УДК [004.78:33](075.8)

ИССЛЕДОВАНИЕ ДОСТОВЕРНОСТИ ОПТИМИЗИРОВАННОЙ МОДЕЛИ СКОРИНГА ПУТЕМ ПРОГНОЗИРОВАНИЯ КРЕДИТНЫХ ИСТОРИЙ ЗАЕМЩИКОВ, ДАННЫЕ КОТОРЫХ НЕ ИСПОЛЬЗОВАЛИСЬ ПРИ СИНТЕЗЕ МОДЕЛИ

Лебедев Е.А., – соискатель
Кубанский государственный аграрный университет

В статье рассматривается актуальная проблема прогнозирования рисков кредитования физических лиц, и предлагаются пути решения поставленной задачи. Проводится прогнозирование кредитных историй на двух тестирующих групп заемщиков, данные которых не использовались при синтезе модели: которым были выданы кредиты и которым в этом было отказано. Результаты прогнозирования анализируются с целью установить являются ли тестирующие выборки частью той генеральной совокупности, по отношению к которой репрезентативна обучающая выборка, использованная при синтезе модели.

Ключевые слова: ДОСТОВЕРНОСТЬ ОПТИМИЗИРОВАННАЯ СЕМАНТИЧЕСКАЯ ИНФОРМАЦИОННАЯ МОДЕЛЬ СКОРИНГ ПРОГНОЗИРОВАНИЕ КРЕДИТНАЯ ИСТОРИЯ ЗАЕМЩИК СИНТЕЗ ВНУТРЕННЯЯ И ВНЕШНЯЯ ВАЛИДНОСТЬ АДЕКВАТНОСТЬ

Данная статья является продолжением статьи автора (7). В работе автор описал применение метода системно-когнитивного анализа (СК-анализ) для определения кредитоспособности потенциальных заемщиков. С помощью специального программного инструмента СК-анализа – универсальной когнитивной аналитической системы “Эйдос” было осуществлено решение следующих задач:

1. Формализация предметной области.
2. Формирование обучающей выборки.
3. Синтез модели
4. Оптимизация.
5. Верификация модели.

Для синтеза модели были использованы данные из 400 кредитных досье заемщиков получивших кредит в Краснодарском отделении

Сбербанка России №8619 в период с 2002 по 2006 гг. и имеющих кредитную историю.

1. Формализация предметной области.

Исходя из определения положительной кредитной истории изложенного в Правилах кредитования физических лиц Сбербанком России и его филиалами от 30.05.2003 №229-3р., было принято решение о формировании двух классов заемщиков с “положительной” и “отрицательной” кредитной историей.

Для решения задачи формализации предметной области решено использовано 17 описательных шкалах и 412 градациях. Описательные шкалы представлены в таблице 1. Так как количество градаций слишком велико, в рамках данной статьи градации расшифровываться не будут.

ТАБЛИЦА 1 – ОПИСАТЕЛЬНЫЕ ШКАЛЫ И ГРАДАЦИИ ИСПОЛЬЗУЕМЫЕ ДЛЯ ФОРМАЛИЗАЦИИ ПРЕДМЕТНОЙ ОБЛАСТИ

№ п./п.	Наименование описательной шкалы	Кол-во градаций описательной шкалы
1.	Пол	2
2.	Возраст	58
3.	Место рождения	5
4.	Семейное положение	8
5.	Наличие иждивенцев	4
6.	Округ проживания	4
7.	Продолжительность проживания на последнем месте	21
8.	Продолжительность проживания на предпоследнем месте	9
9.	Образование	5
10.	Место работы	188
11.	Сфера деятельности работодателя	39
12.	Организационно-правовая форма работодателя	5
13.	Должность	14
14.	Стаж на последнем месте работы	21
15.	Доходы	15
16.	Коэффициент долговой нагрузки	11
17.	Наличие собственности	3

2. Формирование обучающей выборки

Разработав описательные и классификационные шкалы, переходим к формированию обучающей выборки, которая включает в себя информацию о факторах влияющих на состояние объекта управления и о состоянии объекта. Информация в обучающей выборке была зашифрована согласно справочникам классов и признаков и приняла вид, показанный в таблице 2.

