ВІСНИК ПРИАЗОВСЬКОГО ДЕРЖАВНОГО ТЕХНІЧНОГО УНІВЕРСИТЕТУ
2011 р. Серія: Економічні науки № 2 (22)
ISSN 2225-6725

ФІНАНСИ І БАНКІВСЬКА СПРАВА

УДК 336.051:303.7

© Минц А.Ю.[1] Хаджинова Е.В.[2]

СОВРЕМЕННЫЕ МЕТОДЫ АНАЛИЗА ДАННЫХ В ФИНАНСОВО-КРЕДИТНОЙ СФЕРЕ

В статье проанализированы современные методы математического анализа зависимостей в финансовых данных. Рассмотрены программные продукты для автоматизации анализа данных, а также рассмотрен пример использования современных инструментальных средств для решения задачи анализа данных в коллекторском скоринге.
Ключевые слова: *Data Mining, Deductor, поиск зависимостей, коллекторский скоринг, логистическая регрессия.*

Мінц О.Ю., Хаджинова О.В. Сучасні методи аналізу даних в фінансово-кредитній сфері. *У статті проаналізовано сучасні методи математичного аналізу залежностей у фінансових даних. Розглянуто програмні продукти для автоматизації аналізу даних, а також розглянуто приклад використання сучасних інструментальних засобів для вирішення завдання аналізу даних у колекторському скорінгу.*
Ключові слова: *Data Mining, Deductor, пошук залежностей, колекторський скорінг, логістична регресія.*

O.Y. Mints, O.V. Khadzhinova. Modern methods of data processing in the domain of finance and credit. *The article analyzes the modern methods of mathematical analysis of the dependence in financial data. Reviewed software products to automate data analysis, and considered an example of using modern tools to solve the problem of data analysis in the debt collection scoring.*
Keywords: *data processing, Deduction data analysis, debt collection scoring, logistic regression.*

Постановка проблемы. Несмотря на то, что экономические задачи становятся всё разнообразнее и сложнее, инструментарий их решения остается практически без изменений с середины 1960-х годов и содержит преимущественно методы линейного анализа. В то же время известно, что линейные зависимости не всегда подходят для описания реальных экономических процессов.

Основной причиной использования линейных методов, с точки зрения авторов, является их достаточно простой математический аппарат, доступный для понимания любому квалифицированному экономисту. При этом бытует мнение, что для эффективного использования метода необходимо обязательно знать и понимать все его математические компоненты. Это вызывает парадоксальную ситуацию: повышение эффективности современных методов достигается путем усложнения их математического аппарата. В то же время это усложнение вызывает сокращение количества пользователей метода, ввиду того, что понимание всех математических процедур требует специальной подготовки, далеко выходящей за рамки экономического образования.

Анализ исследований и публикаций. Развитие методов анализа данных может быть отслежено с XVIII века. Так, задачи восстановления зависимостей в рамках методов экономиче-

[1] *канд.экон.наук, доцент, ГВУЗ «Приазовский государственный технический университет», г. Мариуполь.*
[2] *канд.экон.наук, доцент, ГВУЗ «Приазовский государственный технический университет», г. Мариуполь.*

ISSN 2225-6725

ской статистики изучаются с момента разработки К. Гауссом в 1794 г. метода наименьших квадратов. В 1890-х годах К. Пирсоном, Ф. Эджуортом и Р. Уэлдоном был введен линейный коэффициент корреляции. Разработка методов аппроксимации данных и сокращения размерности описания была начата в начале XX века, когда К. Пирсон создал метод главных компонент.

В начале XX века в работах А.А. Маркова было сформулировано определение процессов, получивших в дальнейшем названия «Марковских». Теория Марковских процессов позволяет описать ситуации, в которых «будущее» не зависит от «прошлого» при известном «настоящем».

Важнейшая задача линейной оптимизации экономических процессов была решена в 1939 году, когда Л. В. Канторович опубликовал работу «Математические методы организации и планирования производства», в которой заложил основы линейного программирования. Сам термин «программирование» был предложен в середине 1940-х годов другим основоположником метода Дж. Данцигом и в английском языке является синонимом слова «планирование».

