УДК 004.89
DOI: 10.17223/19988648/44/12

А.Л. Богданов, И.С. Дуля

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ В РЕШЕНИИ ЗАДАЧИ КРЕДИТНОГО СКОРИНГА

В статье рассматривается применение нейронных сетей в решении задачи кредитного скоринга и демонстрируется эффективность их применения для построения модели оценки рейтинга кредитоспособности заемщика на основе реальных данных платформы взаимного кредитования Lending Club. Показано применение многослойного персептрона в решении задачи скоринга заявителя. В ходе работы была осуществлена подготовка данных, построена модель классификации заемщиков и проведена оценка ее точности.
Ключевые слова: *кредитный скоринг, классификация, многослойный персептрон, нейронная сеть, машинное обучение, анализ данных.*

По данным Росстата [1], за последние четыре года в банковском секторе объем размещенных денежных средств вырос с 38767,9 до 52816 млрд руб., или на 36%. Помимо роста объемов кредитования, наблюдался значительный рост конкуренции между финансовыми организациями, проявляющийся в предложении новых продуктов, увеличении скорости обработки запросов, заключении сделок и т.д. Одной из нетривиальных проблем финансовых организаций является кредитный скоринг – задача оценки кредитоспособности заемщика по имеющимся данным о самом заемщике и исторических сведениях о поведении других заемщиков. Имея такой инструмент, банки могут значительно повысить отдачу от своих вложений и свести к минимуму риск финансовых операций. Одним из подходов к решению таких задач являются искусственные нейронные сети.

Постановка задачи

Объектом данного исследования является решение задачи кредитного скоринга с помощью аппарата нейронных сетей, целью – построение и реализация модели, способной выявить заемщиков, которые с высокой долей вероятности выполнят условия договора. В ходе решения данной задачи были использованы реальные данные, полученные с платформы p2p-кредитования Lending Club [2] – места встречи кредиторов и заемщиков, являющихся физическими лицами. Данный выбор был обусловлен тем, что данные находятся в свободном доступе, в отличие от банковских данных. Результаты работы могут быть применены к банковскому скорингу после соответствующей корректировки.

Кредитный скоринг

Под *кредитным скорингом* (*credit scoring*) понимают процедуру оценки вероятности банкротства потенциального заемщика при рассмотрении возможности его кредитования. По сути скоринг представляет собой инструмент классификации потенциальных заемщиков на различные группы по уровню кредитоспособности [3]. В основе скоринга лежит математическая модель, основывающаяся на целостной системе показателей, по значениям которых принимается решение об отнесении заемщика к определенному классу, отражающему уровень риска его банкротства.

Скоринговые модели применяются банковскими организациями для кредитования как физических, так и юридических лиц. На практике большее распространение получило потребительское кредитование – кредитование физических лиц на небольшие суммы. Обычно данный вид кредитования осуществляется в сжатые сроки, поэтому имеется необходимость в методах оценки кредитоспособности заемщиков, способных быстро и с высокой точностью выполнить данную задачу.

В настоящий момент в международной банковской практике принято выделять несколько разновидностей кредитного скоринга (табл. 1) [4]. Особый интерес представляют скоринговые модели оценки заемщика и его поведения. Технологии построения данных моделей являются практически идентичными. Остальные виды скоринга опираются на специализированные методики. В терминах табл. 1 цель данной работы может быть сформулирована как построение скоринговой модели заявителя.

В банковской практике предварительная скоринговая оценка и поведенческая скоринговая оценка обычно проводятся по одной и той же модели.