ТАБЛИЦА 2 – ОБУЧАЮЩАЯ ВЫБОРКА (ФРАГМЕНТ)

№	Класс. Шкалы	Описательные шкалы																
	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	2	1	7	64	66	77	78	87	0	115	268	311	344	355	364	390	403	412
2	2	1	14	63	68	74	78	92	0	116	270	311	344	350	366	391	406	410
3	2	2	14	64	73	74	81	82	0	116	119	305	347	357	367	390	405	412
4	2	1	11	64	68	77	79	91	0	116	211	315	344	354	364	392	407	410
5	2	1	40	61	72	74	81	102	0	116	269	311	344	355	364	385	405	410
6	1	2	42	61	73	77	79	96	0	114	0	0	0	362	370	384	402	410
7	1	1	10	64	66	77	81	89	0	114	263	311	344	355	364	388	402	410
8	2	1	11	61	68	74	81	92	0	114	145	339	347	355	365	389	405	412
9	2	1	11	64	68	75	81	91	0	114	188	311	345	354	364	390	407	410
10	1	1	8	64	66	77	80	88	0	115	299	342	344	351	364	396	403	412

3. Синтез модели

С помощь системы “Эйдос” проведен синтез модели, который включает в себя расчет матрицы абсолютных частот, поиск и исключение из дальнейшего анализа артефактов, расчет матрицы информативностей, расчет матрицы условных процентных распределений.

4. Оптимизация

Полученная модель была оптимизирована с помощью удаления признаков, по которым имеется недостаточно данных. За пороговое

значение встреч признаков в модели приняли 5%. После оптимизации количество градаций описательных шкал уменьшилось с 412 до 197.

5. Верификация модели

Измерение внутренней валидности.

Для измерения внутренней валидности полученную обучающую выборку скопировали в распознаваемую, после чего провели пакетное распознавание. Измерение внутренней валидности показало, что из 400 анкет выборки, верно идентифицировалось 84,3% анкет, верно не идентифицировались 65,6% анкет, ошибочно не идентифицировались 15,7%, ошибочно идентифицировались 34,4%. Анализируя полученные данные можно предположить, что не все заемщики представленные в выборке сходны по своим признакам в разрезе классов. Так, не смотря на принадлежность заемщиков к одному из существующих классов, 15,7% анкет не были идентифицированы.

Для решения задачи 100% идентификации анкет этап оптимизации модели был повторен. Суть оптимизации состояла в сохранении существующих классов состоящих из верно идентифицирующихся типичных анкет заемщиков и добавлении новых классов состоящих из не идентифицирующихся нетипичных анкет из старых классов модели. Данная процедура была проделана до полной идентификации распознаваемой выборки. После каждого разделения классов для измерения внутренней валидности создается новая итерация модели. Процесс разделения классов показан на рисунке 1. Для решения поставленной задачи процесс оптимизации (разделения классов) был повторен 14 раз, результатом чего стало увеличение количества классов с 2-х до 37-и.

