP+FN},$$

где *TP* – количество истинноположительные случаев; *FN* – количество ложноотрицательных случаев.

Показатель *FPR* и специфичность связаны следующим выражением:

$$FPR = 1 - Specificity,$$

где *FPR* – доля ложноположительных случаев от общего числа отрицательных значений.

В контексте текущей задачи под чувствительностью понимается доля одобренных запросов на кредит от благонадежных заемщиков, под специфичностью – доля откло-

ненных запросов от заемщиков, которые были классифицированы как неблагонадежные.

Для построения ROC-кривой на графике по оси ординат откладываются значения *TPR* (или же чувствительности классификатора), а по оси абсцисс – значения показателя *FPR* (который связан с величиной специфичности по формуле). Таким образом, ROC-кривая лучшей модели имеет наиболее сильный изгиб в сторону верхнего левого угла графика, где доля истинно-положительных случаев равна единице, а доля ошибочных положительных классификаций равна нулю [16].

Построим ROC-кривые для рассматриваемых алгоритмов (рис. 3).

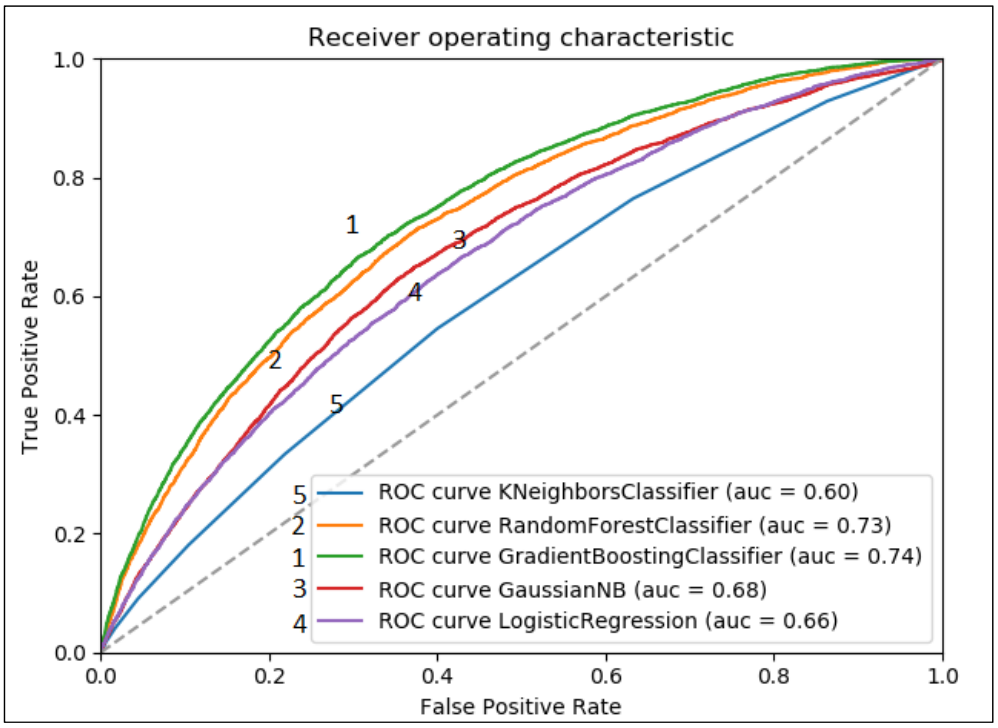

Рис. 3. ROC-кривые и меры AUC для построенных моделей

Помимо ROC-кривых на графике также вычислены значения показателя AUC для каждой модели. Самое низкое значение критерия AUC у алгоритма, реализующего метод ближайших соседей (60 %). Распространенная среди скоринговых систем логистическая регрессия имеет значение AUC 66 %, что немного ниже качества модели наивного байесовского классификатора.

Самые лучшие результаты показали модели случайного леса (73 %) и градиентного бустинга (74 %).

Для определения степени влияния факторов моделей случайного леса и градиентного бустинга на результат прогнозирования целевой переменной воспользуемся функцией model.feature_importances_.