Таблица 1. **Виды кредитного скоринга**

Вид кредитного скоринга	Направление деятельности
Скоринг заявителя (Application scoring)	Определение уровня кредитоспособности заявителя
Поведенческий скоринг (Behavioral scoring)	Определение уровня риска существующих должников на основе имеющихся данных о их поведении
Скоринг по работе с просроченной задолженностью (Collection)	Определение методов воздействия в отношении неплательщиков
Скоринг мошенничества (Fraud scoring)	Оценка вероятности того, что новый клиент является мошенником
Скоринг отклика (Response scoring)	Оценивание возможной реакции потребителя на направленное ему предложение
Скоринг потерь (Attrition scoring)	Прогнозирование оттока клиентов

Кредитный скоринг является основой для решения возникающих задач по управлению кредитной деятельностью в банковской организации. При этом решаются следующие задачи: создание модели принятия решений о выдаче кредита, максимизация эффективности взаимодействия с клиентом,

построение централизованной системы управления кредитной политикой, а также прогнозирование качества кредитных активов банка.

Очистка и предобработка данных

В качестве исходных использовались данные компании Lending Club по ссудам, выданным за 2014 г. Указанный период был выбран по причине того, что у большинства ссуд, выданных позднее, еще не наступил срок погашения. Статистическая совокупность содержала 235 629 записей о выданных ссудах. Каждая запись включала 135 признаков (134 входных и один целевой – оценку заемщика). Признаки были представлены как числовыми, так и атрибутивными (нечисловыми) значениями. Подробное описание признаков доступно по адресу https://github.com/dulyaivan/credit_scoring/description.pdf.

Первичный анализ показал, что загруженные данные требуют предобработки, так как содержат многочисленные пропуски, пустые признаки, а атрибутивные признаки представлены в строчном формате. С этой целью были проделаны следующие шаги:

Во-первых, были удалены 29 признаков, которые не содержали ни одного значения, и признаки, которыми не мог располагать инвестор на момент рассмотрения заявки. Описательная статистика оставшихся признаков доступна по адресу https://github.com/dulyaivan/credit_scoring/blob/master/statistics.pdf. По указанной ссылке представлено описание каждого признака в формате: тип переменной, количество пропусков, доля пропусков, количество уникальных элементов, доля уникальных элементов, среднее значение, медиана, стандартное отклонение, максимум и минимум.

Во-вторых, были удалены признаки, доля пропусков в которых составляла более 70%, и атрибутивные признаки, количество уникальных элементов которых было более 300. Данное решение было продиктовано тем, что менее строгое правило приводило к существенному увеличению объема данных и сложностям в процессе обучения нейронной сети.

В-третьих, был отделен от общей совокупности интересующий признак grade, представляющий класс кредитоспособности заемщика, и удален признак sub_grade, являющийся уточнением к признаку grade. Также был удален признак loan_status (статус ссуды), имеющий прямую связь с целевым признаком. В результате выполненных шагов были получены матрица входных переменных и вектор-столбец значений выходного признака.

В полученной матрице входных переменных все значения были переведены в числовой формат, атрибутивные признаки переведены в булево пространство, причем отсутствие данных интерпретировалось как дополнительное значение признака. Целевой признак grade, принимающий значения A, B, C, D, E, F и G, был представлен в числовом формате. Пропуски в количественных признаках были заполнены выборочным средним соответствующего признака.

На последнем шаге была выполнена стандартизация значений числовых признаков, так как это является важным условием успешного обучения нейронной сети. В результате среднее значение каждого числового признака стало равно 0, а стандартное отклонение – 1.

Построение модели

Для решения задачи классификации заемщиков на классы по платежеспособности (A, B, C, D, E, F и G) была использована нейронная сеть. *Нейронная сеть* – это распределенный параллельный процессор, состоящий из элементарных единиц обработки информации, накапливающих экспериментальные знания и предоставляющих их для последующей обработки [5. С. 30]. Искусственные нейронные сети представляют собой обширный класс методов машинного обучения, позволяющих эффективно решать задачи классификации, кластеризации и регрессии. Объединяющей основой данных методов является то, что алгоритмы из данного класса построены по аналогии с мозгом живого существа.