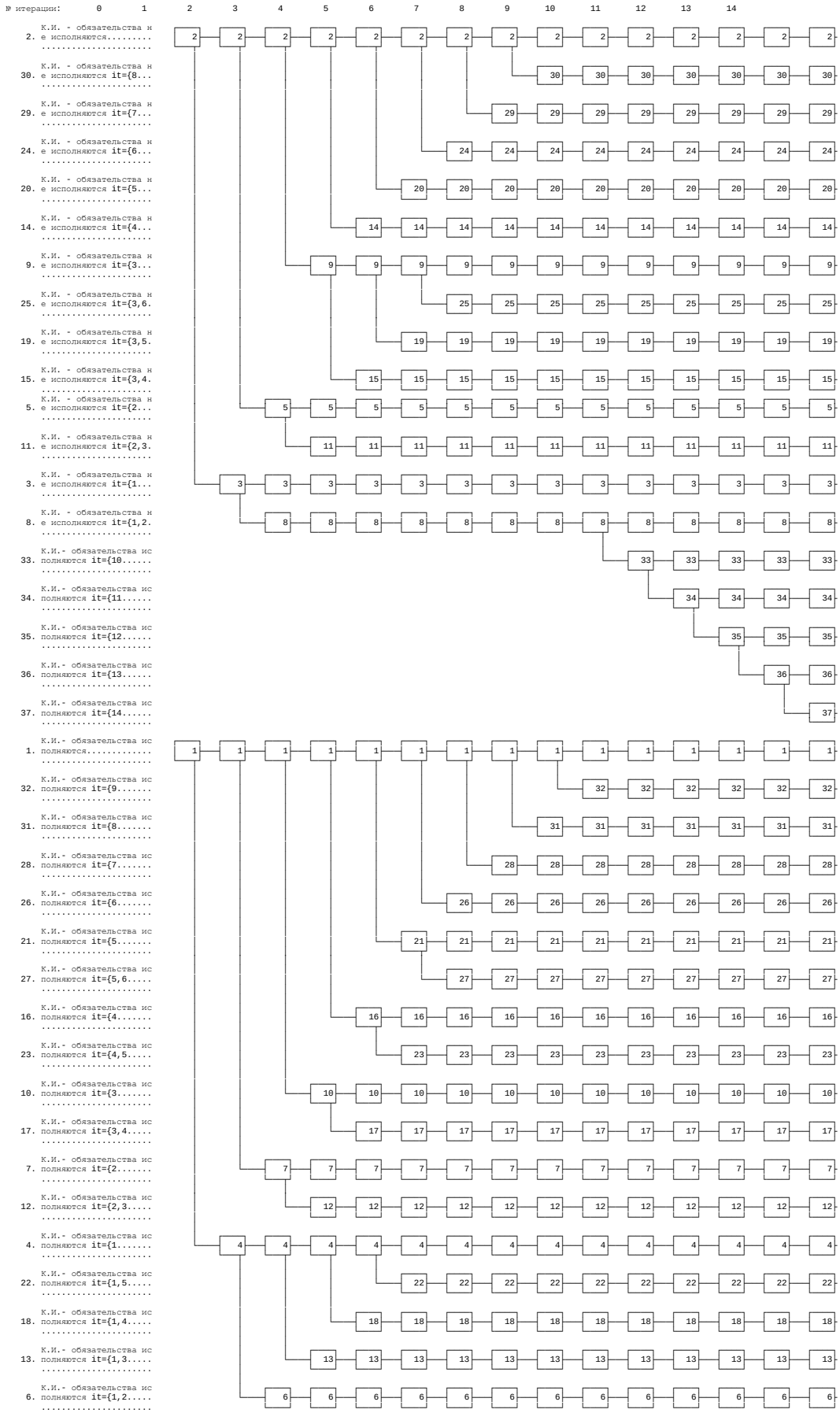

Рисунок 1. Дерево разделения классов.

Результаты оптимизации показаны на рисунке 2. Полученный результат является приемлемым для решения задачи прогнозирования будущих состояний объекта управления, т.к. позволяет производить верную идентификацию заемщиков входящих в обучающую выборку со 100% вероятностью. Также удовлетворительным можно считать процент ошибочной идентификации, который составляет 17,3%.