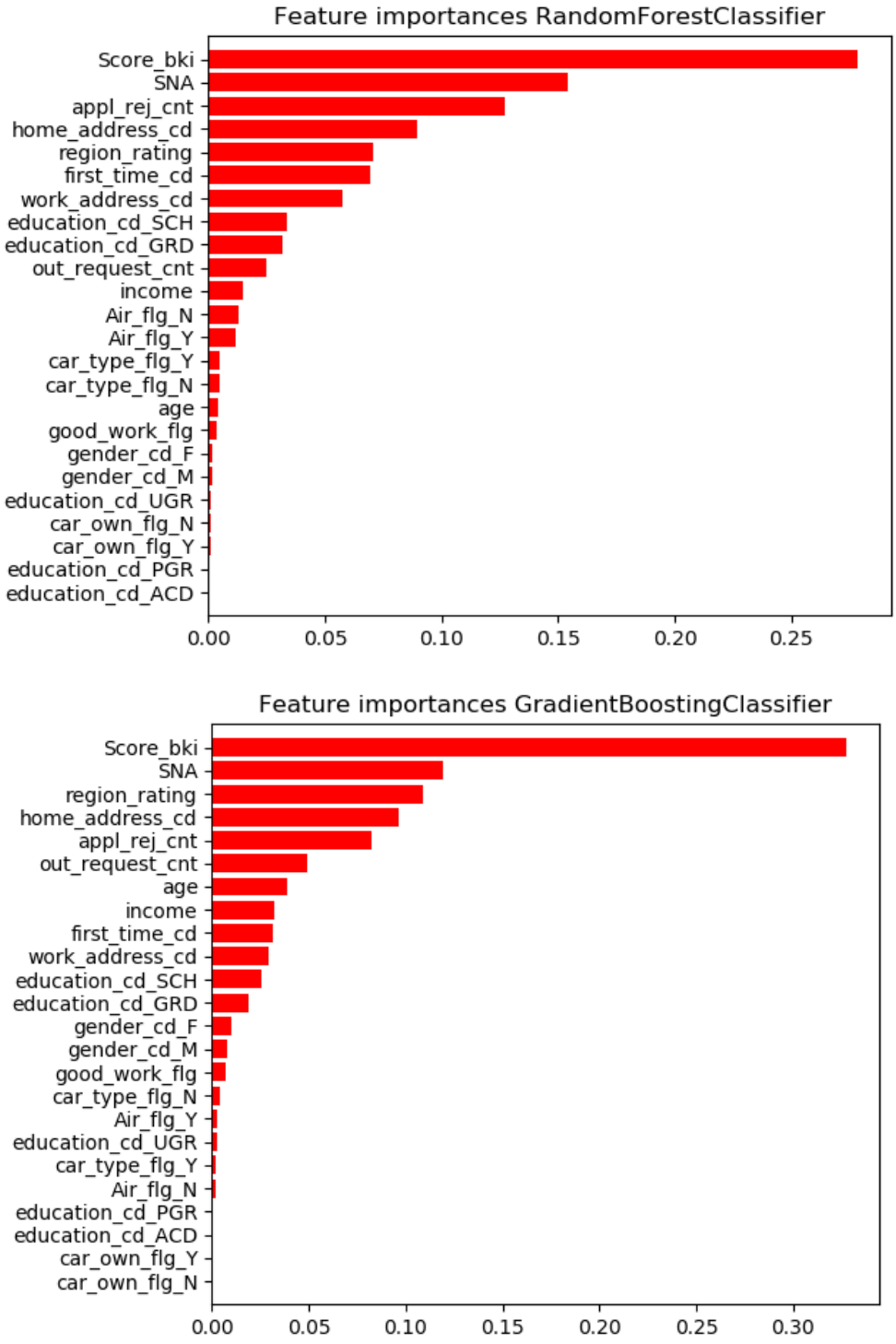

Рис. 4. Степень влияния факторов на целевую переменную

В полученных моделях случайного леса и градиентного бустинга на оценку кредитоспособности наибольшее влияние оказывают следующие характеристики потенциального заемщика (рис. 4): скоринговый балл (по данным из бюро кредитных историй), связь заявителя с клиентами, категоризатор домашнего адреса, число отказов в кредитовании и рейтинг региона проживания заемщика.