Большая группа методов анализа данных основана на моделировании биологических структур. Среди них наиболее известны нейронные сети. Изначально они были предложены в 1950-х годах Ф. Розенблаттом для решения задач распознавания образов, однако впоследствии стали применяться и для решения других задач, в частности экономических.

В 1970х-1990х годах предложен ряд алгоритмов автоматического построения деревьев решений, среди которых наибольшую известность приобрели алгоритмы ID3, C4.5. Деревья решений состоят из некоторого количества условий «если - то» и обычно используются для решения задач классификации данных.

Накопление методов автоматизированного анализа данных привело к выделению их в отдельную область, получившую название Data Mining (буквально – «Раскопка данных»). Концепция данного направления и сам термин были предложены в 1989 году. Тем не менее, до сих пор в публикациях отечественных ученых (за исключением специалистов в области экономико-математического моделирования) эти методы встречаются крайне редко.

Цель статьи – рассмотреть возможности современных методов анализа данных и их роль в повышении эффективности научных исследований в финансово-кредитной сфере. По мнению автора, необходимо расширять использование современных методов в практике экономических исследований. При этом акцент должен ставиться не на вычислительные тонкости, а на постановку задачи и интерпретацию полученных результатов. Действительно, некоторые методы поиска решений экономических задач (например, нейросетевые модели) изначально представляются в виде «черного ящика», осуществляющего преобразование набора входных данных в набор выходных. При этом решение задачи извлечения знаний из нейронных сетей относится к отдельной области соответствующей теории и до настоящего времени в полном объеме не осуществлено.

Изложение основного материала. Одним из оправданий использования только традиционных методов анализа данных может быть то, что до настоящего времени одним из основных программных инструментов, применяемых для экономических расчетов, в том числе в научной среде, является Microsoft Excel. В рамках этого программного продукта реализованы только основные математические и статистические функции, а также средства визуализации. Реализация большинства современных методов анализа данных в среде Excel затруднительна, если не невозможна.

В то же время рядом компаний, специализирующихся на разработке программного обеспечения, предлагаются продукты, позволяющие работать с современными методами анализа данных без специальной математической подготовки, в интерактивном режиме. Тема анализа возможностей этих методов и их программной реализации несомненно является актуальной.

Рассмотрим направления применения современных методов анализа данных в финансово-кредитной сфере и их краткую характеристику. Наиболее общая классификация делит задачи анализа данных на *описательные* и *предсказательные*. Первые связаны с описанием скрытых закономерностей в имеющихся данных и выведении семантических правил, которые могут использоваться в дальнейшем для повышения эффективности работы. К этому классу относится большое количество маркетинговых задач, связанных с анализом различных целевых групп и выявлением предпочтений их участников. Предсказательные задачи связаны с построением модели, которая может использоваться для предсказания результатов поведения анализируемой

ISSN 2225-6725

системы в тех случаях, данных о которых пока нет. К ним относятся задачи предсказания банкротства, прогнозирования финансовых временных рядов и многие другие.

В качестве методов решения описательных задач могут рассматриваться:

поиск ассоциативных правил,

группировка объектов или кластеризация,

построение регрессионной модели.

Для решения предсказательных задач могут использоваться, например, следующие методы:

классификация объектов по заранее заданным классам,

построение регрессионной модели.

Как видно из приведенного перечня, множества методов, используемых для решения описательных и предсказательных задач, пересекаются. Это объясняется тем, что значительную роль в процессе анализа данных играет постановка задачи. При этом одна и та же задача, например, задача анализа биржевых данных, может быть поставлена и как задача классификации [1], и как задача регрессии [2], и как задача кластеризации [3].

Среди других примеров задач анализа данных в финансово-кредитной сфере можно отметить:

выявление мошенничества с кредитными карточками,

сегментация клиентов,

прогнозирование изменений клиентуры,

анализ финансовых рисков,

прогнозирование банкротств,

поиск зависимостей между показателями.

Математический аппарат современных методов решения перечисленных задач, как уже отмечалось, является достаточно сложным и требует математического образования для понимания его специфики. Однако существующее программное обеспечение позволяет исследователю не вникать в тонкости реализации метода и ограничиться постановкой задачи и интерпретацией результатов.