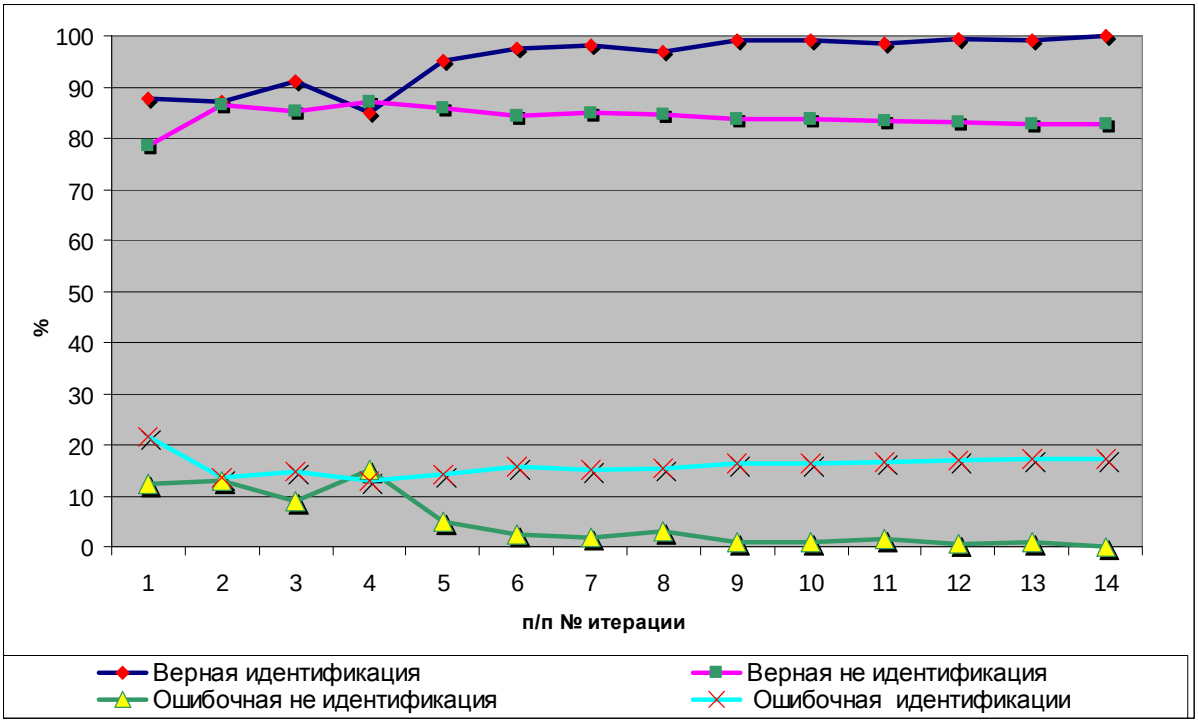

Рисунок 2. Изменение адекватности информационной модели в зависимости от итерации

Измерение внешней валидности.

Теперь можно применить полученную модель для прогнозирования будущей кредитной истории заемщиков не входящих в обучающую выборку на основе, которой производился синтез модели. Для этого сформируем распознаваемую выборку, состоящую из данных 50 заемщиков. Для того, чтобы определить качество будущего прогноза распознаваемая выборка будет состоять из заемщиков получивших кредит в Сбербанке России в 2003 г. и имеющих сложившуюся кредитную

историю. Из 50 заемщиков попавших в выборку положительную кредитную историю имеют 37 человек, у 13 – отрицательная кредитная история. В процентном соотношении только 74% заемщиков получивших кредит и вошедших в распознаваемую выборку имеют положительную кредитную историю.

Для того чтобы начать работу с выборкой зашифруем исходные данные согласно справочникам классов и признаков, после чего проведем пакетное распознавание. Результаты распознавания выводятся в виде карточек. Одну из таких карточек можно видеть на рисунке 3.