При построении нейронной сети необходимо определить ее параметры (скорость обучения сети, количество эпох обучения, размер батчи (batch), значение параметра регуляризации), выбрать количество скрытых слоев и количество нейронов в них, произвести настройку весов нейронов. Построение нейросетевых алгоритмов не предполагает четких требований при выборе конфигурации и архитектуры сети, зачастую выбор конкретной структуры и настройка параметров осуществляются на основании многочисленных тестов, в результате которых делается оптимальный выбор.

Для решения задачи классификации был выбран *многослойный персептрон* (multilayer perceptron, MLP) с двумя скрытыми слоями. *Многослойным персептроном* называется любая нейронная сеть прямого распространения, которая имеет хотя бы один скрытый слой. На рис. 1 представлена архитектура многослойного персептрона, имеющего два скрытых слоя.

Обозначим веса как w_{ij}^k, где i – номер нейрона из предыдущего слоя; j – номер нейрона, на который поступает сигнал, а k – номер слоя.

Сеть работает следующим образом: на вход поступает входной вектор $\bar{x} = [x_1, x_2, \dots, x_n]^T$, а сеть формирует отклик в виде выходного вектора $\bar{y} = [y_1, y_2, \dots, y_m]^T$. Каждый нейрон вычисляет сумму входных значений, умноженных на соответствующие веса, а затем передает значение сумматора на активационную функцию, т.е.

$$u_j^k = \sum_{i=1}^{n} w_{ij}^k x_i + b_j^k, \quad (1)$$

$$y_j^k = \varphi(u_j^k), \quad (2)$$

где u_j^k – значение сумматора j-го нейрона в k-м слое; y_j^k – соответствующий выход данного нейрона; x_i – выходное значение i-го нейрона в предыдущем слое; b_j^k – порог i-го нейрона в k-м слое; $\varphi(.)$ – активационная функция.

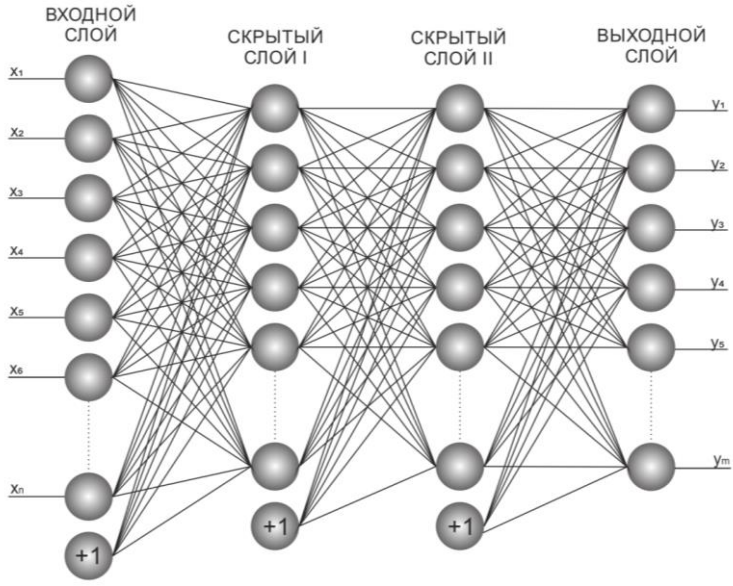

Рис. 1. Многослойный персептрон с двумя скрытыми слоями

Нейронная сеть осуществляет нелинейное преобразование входного вектора $\bar{x} = [x_1, x_2, \dots, x_n]^T$ в выходной вектор $\bar{y} = [y_1, y_2, \dots, y_m]^T$, при этом данное преобразование зависит от значений свободных параметров сети (весов синаптических связей и смещений). Для того чтобы сеть могла решать поставленную задачу, необходимо настроить эти параметры так, чтобы выходные значения соответствовали желаемому результату. Процесс настройки свободных параметров нейронной сети называется *обучением*.