В настоящее время на рынке программного обеспечения представлено достаточно большое количество пакетов обработки данных. Типичным представителем многофункциональных инструментов анализа данных является программный пакет STATISTICA компании StatSoft (США, веб-сайт http: // www. statsoft. ru/), который содержит большинство известных методов анализа данных. Недостатком данного пакета является относительная сложность освоения и высокая стоимость (более 11000 грн за русифицированную однопользовательскую версию). При этом, однако, существует система скидок (до 50%) для образовательных учреждений. Среди других программных средств аналогичной функциональности и сопоставимой стоимости можно отметить SAS (компания SAS Institute, $8300, веб-сайт https: // www. sas. com), IBM SPSS Statistics (IBM-SPSS, $2405, веб-сайт http: // www - 01. ibm. com / software / analytics / spss/), Statgraphics (StatPoint Technologies, $1495, веб-сайт http: // www. statgraphics. com). Все они разработаны зарубежными компаниями.

Среди программных продуктов, разработанных в ближнем зарубежье, можно отметить пакет STADIA (разработчик – НПО «Информатика и компьютеры», веб-сайт http: // statsoft. msu. ru) и аналитическую платформу Deductor (разработка BaseGroup Labs, веб-сайт http: // www. basegroup. ru/). Преимуществом Stadia является относительно невысокая цена (2500 грн. за базовую версию и 3700 грн. за профессиональную), однако возможности этого пакета несколько ниже, по сравнению с другими рассмотренными программными продуктами.

Стоимость аналитической платформы Deductor несколько ниже большинства конкурентов, кроме Stadia, и составляет 7250 грн.. При этом аналитические возможности находятся на уровне зарубежных аналогов. Особенностью продукта являются широкие возможности по анализу адекватности полученных моделей, что определяет практическую направленность платформы.

Каждый из перечисленных пакетов кроме основных (платных) версий имеет также и ознакомительную (бесплатную). В большинстве случаев аналитические возможности ознакомительной версии повторяют возможности базовой, однако программно блокированы возможности сохранения результатов, и (или) ввода пользовательских данных. В некоторых системах

ISSN 2225-6725

(например, Stadia) дополнительно накладывается ограничение на величину анализируемой выборки данных. Единственным исключением из этого правила является аналитическая платформа Deductor, в которой пробная версия формально отсутствует, но есть версия *Academic*, предназначенная для образовательных целей и являющаяся бесплатной. В этой версии присутствует ряд ограничений, препятствующих её эффективному коммерческому использованию, однако ввод пользовательских данных и сохранение результатов возможны, что делает данный пакет наиболее предпочтительным для использования не только в образовательных, но и в научно-исследовательских целях, при отсутствии средств на приобретение полнофункциональных версий аналитических программ.

Рассмотрим некоторые возможности аналитической платформы Deductor на примере анализа данных о банковских заемщиках в задаче коллекторского скоринга. Сущность этой задачи состоит в анализе кредитозаемщиков, допустивших просрочку в погашении своих обязательств по кредитам, и выделении среди них тех, с кем банку необходимо работать в первую очередь, то есть заемщиков, которые после проведения коллекторских мероприятий с наибольшей вероятностью могут возобновить платежи.

Входная выборка данных включает информацию о клиентах банка, с которыми уже была проведена такая работа, а также результаты этой работы, то есть, были ли возобновлены платежи. Вектор входных данных содержит следующие поля: пол, возраст, сумма просрочки, сумма ежемесячного платежа, отношение просрочки к ежемесячному платежу, сумма кредита, информация о возобновлении платежа. Поскольку исходные данные чаще всего находятся в таблице Excel, их необходимо предварительно сохранить в формате «Текстовые файлы (с разделителями табуляции)», который поддерживает система Deductor. При импорте файла система автоматически распознает типы данных в полях таблицы, однако в ряде случаев может потребоваться указать типы вручную (рис. 1).

