УДК 519.24:004.89

О. М. Гусарова, П. И. Комаров, Д. Э. Денисов
ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», Смоленск, e-mail: om.gusarova@mail.ru

НЕЙРОННЫЕ СЕТИ В КРЕДИТНОМ СКОРИНГЕ

Ключевые слова: кредитный скоринг, искусственные нейронные сети, оценка надежности заемщика.

В современных условиях кредитные организации сталкиваются с рядом проблем, одной из которых является риск невозврата кредита. Принятие решения о выдаче заемщику денежных средств в ряде случаев базируется на информации, предоставляемой самим заемщиком, в силу чего может иметь место получение недостоверной информации. Кредитная организация вправе разрабатывать методику оценки платежеспособности заемщика и перечень документов, предоставляемых клиентом. Информация, внесенная в скоринговые карты (досье клиента), оценивается по ряду параметров. Такая система скоринга обладает рядом недостатков: оценка платежеспособности заемщика носит субъективный характер и в значительной степени определяются опытом работы сотрудников банка; незначительное количество параметров, по которым оценивается риск невозврата кредита, также не в полной мере характеризует действительное положение вещей. В рамках настоящего исследования осуществлено построение модели кредитного скоринга на основе искусственных нейронных сетей, нивелирующих фактор субъективности в оценке надежности заемщика. Выполнена методическая проработка подготовительного этапа для построению нейросетевых моделей, в частности, осуществлен анализ входных факторов-аргументов, определяющих значение выходного параметра «благонадежность заемщика». Выполнены расчеты количества синаптических весов нейронной сети и числа нейронов в скрытом слое. Проектирование нейросетевой модели кредитного скоринга осуществлено с использованием программного продукта R Studio. Разработаны нейроетевые модели оценки надежности заемщика с одним и двумя скрытыми слоями. Осуществлена оценка и сравнительный анализ погрешности построенных нейросетевых моделей. Используя нейросетевые модели кредитного скоринга, выявлены ключевые факторы, такие как доход заемщика и заработная плата, оказывающие наибольшее влияние на значение выходного параметра «благонадежность заемщика». Практическая значимость исследования заключается в возможности использования разработанной нейросетевой модели кредитного скоринга для определения надежных заемщиков и минимизации риска невозврата кредита.

Введение

В современных условиях цифровизации всех сфер экономики и управления внедрение систем искусственного интеллекта в одном из самых проблемных направлений банковского сектора – кредитном скоринге, получило дальнейшее развитие. Кредитные организации начали разработку собственных программных продуктов, основанных на собственных методиках. На рынке появились специализированные программные продукты, использующие различные математические модели. Информация, необходимая для оценки кредитоспособности заемщика, бралась не только из документов, предоставленных заемщиком, но также из баз данных кредитных организаций. По данным некоторых исследователей, внедрение таких систем позволило сократить на 50% уровень безнадежного долга [1]. В 80-х годах прошлого века появились первые разработки на основе искусственного интеллекта, в частности, компания HNC разработала нейросетевую модель кредитного скоринга, обладавшую преимуществами по сравнению с моделями статистического анализа, главным из которых была способность к обучению. Ряд научных публикаций посвящены нейросетевому моделированию практических направлений различных сфер деятельности [2, 3]. Это определило начало перехода многих кредитных организаций на системы оценки кредитного риска на основе искусственного интеллекта.

Цель исследования

Целью исследования является разработка модели кредитного скоринга на основе искусственного интеллекта при помощи нейронных сетей, а также анализ возможностей использования нейросетевой модели в практической деятельности организаций банковского сектора.

Материалы и методы исследования

При осуществлении исследования в качестве статистической выборки использовалась информация по 1000 заемщикам [4]. Статистическая выборка рассматривалась как два подмножества: обучающая выборка – информация о 800 заемщиках; тестирующая выборка – информация о 200 заемщиках. Поскольку информация, использующаяся для построения нейросетевой модели, имеет разнообразный характер, а нейронная сеть оперирует только с числовыми данными, вся нечисловая информация была закодирована в соответствии принятыми правилами. Например, пол мужской кодируют числом 1, пол женский – 0 (или наоборот). В табл. 1 приведена структура статистической выборки с учетом кодирования нечисловой информации.