В качестве метода обучения использовался *метод обратного распространения ошибки*, являющийся наиболее популярным методом обучения многослойных нейронных сетей. Метод обратного распространения ошибки позволяет вычислять ошибки нейронов в скрытых слоях, что является ключевой проблемой при построении многослойных нейронных сетей. В методе обратного распространения ошибки применяется следующий прием: ошибка скрытого нейрона рассчитывается как вклад этого нейрона в ошибку на следующем слое, отсюда и название – алгоритм обратного распространения ошибки. То есть сначала считается ошибка на выходном слое, а затем слой за слоем вычисляются ошибки для скрытых нейронов.

Согласно методу ошибки в скрытых и выходных слоях рассчитываются по следующим формулам [6. С. 510]:

$$\delta_j^{hidden} = \varphi'(Z_j) \sum_k \delta_k w_{jk}, \quad (3)$$

$$\delta_j^{out} = \varphi'(Z_j)\,(y_j - a_j), \quad (4)$$

где δ_j^{hidden} – ошибка j-го нейрона в скрытом слое; δ_j^{out} – ошибка j-го нейрона в выходном слое; $\varphi'(.)$ – производная активационной функции;

Z_j – сумма произведений входных значений j-го нейрона и соответствующих весов; δ_k – ошибка k-го нейрона; w_{jk} – вес, соединяющий j-й нейрон и k-й нейрон из следующего слоя; a_j – предсказанное значение на j-ом нейроне, y_j – соответствующее фактическое значение; k – номер нейрона в слое, по которому рассчитывается ошибка нейрона в предыдущем слое.

Величина коррекции w_{ij} нейрона определяется следующим образом:

$$\Delta w_{ij} = \lambda \delta_j x_i, \quad (5)$$

где x_i – i-е входное значение j-го нейрона; δ_j – величина ошибки j-го нейрона; λ – параметр скорости обучения.

В качестве активационных функций на выходном слое была использована функция Softmax, которая на выходе позволяет получить вектор с вероятностями принадлежности записи к конкретному классу. Функция Softmax представляет собой обобщение логистической функции для многомерного случая. Листинг описанного многослойного персептрона доступен по адресу https://github.com/dulyaivan/credit_scoring/blob/master/mlp.py.

В работе был применен следующий метод регуляризации: к вычисленной величине коррекции Δw_{ij} весового коэффициента прибавляется текущее значение w_{ij} этого весового коэффициента, умноженное на значение регуляризации, после чего полученное значение становится величиной коррекции Δw_{ij}.

Скорость обучения нейронной сети может задаваться либо константой, либо убывающей функцией от количества эпох обучения. Задание скорости обучения с помощью убывающей функции, например линейной, позволяет подстраивать скорость обучения. В данной работе скорость обучения была задана константой, равной 0,005, так как такая скорость обучения считается эталонной для выбранной архитектуры сети. Она обеспечивает высокую точность настройки весов. Значение регуляризации также было задано константой, равной 0,005.

Количество эпох обучения нейронной сети может задаваться в зависимости от желаемой точности или определяться заранее. В данной работе использовалось фиксированное количество эпох обучения. Тесты показали, что для обучения достаточно одной эпохи, т.е. сети достаточно один раз обучиться на каждом примере, а последующее обучение приводит лишь к переобучению. Это обусловлено большим количеством записей в обучающем множестве (более 200 тыс.).

Для достижения наибольшей точности было решено передавать сети данные не пакетами (батчами), а отдельными записями. Это привело к значительному увеличению времени обучения, но позволило добиться более высокой точности.

Для рассматриваемой задачи было решено использовать сеть с двумя скрытыми слоями. Количество нейронов на входном слое всегда равно количеству входных признаков, а количество нейронов на выходном слое – количеству возможных значений целевого признака (классов). В скрытых слоях количество нейронов необходимо подбирать самостоятельно. В ра-

боте был использован метод перебора. Количество нейронов в первом и втором скрытых слоях варьировалось в диапазоне от 85 до 115 с шагом 3. Для каждой конфигурации была посчитана *cross-validation точность* сети (*CV-точность*).