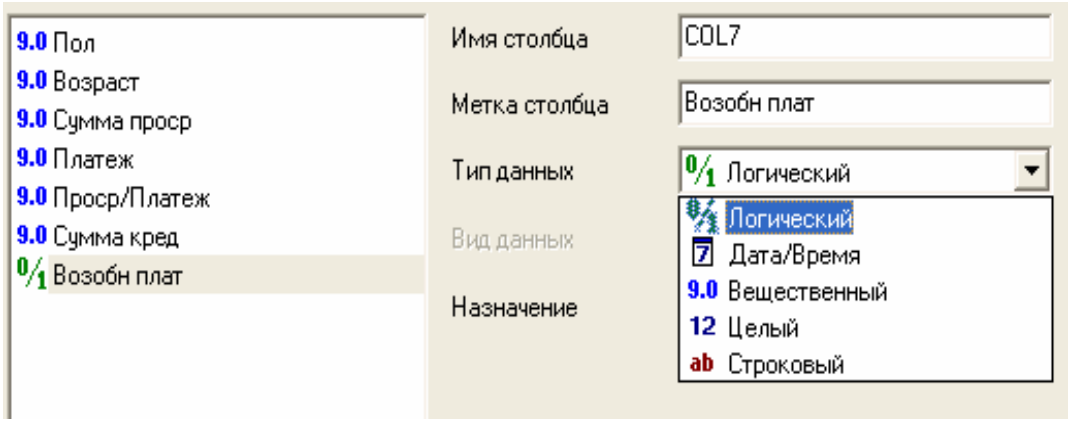

Рис. 1 - Определение типа данных при импорте

После импорта данные могут быть просмотрены в табличном виде. При этом по каждому полю автоматически рассчитываются основные статистические показатели, и строится гистограмма распределения.

Методы обработки данных в Deductor разделены на 4 группы:

очистка данных,

трансформация данных

Data Mining

прочее.

Методы очистки данных позволяют произвести восстановление, редактирование и сглаживание данных, произвести понижение размерности входных факторов, корреляционный анализ, выявление дубликатов и противоречий в данных, фильтрацию строк по заданному условию. В рассматриваемой выборке данных дубликатов и противоречий выявлено не было. Результаты корреляционного анализа между входными полями и полем «возобновление платежа» приведены на рис. 2.

ISSN 2225-6725

На основании корреляционного анализа может быть принято решение об исключении некоторых полей из дальнейшей обработки. Однако это целесообразно делать либо в тех случаях, когда необходимо понижать размерность вектора входных данных, либо в тех случаях, когда сильная корреляция наблюдается между различными входными полями.

Входные поля		Корреляция с выходными полями
№	Поле	Возобн плат
1	Пол	0,032
2	Возраст	0,055
3	Сумма проср	-0,141
4	Платеж	-0,160
5	Проср/Платеж	-0,037
6	Сумма кред	-0,111

Рис. 2 - Результаты корреляционного анализа данных

На основании корреляционного анализа может быть принято решение об исключении некоторых полей из дальнейшей обработки. Однако это целесообразно делать либо в тех случаях, когда необходимо понижать размерность вектора входных данных, либо в тех случаях, когда сильная корреляция наблюдается между различными входными полями.

Методы трансформации данных позволяют осуществить технические действия, связанные с приведением данных в вид, удобный для дальнейшего анализа, в частности, сортировку, группировку, замену данных и т.п. В рассматриваемом случае трансформация данных не используется.

В рамках методов группы Data Mining в системе Deductor реализованы следующие: автокорреляция, линейная регрессия, логистическая регрессия, многослойная нейронная сеть, построение дерева решений по алгоритму C4.5, самоорганизующиеся карты Кохонена, поиск ассоциативных зависимостей, задание модели вручную по формулам и кластеризация алгоритмами k-means или g-means.

Применительно к данной задаче может быть рассмотрено использование бинарной логистической модели. Процедура её построения в системе Deductor включает: определение входных и выходных данных; задание параметров разбиения входного множества на «обучающее» и «тестовое»; настройку параметров обучения, анализ полученной модели.

Окно процедуры назначения данных модели показано на рис. 3.