Как видно из таблицы, значения 9 входных переменных должны определять значение одной выходной переменной «Благонадежный заемщик». Отметим, что объем обучающей выборки удовлетворяет условию репрезентативности, согласно которому этот объем (информация о 800 заемщиках) должен быть больше, чем $7 \cdot N + 15$, где $N = 9$ – число входных переменных [5].

При осуществлении исследования использовались методы системного анализа, методы моделирования на основе искусственных нейронных сетей, методы вероятностно-статистического моделирования.

Результаты исследования и их обсуждение

Проблема кредитного скоринга может быть сформулирована следующим образом: пусть известны ответы заемщика на вопросы анкеты, обозначаемые как $x \in A$, тогда необходимо определить группу, к которой относится заемщик: $x \in A_B$ – «плохие» или $x \in A_G$ – «хорошие». Естественно, «плохой» и «хороший» заемщики могут иметь одинаковые ответы на одни и те же вопросы анкеты, поэтому моделирование носит вероятностный характер.

Разработка модели скоринга требует полной и достоверной информации о заемщиках кредитной организации. Объем данных может меняться в зависимости от конкретной модели, но в любом случае должен удовлетворять условиям случайности и статистической значимости. Для построения модели могут использоваться как внутренняя информация банка, так и внешние данные, предоставляемые, например, Национальным бюро кредитных историй. Модель должна применяться только в отношении тех кредитных продуктов (сектора рынка или экономической ситуации), данные о которых легли в основу проектирования нейросетевой модели. Так сведения по ипотеке не целесообразно использовать при разработке модели скоринга по автокредитам. Важен также период получения информации: например, информацию для построения модели скоринга заявок потребительских кредитов рекомендуют брать за последние 2–5 лет, для проектирования моделей поведенческого скоринга рекомендуется использовать информационный интервал 6–12 месяцев [6]. Как правило, при разработке модели из исходного массива данных исключают информацию о «нетипичных» клиентах (мошенники, сотрудники банка, умершие клиенты, VIP-клиенты, клиенты с аномально большими суммами, нестандартными условиями погашения и целями кредита).

Определение зависимой переменной предусматривает деление всех клиентов на две категории: «хорошие» и «плохие». В категорию «хорошие» попадают клиенты, добросовестно и в полной мере исполняющие свои обязательства

Таблица 1

Структура статистической выборки

Возраст (полных лет)	Пол (мужской – 1, женский – 0)	Состоит в браке (1 – да, 0 -нет)	Иждивенцы (количество)	Доход (рублей)	Опыт работы (полных лет)	Срок проживания (полных лет)	Рыночная стоимость недвижимости (тыс. руб.)	Зарплата (рублей)	Благонадежный заемщик (1 – да, 0 -нет)

перед банком. К категории «плохие» относят мошенников, банкротов, «безнадежных» заемщиков, а также клиентов со следующими параметрами:

- количество дней просрочки платежа превышает установленное значение;
- размер пророченной задолженности превышает величину, установленную банком;
- количество просрочек более установленного числа дней превышает величину, установленную банком.

Возможно введение двух дополнительных категорий клиентов: «отклоненные» заемщики, которым отказано в выдаче кредита; «неопределенные» заемщики – это клиенты с недостаточной кредитной историей, имеющие незначительные просрочки платежа и т. п. По мнению ряда экспертов, учет отклоненных заявок, требует значительно больше ресурсов и не всегда приводит к качественному улучшению модели. Таким образом, при проектировании модели кредитного скоринга целесообразно рассматривать зависимую переменную с двумя категориями: «плохой» и «хороший».