Номер анкеты: 32 Наим.физ.источника: Качество результата распознавания: 3.314%

Код	Наименование класса распознавания	% Сх	Гистограмма сходств/различий
37	К.И.- обязательства исполняются it={14}	14	
33	К.И.- обязательства исполняются it={10}	11	
22	К.И.- обязательства исполняются it={1,5}	7	
1	К.И.- обязательства исполняются	√ 6	
35	К.И.- обязательства исполняются it={12}	6	
23	К.И.- обязательства исполняются it={4,5}	6	
11	К.И. - обязательства не исполняются it={2,3}	5	
32	К.И.- обязательства исполняются it={9}	5	
13	К.И.- обязательства исполняются it={1,3}	3	
25	К.И. - обязательства не исполняются it={3,6}	3	
31	К.И.- обязательства исполняются it={8}	2	
15	К.И. - обязательства не исполняются it={3,4}	2	
30	К.И. - обязательства не исполняются it={8}	1	
27	К.И.- обязательства исполняются it={5,6}	1	
26	К.И.- обязательства исполняются it={6}	-0	
17	К.И.- обязательства исполняются it={3,4}	-0	
12	К.И.- обязательства исполняются it={2,3}	-0	
29	К.И. - обязательства не исполняются it={7}	-1	
34	К.И.- обязательства исполняются it={11}	-1	
36	К.И.- обязательства исполняются it={13}	-1	
19	К.И. - обязательства не исполняются it={3,5}	-1	
3	К.И. - обязательства не исполняются it={1}	-2	
28	К.И.- обязательства исполняются it={7}	-4	
18	К.И.- обязательства исполняются it={1,4}	-4	
9	К.И. - обязательства не исполняются it={3}	-5	
8	К.И. - обязательства не исполняются it={1,2}	-6	
24	К.И. - обязательства не исполняются it={6}	-7	
5	К.И. - обязательства не исполняются it={2}	-8	
10	К.И.- обязательства исполняются it={3}	-8	
20	К.И. - обязательства не исполняются it={5}	-8	
21	К.И.- обязательства исполняются it={5}	-10	
14	К.И. - обязательства не исполняются it={4}	-14	
16	К.И.- обязательства исполняются it={4}	-15	
6	К.И.- обязательства исполняются it={1,2}	-17	
7	К.И.- обязательства исполняются it={2}	-18	
2	К.И. - обязательства не исполняются	-21	
4	К.И.- обязательства исполняются it={1}	-30	

Рисунок 3. Результат идентификации информационного источника с классами распознавания

Карточка состоит 3-х столбцов. В первом столбце перечислены все 37 классов имеющиеся в нашей модели. Во втором столбце, напротив каждого класса, указан процент сходства/различия заемщика с данным классом, и, наконец, в третьем столбце можно видеть графическое

отображение сходства/различия заемщика с представленными классами. Специальным значком (галочкой) отмечен класс, к которому нами заемщик был отнесен.

Как вы помните, в ходе оптимизации модели мы разделили имеющиеся 2 класса на 37, отличающихся друг от друга по признакам, характеризующим заемщиков представленных в данных классах, но по своему смыслу являющихся подклассами класса «положительная» и «отрицательная» кредитная история. Так как подкласс заемщиков входящих в распознаваемую выборку нам не известен, респонденты были отнесены к первоначальным классам. Анализируя представленную карточку можно интерпретировать полученный результат, как удачный прогноз, так как наибольший процент сходства с данным заемщиком имеет 37 класс «Обязательства исполняются (14 итерация)» который в свою очередь по своему смыслу является подклассом класса 1 «обязательства исполняются».

Полученный результат прогнозирования кредитной истории у респондентов входящих в распознаваемую выборку выглядит следующим образом: из 50 заемщиков верный прогноз получен по 38, что составляет 76% идентификации выборки; из 37 заемщиков имеющих положительную кредитную историю, верно идентифицировано 32, что составляет 86,5% идентификации; из 13 заемщиков имеющих отрицательную кредитную историю верно идентифицировано 6, что составляет 46,2% идентификации.