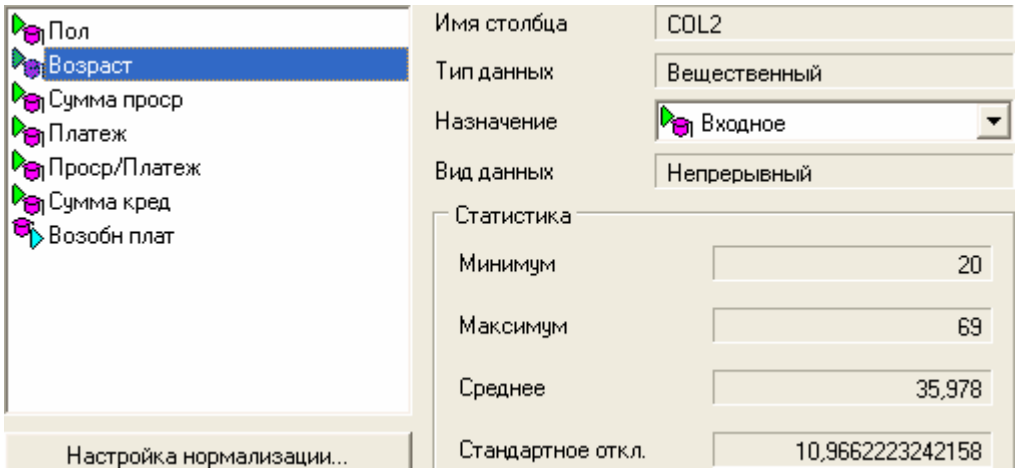

Рис. 3 - Определение назначения данных модели

Выделение во входном множестве данных тестовой и обучающей выборки позволяет организовать проверку качества обучения и избежать ситуации, когда построенная модель «запоминает» входную выборку данных. Практика показывает, что для тестовой выборки достаточно

ISSN 2225-6725

выделить 10% всего массива данных.

Настройка параметров обучения позволяет уточнить некоторые параметры обучающих алгоритмов. При этом изначально система Deductor предлагает параметры по умолчанию, которые обычно обеспечивают достаточную сходимость алгоритма.

После завершения автоматической процедуры обучения модели пользователю предлагается настроить параметры визуализации результатов и выбрать, какие инструменты визуализации будут использоваться. Параметры построенной модели показаны на рис. 4.

Атрибут	Коэффициент	Отношение шансов
9.0 <Константа>	-0,856	
0/1 Пол		
False	0	1
True	0,24481	1,2774
9.0 Возраст	-0,0005101	0,99949
9.0 Сумма проср	-6,9161E-5	0,99993
9.0 Платеж	-0,00092628	0,99907
9.0 Проср/Платеж	-0,02748	0,97289
9.0 Сумма кред	7,5347E-5	1,0001

Рис. 4 - Параметры модели логистической регрессии

Построенная модель может использоваться для анализа данных, не входящих в исходную выборку. Для этого используется инструмент анализа «Что-если», позволяющий интерактивно вводить пользовательские данные и получать результат их обработки моделью.

Для анализа адекватности модели могут использоваться как встроенные средства (таблица сопряженности результатов, ROC-кривая), так и задаваемые пользователем.

Таблица сопряженности результатов (рис. 5) применяется при анализе классифицирующих моделей и позволяет оценить ошибки классификации первого и второго рода. Ошибка первого рода соответствует положительной классификации отрицательного исхода. Ошибка второго рода – отрицательной классификации положительного исхода.

	Классифицировано		
Фактически	False	True	Итого
False	257	160	417
True	33	50	83
Итого	290	210	500

Рис. 5 - Матрица сопряженности результатов

В реальной экономике ошибка первого рода соответствует убытку по совершенной операции, а ошибка второго рода – недополученной прибыли по несовершенной операции. При этом стоимость ошибок первого рода обычно выше, чем ошибок второго рода.

Аналогичной цели служит такой визуальный инструмент анализа, как ROC-кривая (рис.6), которая показывает соотношение количества верно классифицированных положительных примеров и неверно классифицированных отрицательных примеров, в зависимости от значения порога отсечения.

Важным параметром в ROC-анализе является площадь под кривой, которая для рассматриваемого примера равна 0.652. При визуальном анализе преимущественно сопоставляют взаимное расположение положение двух и более ROC-кривых. Кроме того, система Deductor позволяет строить разновидность ROC-кривой, кривые баланса, с помощью которых может быть найдено оптимальное значение порога отсечения бинарной модели.