При проектировании модели кредитного скоринга могут быть использованы различные методы: статистики (линейная регрессия, дискриминантный анализ), эконометрического моделирования (корреляционно-регрессионный, дисперсионный, факторный анализ); методы оптимизации, методы экспертных оценок, методы искусственного нейросетевого моделирования. Ряд авторов научных исследований оценивают точность методов моделирования кредитного скоринга и отдают предпочтение тому или иному методу моделирования (табл. 2) [7, 8].

При осуществлении данного исследования проектирование модели кредитного скоринга было осуществлено с использованием искусственных нейронных сетей в среде R Studio.

При построении нейросетевой модели кредитного скоринга для определения значения выходной функции «Благонадежный заемщик» в качестве аргументов используют следующие входные данные:

- социально-экономические (пол, возраст, семейное положение, стаж работы общий и на последнем месте работы, состав семьи, доход личных и семьи в целом, наличие депозитов и их сумма);
- информация о кредите (сумма, назначение, обеспечение, срок погашения и т. д.);
- кредитная история (рейтинг, информация о взятых и погашенных кредитах, об имеющихся кредитах, в том числе просроченных, наличие других банковских продуктов).

Как видно из приведенного выше перечня аргументы могут быть как количественными (размер кредита, срок погашения и т. д.), так и качественными (пол, назначение кредита). Для использования качественных переменных для построения модели кредитного скоринга осуществлено их нормирование.

Статистическая выборка разделена на две части: обучающая выборка (данные о 800 заемщиках) использована для расчета числовых параметров модели; тестовая выборка (информация о 200 заемщиках) – для проверки адекватности модели, т. е. способности построенной нейросетевой модели отличать «хороших» заемщиков от «плохих». Для этого в нейросетевую модель подставляют данные тестовой выборки с заранее

Таблица 2

Оценка точности методов построения модели кредитного скоринга

Автор	Линейная регрессия (%)	Логистическая регрессия (%)	Рекурсионно-партиционный анализ (%)	Линейное программирование (%)	Нейронные сети (%)	Генетический алгоритм (%)
Хенли (1995)	43,40	43,30	43,80			
Бойл (1992)	77,50		75,00	74,70		
Шринивисан (1987)	87,50	89,30	93,20	86,10		
Йобас (1997)	68,40		62,30		62,00	64,50
Десаи (1997)	66,50	67,30	67,30		64,00	

известными критериями «хороший», «плохой» заемщик; если тестовая выборка соответствует «плохому» заемщику, и модель дает такой же результат, следовательно, на этом тестовом наборе модель адекватна и данную нейросетевую модель можно использовать для определения благонадежности заемщиков.

Естественно, что на всей тестовой выборке нейростевая модель может и не дать полного совпадения результатов и заранее известных критериев заемщиков, однако, если различия будут иметь место в пределах заданной погрешности, можно говорить об адекватности нейросетевой модели и ее возможности для практического применения оценки надежности заемщиков.

Расчет необходимого количества синаптических весов нейронной сети осуществлен с использованием формулы, вытекающей из теоремы Колмогорова – Арнольда-Хехт-Нильсена:

$$\frac{N_y \cdot Q}{1+\log_2 Q} \le N_w \le N_y \times \times\left(\frac{Q}{N_x}+1\right)\cdot\left(N_x+N_y+1\right)+N_y, \quad (1)$$

где N_x – количество входных факторов-аргументов, определяющих число нейронов входного слоя (V1–V9); N_y – количество нейронов выходного слоя, определяемое числом выходных переменных (V10); Q – размерность обучающей выборки (информация о 800 заемщиках); N_w – необходимое число синаптических связей.

Число нейронов скрытого слоя N может быть определено по формуле:

$$N = \frac{N_w}{N_x + N_y}, \quad (2)$$

Подставив в формулу (1) соответствующие значения переменных, получим, что число синаптических связей нейронной сети принадлежит интервалу [75; 989]. Количество нейронов скрытого слоя модели ИНС, рассчитанное по формуле (2), должно находиться в интервале [7; 99].

Авторы исследования ранее отмечали, что «на практике число нейронов в скрытых слоях выбирают в пределах от $N_x/2$ до $3 \cdot N_x$, и, как правило, их число определяется ошибкой, получаемой на этапе обучения сети. В нашем случае число нейронов в скрытом слое начнем изменять от 2 в сторону увеличения» [9].