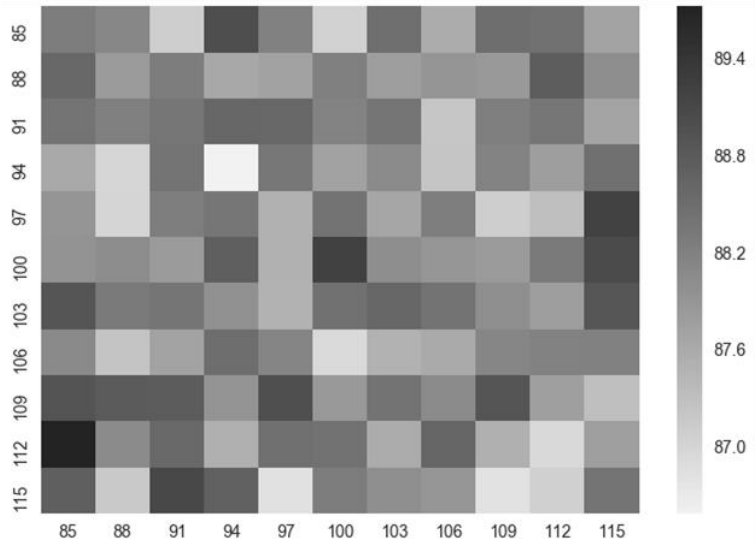

Рис. 2. Тепловая карта cross-validation точности различных конфигураций нейронной сети (OSR)

На рис. 2 представлена тепловая карта CV-точности для каждой конфигурации. Согласно данным карты была выбрана конфигурация, обеспечивающая наивысшую точность (112 нейронов в первом скрытом слое и 85 нейронов во втором).

Оценка точности модели

Одним из важнейших этапов построения классификационных моделей является оценка ошибки, именно на данном этапе подтверждается значимость модели и ее эффективность. В качестве показателя точности использовался *общий показатель успеха* (overall success rate, OSR) – отношение правильно классифицированных наблюдений к их общему числу.

Процесс оценки качества модели сводится к тестированию предикативной способности на тестовом множестве, т.е. на тех данных, которые модель еще не видела. В связи с этим обычно исходные данные разделяют на три множества: *обучающее*, *валидационное* и *тестовое*.

После того как модель прошла тестирование и показала хороший результат, тестовое множество может быть использовано для повышения качества модели. При этом считается, что дополнительное обучение не может снизить точность модели.

Для оценки точности была использована *матрица классификации*. Матрица классификации содержит две оси: фактический класс и предсказанный. Обычно вдоль строк располагаются значения фактического класса, а вдоль столбцов – значения предсказанного класса. То есть элемент матрицы, находящийся на пересечении *i*-й строки и *j*-го столбца, показывает количество результатов классификаций, которые имели фактически *i*-й класс, а модель поставила данным наблюдениям в соответствие *j*-й класс. Значения, находящиеся на главной диагонали, означают правильные классификации, а сумма значений главной диагонали (след матрицы) отражает количество правильно классифицированных записей.

Была рассмотрена динамика обучения модели (рис. 3). В результате анализа было выявлено, что для обучения нейронной сети достаточно показать 40 тыс. записей, потери на обучающей выборке и на тестовой примерно одинаковы на протяжении всего процесса обучения, отсутствует эффект переобучения.