Таким образом, в данном исследовании было доказано, что при создании моделей оценки кредитоспособности физического лица стоит принять во внимание алгоритмы случайного леса и градиентного бустинга, которые показывают более высокие результаты классификации физических лиц на благонадежных и неблагонадежных заемщиков, чем широко распространенная в скоринговых системах логит-модель.

Литература

1. *Лаврушин О.И., Афанасьева О.Н., Корниенко С.Л.* Банковское дело: современная система кредитования. – М.: КноРус, 2007. – С. 39–45.

2. *Лаврушин О.И., Валенцева Н.И.* Банковское дело: учебник. – М.: КноРус, 2016. – С. 353–360.

3. *Лаврушин О.И.* Банковское дело: экспресс-курс. – М.: КноРус, 2009. – С. 164–173.

4. *Ильясов С.М.* Об оценке кредитоспособности банковского заемщика // Деньги и кредит. – 2011. – № 9. – С. 28–34.

5. *Пещанская И.В*. Организация деятельности коммерческого банка: учеб. пособие. – М.: ИНФРА-М, 2001. – С. 146–154.

6. *Абалакин А.А., Соболева Е.С., Османова А.Э*. Оценка кредитоспособности физических лиц на основе современных банковских технологий // Интернет-журнал «Науковедение». – 2015. – Т. 7, № 5 (30). – С. 1–7.

7. *Ворошилова И.В., Сурина И.В.* К вопросу о совершенствовании механизма оценки кредитоспособности индивидуальных заемщиков // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. – 2005. – № 16. – С. 2–11.

8. *Рыкова И.Н*. Методика оценки кредитоспособности заемщиков // Банковское кредитование. – 2015. – № 6. – С. 7–9.

9. *Ленская Н.В., Чернышева Т.Ю.* Методы оценки кредитоспособности заемщика банком // Современные технологии поддержки принятия решений в экономике: сборник трудов Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых (г. Юрга, 28–29 апреля 2014 г.). – Томск: ТПУ, 2014. – С. 73–75.

10. *Алёшин В.А., Рудаева О.О.* Кредитный скоринг как инструмент повышения качества банковского риск-менеджмента в современных условиях // Пространство экономики. – 2012. – № 2–3. – С. 27–30.

11. *Самойлова С.С., Курочка М.А.* Скоринговые модели оценки кредитного риска // Социально-экономические явления и процессы. – 2014. – Т. 9, № 3. – С. 99–102.

12. *Яковлева А.Ю.* Анализ скоринговой модели оценки кредитоспособности заемщиков // Научный альманах. Экономические науки. – 2018. – № 6–1(44). – С. 119–121.

13. *Данилович В.Ю., Курганская Г.С*. Скоринговые модели как средство управления кредитными рисками в российских банках // Бизнес-образование в экономике знаний. – 2017. – № 1. – С. 29–32.

14. Кредитный скоринг. Режим доступа: https://www.kaggle.com/c/credit-scoring/data (дата обращения: 29 апреля 2019).

15. Деревья решений. Режим доступа: https://basegroup.ru/community/articles/math-cart-part1 (дата обращения: 7 мая 2019).

16. ROC-анализ. Режим доступа: https://basegroup.ru/community/articles/logistic (дата обращения: 16 мая 2019).

Поступила в редакцию 9 октября 2019 г.

UDC 33.061.067.21

DOI: 10.21779/2542-0321-2020-35-2-18–26

Scoring Credit Rating of Individuals

P.M. Simonov, E.V. Zboev

Perm State National Research University; Russia, 614990, Perm, Bukirev st., 15; simpm@mail.ru, egor7796@yandex.ru

The effective strength of scoring systems in the banking sector directly depends on the possibility of making changes and reconfiguring of the model used, as well as on the periodic checks of the quality of the system. The transformation of the closed Western systems used in many Russian banks is both costly and time-consuming. For this reason, the use of Russian counterparts should reduce the costs of introducing credit rating systems. However, today, the solutions presented by the developers from the CIS countries are inferior to their Western counterparts. This fact determines the relevance of the implementation of models for assessing creditworthiness based on random forest algorithms and gradient boosting, which are superior in efficiency to common logit-models.

Keywords: *scoring, random forest, gradient boosting, logit-model.*

Received 9 October 2019