При проектировании нейросетевой модели кредитного скоринга в R Studio использовались следующие возможности программного продукта: считывание подготовленных исходных данных; формирование структуры нейронной сети при обучении и тестировании; нормирование входных и выходных данных для этапа обучения; нормирование входных данных при тестировании и представление выходных данных в реальном диапазоне.

В результате проектирования нейросетевой модели получены следующие результаты (рис. 1).

Соответствие между переменными модели и переменными предметной области представлено в табл. 3.

Таблица 3

Соответствие между переменными модели и переменными предметной области

Переменная модели	Переменная предметной области
V1	Возраст (полных лет)
V2	Пол
V3	Состоит в браке
V4	Иждивенцы (количество)
V5	Доход (рублей)
V6	Опыт работы (полных лет)
V7	Срок проживания (полных лет)
V8	Рыночная стоимость недвижимости (тыс. рублей)
V9	Зарплата (рублей)
V10	Благонадежный заемщик

С целью определения параметров ИНС с наименьшей погрешностью будем менять число нейронов в скрытом слое и оценивать погрешность, получаемую на тестирующей выборке.

В качестве меры погрешности оценки выберем среднее квадратичное отклонение:

$$\delta = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(y_i^t - y_i^p\right)^2}, \quad (3)$$

где δ – погрешность оценки; n – объем тестирующей выборки; y_i^t – значение

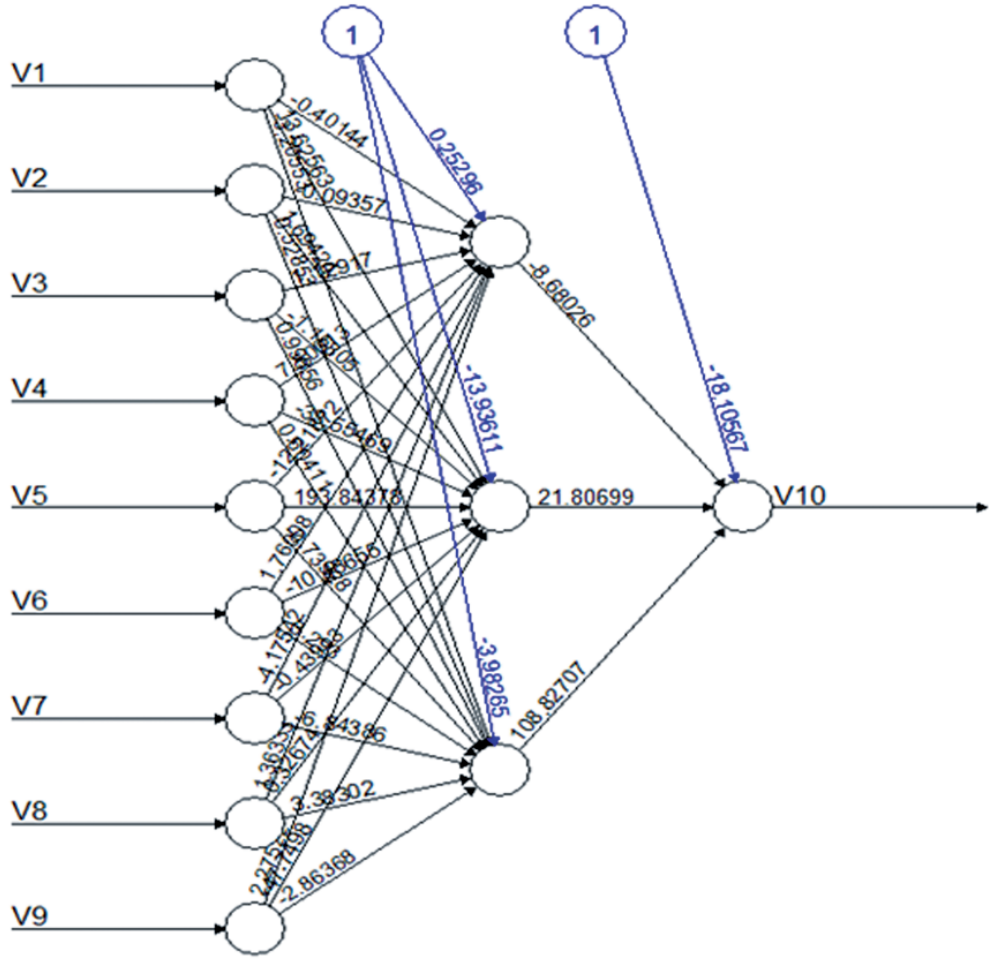

Рис. 1. Нейросетевая модель кредитного скоринга с тремя нейронами в скрытом слое

выходной переменной из тестирующей выборки (заемщик «хороший» или «плохой»); y_i^p – значение выходной переменной, полученное ИНС на тестирующем наборе (прогнозная оценка заемщика).

График зависимости ошибки модели от числа нейронов ИНС в скрытом слое представлен на рис. 2.

Визуальный анализ представленного графика позволяет утверждать, что наименьшее значение ошибки модели достигается при количестве нейронов в скрытом слое равное 3.

В ходе исследования для повышения точности нейросетевой модели была разработана искусственная нейронная сеть (ИНС) с двумя скрытыми слоями (рис. 3).