Попытаемся сравнить полученный результат с тем, прогнозом кредитной истории который дал кредитный комитет банка, принимая решение о выдаче кредита. В данном случае можно расценивать выдачу кредитов, которые вошли в распознаваемую выборку как прогноз о принадлежности заемщиков к классу «обязательства исполняются». Тогда выходит, что кредитный комитет дал правильный прогноз по 37

заемщикам из 50, что соответствует 74% верной идентификации. Это меньше чем прогноз, полученный с помощью скоринговой модели (86,5% верной идентификации).

Для того, чтобы определить точность прогноза скоринговой модели по заемщикам относящимся к классу «Обязательства не исполняются» проанализируем еще одну распознаваемую выборку из 50 человек в состав которой войдут клиенты, получившие отказ со стороны банка в выдаче кредита. Данных респондентов кредитный комитет своим решением отнес к классу «Обязательства не исполняют».

Распознавание выборки с помощью скоринговой модели показало, что из 50 респондентов только 19 относятся к классу «Обязательства не исполняются», оставшиеся 31 – не относится к данному классу. Достоверность данного прогноза трудно проверить, так как клиенты, попавшие в выборку, не имеют кредитной истории. Однако, учитывая результаты прогноза, полученный по предыдущей выборке, с помощью модели, по нашему мнению, имеет смысл применить полученный результат к данной выборке.

Учитывая, что к классу «Обязательства не исполняются» с помощью модели можно верно отнести 46,2% заемщиков, предположим, что процент идентификации клиентов относящихся к данному классу в анализируемой выборке будет примерно таким же. Из этого следует, что из 50 клиентов к данному классу в действительности относится 41 клиент.

Результаты прогнозирования кредитной истории заемщиков вошедших в выборки выданных и отказных дел представлены в таблице 3.

Из таблицы можно сделать вывод о том, что прогноз, полученный с помощью модели адекватен изменениям в распознаваемой выборке. Так в выборке большинство в которой составляют заемщики с положительной кредитной историей с помощью модели мы смогли верно спрогнозировать 32 заемщика с положительной и 6 с отрицательной. В выборке из 50

заемщиков, состоящей в основном из отказных дел (для которых кредитным комитетом прогнозировалась отрицательная кредитная история) с помощью модели мы смогли спрогнозировать увеличение на 3 (до 9 заемщиков) клиентов с отрицательной кредитной историей и уменьшение на 1 (31 заемщик) с положительной кредитной истории по сравнению с выборкой по выданным кредитам.

ТАБЛИЦА 3 – ДОСТОВЕРНОСТЬ ПРОГНОЗИРОВАНИЯ С ПОМОЩЬЮ ОПТИМИЗИРОВАННОЙ СКОРИНГОВОЙ МОДЕЛИ

	Выборка, состоящая из выданных кредитов		**Выборка, состоящая из отказных дел**	
	Кол-во	**%**	**Кол-во**	**%**
Фактическое количество заемщиков по классу: «Обязательства исполняются»	37	**74**	9	8
Результаты прогнозирования по классу: «Обязательства исполняются»	**32**	**86,5**	**31**	29
Фактическое количество заемщиков по классу: «Обязательства **не** исполняются»	13	26	41	82
Результаты прогнозирования по классу: «Обязательства не исполняются»	**6**	46,2	**19**	46,3
Всего количество достоверных прогнозов с помощью модели	38	**76,5**	28	56

Из представленной таблицы видно, что для выборки, состоящей из отказных дел, в соответствии с созданной моделью прогнозируется более чем в 3 раза высокий уровень отрицательных кредитных историй, чем для выборки, включающей данные о выданных кредитах, а также меньшее количество положительных кредитных историй. Эти параметры разумны и ожидаемы нами и, на наш взгляд, подтверждают адекватность работы как кредитного комитета, так и созданной нами модели. Необходимо отметить, по выборке из выданных кредитов применение созданной модели на 12.5% повышает достоверность прогнозирования положительной кредитной истории по сравнению с традиционным подходом. По выборке из отказных

дел согласно модели прогнозируется довольно значительное количество положительных кредитных историй (31), а мы знаем, что достоверность подобных прогнозов составляет 86.5%.