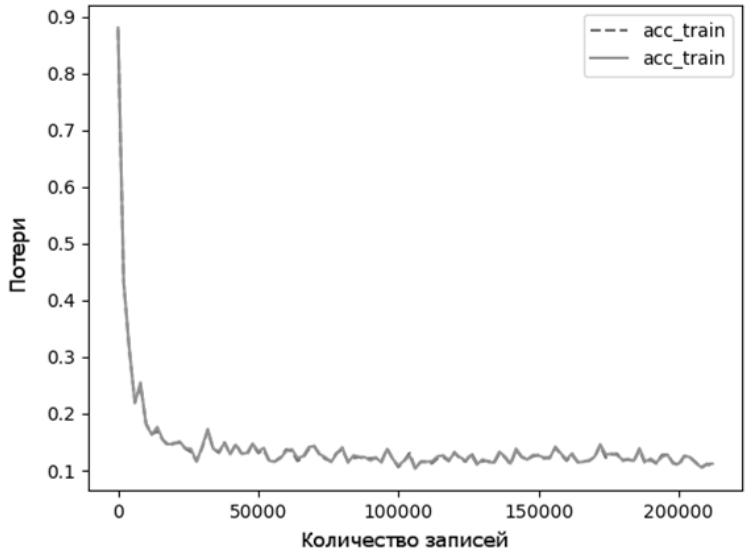

Рис. 3. Потери обучения на обучающем и тестовом множествах

Таблица 2. **Матрица классификации**

		Предсказанный класс						
		A	B	C	D	E	F	G
Фактический класс	A	3028	0	0	576	0	0	0
	B	1	6059	204	318	5	2	0
	C	1	340	3773	1	219	5	0
	D	72	550	0	5661	0	0	0
	E	1	1	173	0	1794	26	0
	F	0	0	0	0	104	502	0
	G	1	0	0	0	6	140	0

На тестовом множестве модель смогла определить 20 817 из 23 563 элементов, т.е. точность составила 88,35% (табл. 2). Большая часть ошибочных классификаций сосредоточена вдоль главной диагонали, т.е. если модель и ошибается, то в основном она относит элемент одного класса к соседнему, что для многоклассовой классификации не столь критично. Низкие результаты для класса G обусловлены относительно малым количеством записей данного класса в исходной выборке. С практической точки зрения наибольший интерес представляют элементы матрицы классификации, расположенные ниже главной диагонали, т.е. эти исходы обладают высокими издержками классификации. Точность построенного классификатора, вычисленная методом 10-блочной кросс-валидации [7. С. 172], равна 88,23%.

Заключение

В работе показано применение нейронной сети в решении задачи кредитного скоринга. Полученные результаты позволяют сделать вывод, что многоклассовая классификация заемщиков по рейтингу кредитоспособности может быть эффективно решена простым многослойным персептроном.

В ходе решения задачи использовалось программное обеспечение с открытым исходным кодом. Подготовка данных, построение модели и ее оценка были осуществлены на языке Python в среде разработки PyCharm. Для работы были использованы базовые библиотеки NumPy [8] и Pandas [9]. Стандартизация распределения признаков, разбиение набора данных на обучающее и тестовое множества, построение матрицы классификации осуществлялись с помощью библиотеки Scikit-learn [10]. Листинг многослойного персептрона был написан самостоятельно.

Нейронные сети могут выявлять сложные, нетривиальные связи между входными и выходными переменными, что позволяет существенно повысить эффективность принимаемых решений, сократить время рассмотрения заявок, снизить влияние человеческого фактора. Автоматизация такой рутинной процедуры, как скоринг, позволяет банкам сократить затраты на соответствующие операции, а освободившиеся трудовые и финансовые ресурсы направить на решение иных задач.

Применение нейронных сетей банками не ограничивается одним скорингом заявителя, но также используется для выявления случаев мошенничества, при работе с должниками и в маркетинге. Это обусловлено высокой универсальностью нейронных сетей.

Литература

1. *Официальная* статистика. Финансы / Росстат. М., 1999–2018. URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/finance (дата обращения: 01.11.2017).

2. *LendingClub:* официальный сайт. San Francisco, 2006–2018. URL: https://www.lendingclub.com (дата обращения: 10.10.2017).

3. *Глинкина Е.В.* Кредитный скоринг как инструмент повышения эффективной оценки кредитоспособности // Банковское дело. 2011. № 16. С. 43–47.

4. *Алешин В.А., Рудаева О.О.* Кредитный скоринг как инструмент повышения качества банковского риск-менеджмента в современных условиях // Terra Economicus. 2012. Т. 10, № 2 (3). С. 27–30.