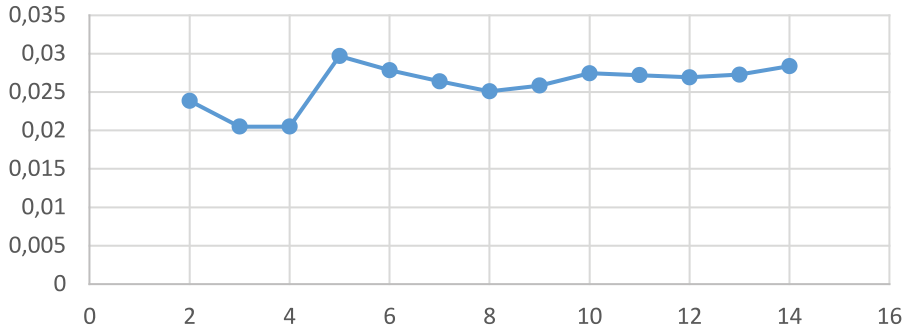

Рис. 2. График зависимости ошибки ИНС от числа нейронов в скрытом слое

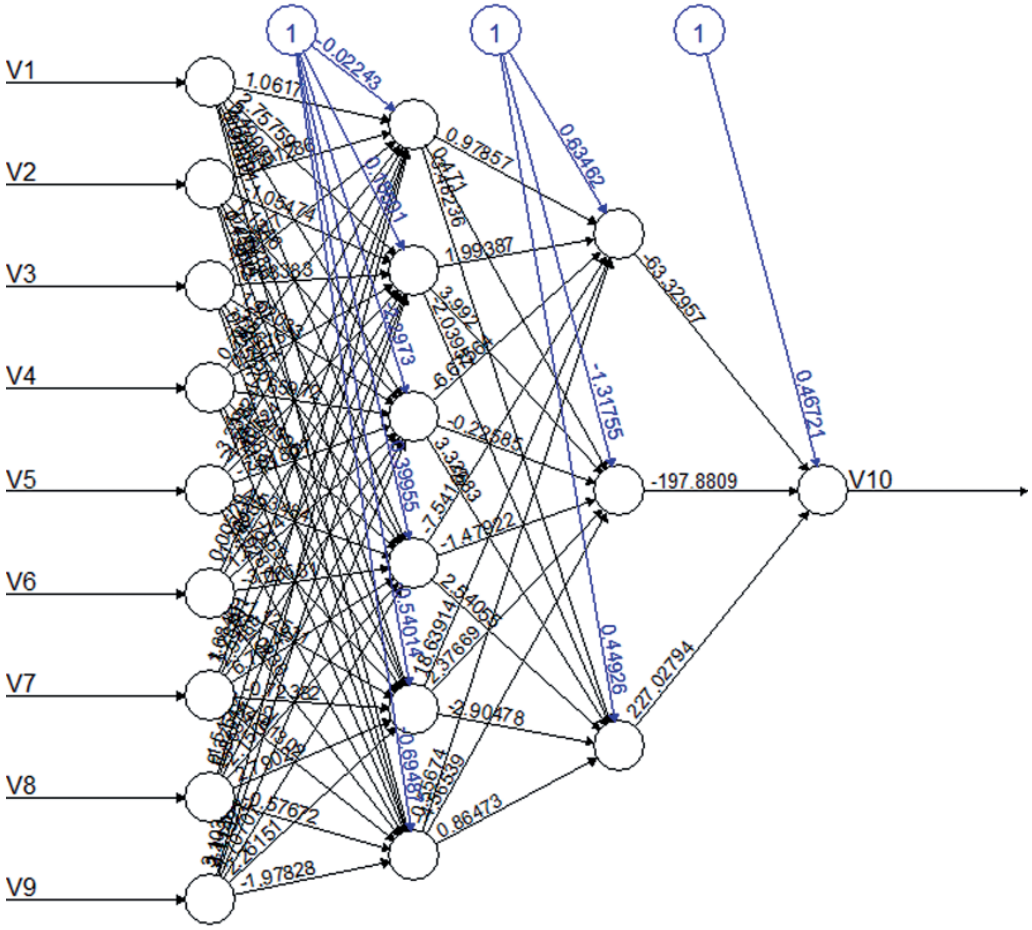

Рис. 3. Искуственная нейронная сеть с двумя скрытыми слоями

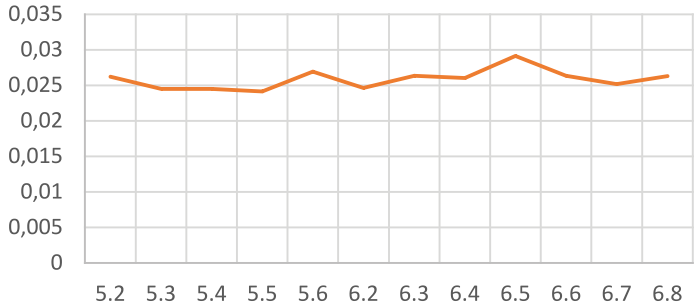

Рис. 4. График зависимости ошибки ИНС от числа нейронов в скрытых слоях

График зависимости ошибки ИНС от числа нейронов в скрытых слоях представлен на рис. 4 (первая цифра – число нейронов в первом скрытом слое, вторая – во втором).

Одним из направлений исследования является оценка влияния входных факторов на значение выходной переменной. Для решения данной проблемы могут быть использованы различные методы, в частности, методы эконометрического моделирования, рассмотренные в [10]. В рамках настоящего исследования, используя модель кредитного скоринга, полученную на основе ИНС, осуществим ее анализ, в частности, определим, от каких факторов в больней степени зависит значение выходной переменной «благонадежность заемщика». Для этого будем поочередно исключать входные факторы и оценивать точность модели. Графическая интерпретация полученных результатов исследования представлена на рис. 5.