Однако в рассматриваемой задаче целью является не получение ответа: возвратит, или не возвратит заемщик кредит (это соответствует задачам аппликационного скоринга), а ранжирование заемщиков в порядке убывания вероятности возврата.

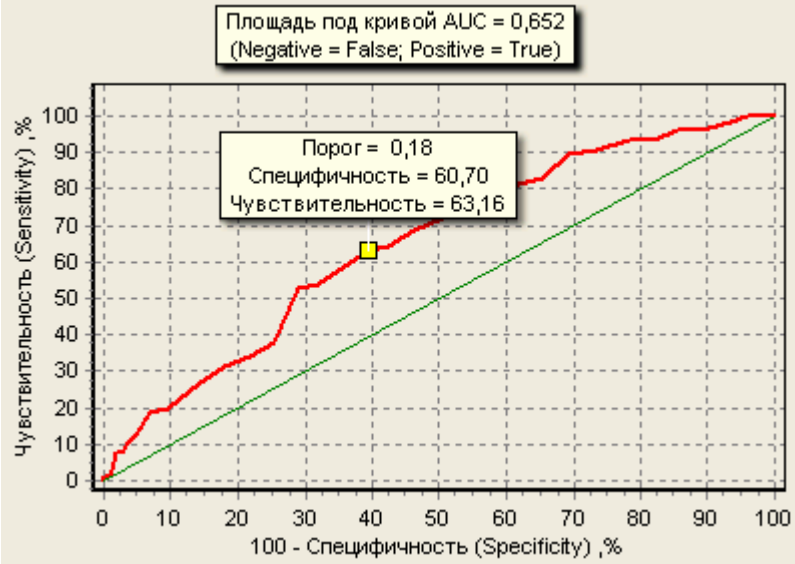

Рис. 6 - ROC-кривая построенной модели

Это значение рассчитывается моделью логистической регрессии и может быть отслежено в окне инструмента «Что-если». Однако для более наглядного анализа эффективности работы модели могут быть использованы дополнительные инструменты, такие как Lift-кривые. Данный инструмент отсутствует среди встроенных в систему Deductor, однако может быть создан пользователем (рис. 7).