5. *Хайкин С.* Нейронные сети: полный курс. 2-е изд. / пер. с англ. М. : Вильямс, 2006. 1104 с.

6. *Паклин Н.Б., Орешков В.И.* Бизнес-аналитика: от данных к знаниям. СПб. : Питер, 2013. 704 с.

7. *Рашка С.* Python и машинное обучение / пер. с англ. А.В. Логунова. М. : ДМК Пресс, 2017. 418 с.

8. NumPy // NumPy developer. 2005–2018. URL: http://www.numpy.org (Accessed: 29.04.2018).

9. Pandas // AQR Capital Management. 2008–2018. URL: http://pandas.pydata.org (Accessed: 29.04.2018).

10. Scikit-learn // Scikit-learn developer. 2007–2018. URL: http://scikit-learn.org (Accessed: 29.04.2018).

Bogdanov A.L., Department of Information Technologies and Business Analytics, Institute of Economics and Management, National Research Tomsk State University (Tomsk, Russian Federation). E-mail: bogdanov.al@mail.tsu.ru

Dulya I.S., Department of Information Technologies and Business Analytics, Institute of Economics and Management, National Research Tomsk State University (Tomsk, Russian Federation). E-mail: idulya7@gmail.com

AN APPLICATION OF NEURAL NETWORKS TO SOLUTION OF THE CREDIT SCORING TASK

Keywords: classification, multilayer perceptron, neural network, machine learning, data analysis.

The article describes the application of neural networks to solve the task of credit scoring. Also shows the effectiveness of their application to real tasks in example of building model determines the rating of borrower creditworthiness. The statistic data received from platform of mutual crediting Lending Club was used as an initial data. Application of multilayer perceptron was shown in the task of application scoring. The article contains a detailed description of process of knowledge extraction from data: from data preprocessing to building a model and assessing its accuracy.

References

1. Official statistics. Finance: [Electronic resource] // Rosstat. electronic data. Moskow, 1999-2018. URL: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/finance (circulation date: 01.11.2017)/

2. LendingClub [Electronic resource] // Official site - Electronic. San Francisco, 2006-2018 - URL: https://www.lendingclub.com (circulation date: 10.10.2017)/

3. Glinkina E.V. Kreditnyy skoring kak instrument povysheniya effektivnoy otsenki kreditosposobnosti // Bankovskoye delo. 2011. № 16. P. 43–47.

4. Aleshin V.A. Kreditnyy skoring kak instrument povysheniya kachestva bankovskogo risk-menedzhmenta v sovremennykh usloviyakh / V.A. Aleshin, O.O. Rudayeva // Terra Economicus. 2012. T. 10, № 2 (3). P. 27-30.

5. Khaykin S. Neyronnyye seti: polnyy kurs, 2-ye izdaniye. M .: Williams Publishing House, 2006. 1104 p.

6. Paklin N.B. Biznes-analitika: ot dannykh k znaniyam / N. B. Paklin, V. I. Oreshkov. - SPb.: Piter, 2013. 704 p.

7. Rashka S. Python i mashinnoye obucheniye / S. Rashka. M.: DMK Press, 2017. 418 p.

8. NumPy: [Electronic resource] // NumPy developer, Electronic data, 2005-2018. URL: http://www.numpy.org (circulation date: 29.04.2018).

9. Pandas: [Electronic resource] // AQR Capital Management, Electronic data, 2008-2018 URL: http://pandas.pydata.org (circulation date: 29.04.2018).

10. Scikit-learn: [Electronic resource] // Scikit-learn developer, Electronic data, 2007-2018. URL: http://scikit-learn.org (circulation date: 29.04.2018).

For referencing:
Bogdanov A.L., **Dulya I.S.** Primenenie nejronnyh setej v reshenii zadachi kreditnogo skoringa **[**An application of neural networks to solution of the credit scoring task]. Vestnik Tomskogo gosudarstvennogo universiteta. Ekonomika – Tomsk State University Journal of Economics, 2018, no 44, pp. 173–183.