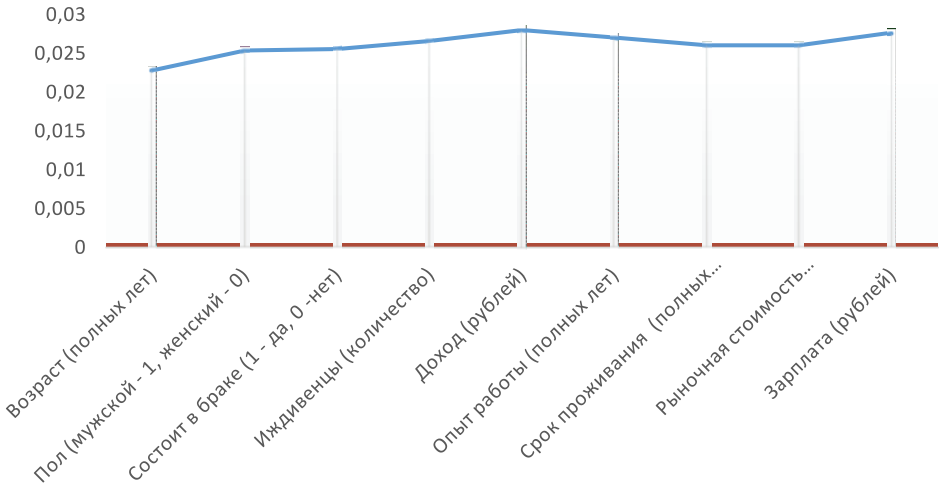

Рис. 5. График оценки точности модели ИНС

Анализ представленного графика позволил выявить наиболее значимые факторы, определяющие благонадежность заемщика, которыми являются доход заемщика и зарплата.

Выводы по результатам исследования:

● разработаны модели кредитного скоринга на основе ИНС с одним и двумя скрытыми слоями, позволяющие прогнозировать «надежность» заемщика;

● нейросетевые модели кредитного скоринга позволяют нивелировать фактор субъективности в оценке надежности заемщика;

● по результатам сравнения точности (величины ошибки) ИНС с одним скрытым слоем и ИНС с двумя скрытыми слоями, можно сделать вывод о том, что погрешность для ИНС второго вида в меньшей степени зависит от числа нейронов в скрытых слоях, хотя в целом имеет большую погрешность, по сравнению с ИНС с одним скрытым слоем (минимальное значение погрешности ИНС с одним скрытом слоем равно 0,02050, с двумя скрытыми слоями – 0,02412);

● на основе построенных ИНС выявлены ключевые факторы, в наибольшей степени оказывающие влияние на «надежность» заемщика – доход заемщика и его зарплата;

● практическая значимость осуществленного исследования заключается в возможности использования нейросетевой модели кредитного скоринга для оценки надежности заемщика и снижения риска потерь и/или сокращение числа дефолтов по выданным кредитам.

Библиографический список

1. Степанов П.П. Искусственные нейронные сети // Молодой ученый. – 2017. – №4 (138). – С. 185–187.

2. Сорокин А.С. Построение скоринговых карт с использованием модели логистической регрессии // Науковедение. – 2014. – № 2. – URL: http://www.naukovedenie.ru (дата обращения 28.11.2018).

3. Мисник А.Е., Борисов В.В. Композиционное нейросетевое моделирование сложных технических систем // Нейрокомпьютеры: разработка, применение. – 2016. – № 7. – С. 39–46.

4. Ясницкий Л.Н. Интеллектуальные системы. – М.: Лаборатория знание, 2016. – 221 с.

5. Haykin S. Neural networks: A comprehensive foundation (2nd ed.). – New Jersey: Prentice Hall International, Inc. 1999. – 1103 p.

6. Марк Андрессен. Why Bitcoin Matters // The New York Times 21.01.2014 [Электронный ресурс]. – URL: https:// jscc-is.ru/projects/oblachnyie-shkolnyie-biblioteki (дата обращения 29.11.2018).

7. Паклин Н. Логистическая регрессия и ROC-анализ – математический аппарат [Электронный ресурс]. – URL: https://basegroup.ru/community/articles/logistic (дата обращения 05.12.2018).

8. Паклин Н. Применение логистической регрессии в медицине и скоринге [Электронный ресурс]. – URL: https://basegroup.ru/community/articles/logis-medic-scoring (дата обращения 05.12.2018).

9. Комаров П.И., Гусарова О.М., Таранец С.А. Нейросетевые модели оценки стоимости бренда компании // Современные наукоемкие технологии. – 2018. – № 12. – С. 128–132.

10. Гусарова О.М. Информационно-аналитические технологии прогнозирования деятельности организаций // Международный журнал прикладных и фундаментальных исследований. – 2015. – № 12(3). – С. 492–495.