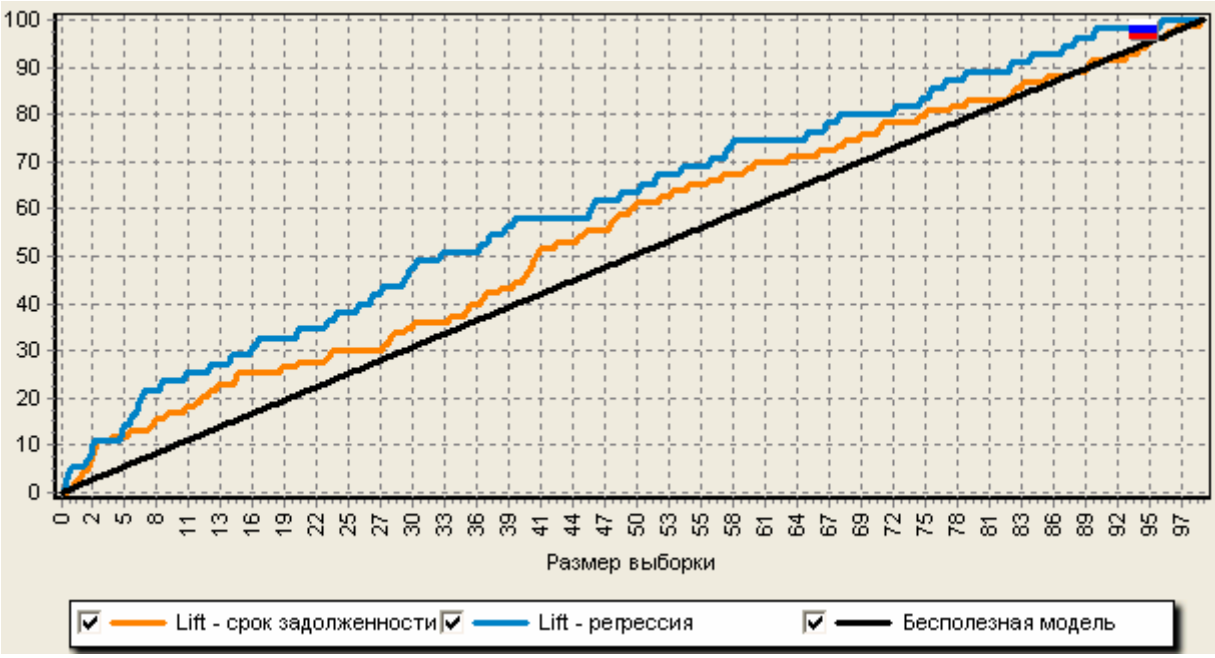

Рис. 7 - Lift-кривые анализа различных моделей ранжирования заемщиков

Lift-кривая интерпретируется, как показатель, отражающий увеличение числа откликов относительно числа действий. При построении Lift-кривой выборка упорядочивается по убыванию вероятности положительного события. После этого на графике по горизонтали откладывается размер выборки в процентах от общего числа наблюдений, а по вертикали фиксируется количество положительных исходов, взятое нарастающим итогом. На рис. 7 показаны Lift-кривые моделей логистической регрессии (верхняя кривая), популярной в коллекторском скоринге эмпирической модели ранжирования заемщиков по сроку задолженности (средняя кри-

ISSN 2225-6725

вая) и бесполезной модели, которой соответствует случайный выбор заемщиков из списка.

Из рис. 7 видно, что лучшие результаты показывает модель логистической регрессии. Практическое значение полученных результатов проявляется в уменьшении количества кредитных дел, которые необходимо обработать, чтобы достичь некоторого результата. Так, для того, чтобы провести работу с половиной должников, которые потенциально могут возобновить выплаты по кредиту, в «бесполезной» модели необходимо обработать 50% кредитных дел, с использованием модели логистической регрессии – 31% кредитных дел, а при сортировке заявок по сроку просрочки – 41% дел.

Кроме модели логистической регрессии, для решения этой же задачи в рамках пакета Deductor с различным успехом могут быть использованы нейросетевые модели, самоорганизующиеся карты Кохонена, деревья решений, ассоциативные правила. Общие принципы использования всех перечисленных методов от рассмотренных выше отличаются незначительно.

Выводы

Возможности современного программного обеспечения позволяют исследователю дистанцироваться от изучения математической сущности методов анализа данных и сосредоточиться на постановке задачи и анализе полученных результатов. Это существенно повышает производительность и эффективность труда исследователя за счет уменьшение временных затрат на процесс моделирования и расширения спектр используемых методов анализа.

Список використаних джерел:

1. Лысенко Ю.Г. Поиск эффективных решений в экономических задачах / Ю.Г. Лысенко, А.Ю. Минц, В.Г. Стасюк. – Донецк : ДонНУ; ООО “Юго-Восток, Лтд”, 2002. – 101 с.
2. Ежов А.А. Нейрокомпьютинг и его применения в экономике и бизнесе / А.А. Ежов, С.А. Шумский. - (серия «Учебники экономико-аналитического института МИФИ» под ред. проф. В.В. Харитонова). М. : МИФИ, 1998. – 224 с.
3. Минц А.Ю. Прогнозирование валютных рынков с использованием самоорганизующихся нейронных сетей / А.Ю. Минц // Вісник СНУ ім. В.Даля. – 2004. №4(74). – С. 184-193.

Bibliography:

1. Lysenko Y.G. Searching for effective solutions to economic tasks. / Y.G. Lysenko, O.Y. Mints, V.G. Stasyuk. – Donetsk : DonNU; Yugo-Vostok Ltd., 2002. – 101p. (Rus.)
2. Ezhov A.A. Neurocomputing and its applications in economics and business / A.A. Ezhov, S.A. Shumskiy. - M. : MIFI, 1998. – 224p. (Rus.)
3. Mints O.Y. Prediction of foreign currency markets using self-organizing neural networks / O.Y. Mints // Visnyk SNU im V. Dalya. – 2004. №4(74). – P.184-193. (Rus.)

Рецензент: Т.Г. Логутова
д-р економ. наук, проф., ГВУЗ «ПГТУ»

Стаття надійшла15.11.2011