Итак, на основе обучающей выборки из 400 заемщиков, которым были выданы кредиты, модель выявила зависимости между анкетными данными этих заемщиков и их кредитной историей, и эти зависимости имеют силу в более широкой генеральной совокупности, по отношению к которой обучающая выборка репрезентативна. Для проверки адекватности модели и измерения ее достоверности были использованы две выборки, содержащие данные по 50 заемщикам, которым были выданы кредиты, и по 50 заемщикам, которым было оказано в этом. Данные этих заемщиков не были использованы при синтезе модели. По результатам прогнозирования кредитных историй по этим двум тестирующим выборкам получены результаты, позволяющие сделать следующие ***выводы***:

1. С достоверностью 95% можно утверждать, что созданная модель позволяет отличать выборки из выданных кредитов и отказных дел.

2. Применение модели обеспечивает общую средневзвешенную достоверность прогнозирования кредитной истории заемщиков, данные которых не использовались при синтезе модели, на уровне **76,5%**, а положительной кредитной истории **86,5**%, что на **12,5%** выше, чем при традиционном подходе.

3. *Представленная скоринговая модель может применяться в банковской работе на этапе рассмотрения кредитной заявки в качестве консультационной программы, что позволит улучшить качество кредитного портфеля и одновременно сократить издержки банка связанные с проверкой заемщика.*

Необходимо также отметить, что при эксплуатации модели в адаптивном режиме, когда кредитные истории заемщиков, которым выданы кредиты, будут использоваться для *дополнения* обучающей выборки и пересинтеза модели, достоверность и экономическая эффективность модели будет повышаться.

Литература

1. Луценко Е.В. Автоматизированный системно-когнитивный анализ в управлении активными объектами (системная теория информации и ее применение в исследовании экономических, социально-психологических, технологических и организационно-технических систем): Монография (научное издание). – Краснодар: КубГАУ. 2002. –605с.

2. Лебедев Е.А. Оценка рисков кредитования физических лиц (проблема исследования, ее актуальность, идея решения) / Лебедев Е.А. // Научный журнал КубГАУ [Электронный ресурс]. - Краснодар: КубГАУ, 2006. - № 01(17). - Режим доступа: http://ej.kubagro.ru/2006/01/13/p13.asp.

3. Луценко Е.В. Теоретические основы и технология адаптивного семантического анализа в поддержке принятия решений (на примере универсальной автоматизированной системы распознавания образов "ЭЙДОС-5.1"). – Краснодар: КЮИ МВД РФ, 1996. – 280 с.

4. Луценко Е.В. Интеллектуальные информационные системы: Учебное пособие для студентов специальности: 351400 "Прикладная информатика (по отраслям)". – Краснодар: КубГАУ. 2004. – 633 с.

5. Луценко Е.В., Лебедев Е.А. Определение кредитоспособности физических лиц и риски их кредитования. – М.: Финансы и кредит, ноябрь 2006 – № 32(236).

6. Лебедев Е.А. Прогнозирование рисков кредитования физических лиц с применением системно-когнитивного анализа. Научное обеспечение агропромышленного комплекса: материалы 7-й региональной научно-практической конференции молодых ученых. - Краснодар: КубГАУ, 2005 - 450 с.

7. Лебедев Е.А. Синтез скоринговой модели с помощью системно-когнитивного анализа / Лебедев Е.А. // Научный журнал КубГАУ [Электронный ресурс]. - Краснодар: КубГАУ, 2007. - № 29(05).

8. Лебедев Е.А. Прогнозирование рисков при кредитовании физических лиц на основе применения новых математических и инструментальных методов экономики (скоринг) / Лебедев Е.А. // Научное издание «Математические методы и информационно-технические средства» Труды 2 Всероссийской научно-практической конференции 23 июня 2006г. – Краснодар: Краснодарский университет МВД России, 2006. С.45-46.