Лабораторная работа №7

Повышение эффективности массовой рассылки клиентам

Постановка задачи. Торговая компания, осуществляющая продажу товаров, располагает информацией о своих клиентах и их покупках. Компания провела рекламную рассылку 13 504 клиентам и получила отклик в 14,5 % случаев. Необходимо построить модели отклика и проанализировать результаты, чтобы предложить способы минимизации издержек на новые почтовые рассылки.

Исходные данные. Набор данных содержит информацию о 13 504 клиентах, включая известные отклики на рекламную рассылку и такие сведения, как пол; возраст; сколько лет данный человек является клиентом компании; суммарная стоимость всех заказов клиента; общее число покупок; факты обращений в службу поддержки и др. Всего для анализа доступно 9 независимых и 1 зависимая переменная. Имеется также следующая информация (будем использовать обозначения, аналогичные принятым в разделе 12.4):

- расходы на одну рассылку $CM = 1{,}0$ ед.;
- издержки на обслуживание клиента $CR = 9{,}0$ ед.;
- ожидаемая выручка с 1 заказа $R = 20{,}0$ ед.

Решение задачи

Сформулированная выше бизнес-задача сводится к бинарной классификации. Но сначала разберемся в типах ошибок классификации применительно к массовой рассылке (табл. 19.1).

Таблица 19.1. Доход-издержки в массовой рассылке

Исход	Предсказано моделью	Фактически	Доход, ед.	Комментарий
TN	Нет отклика	Нет отклика	0	Нет контакта, нет издержек
TP	Есть отклик	Есть отклик	$(R - CM - CR) = 10{,}0$	Ожидаемая выручка минус расходы на обслуживание и рассылку
FN	Нет отклика	Есть отклик	0	Нет контакта, нет издержек
FP	Есть отклик	Нет отклика	$CM = -1{,}0$	Печать, упаковка и доставка по почте рассылки

Заметим, что в столбце «Доход, ед.» издержки записаны как отрицательный доход.

Формула для расчета дохода будет иметь вид:

$$P = TP \cdot (R - CM - CR) - FP \cdot CM.$$

Чтобы оценить прогностическую силу классификационной модели, необходимо с чем-то сравнить эффективность ее работы — в данном случае с моделью «разослать всем». То есть мы рассчитаем, какую прибыль получим, если проведем рассылку всем клиентам (при этом моделирования не потребуется). Более того, такой расчет нужно выполнять не на обучающем, а на тестовом множестве. Объем тестового множества обычно варьируется от 10 до 50 %; мы выберем 40 % (5402 записи, из которых в 781 случае есть отклик, в 4621 — отклик отсутствует). Тогда доход составит $(781 \cdot 10{,}0 - 4621 \cdot 1{,}0) = 3189$ ед.

Нам нужна модель, которая будет давать доход больше этой величины.

Импортируем в Deductor два текстовых файла — `responses1.txt` и `responses2.txt` с обучающим и тестовым множествами соответственно. Открыв статистические характеристики, обнаружим, что доля клиентов с положительным откликом в обучающем множестве составляет 14,5 %. Иначе говоря, распределение классов в выходной переменной неравномерное. В таких случаях нежелательно строить модель на всем доступном множестве примеров, а рекомендуется предварительно уравновесить их (см. раздел 12.6).

рассматривая бизнес-задачу о кредитном скоринге, мы уже сталкивались с несбалансированностью классов, для борьбы с которой использовали правило Байеса для расчета порога в логистической регрессии и процедуру отбора

с дублированием миноритарного класса (oversampling) в дереве решений. В задаче о массовой рассылке мы построим предсказательные модели двумя методами — снова с помощью логистической регрессии и нейронной сети (многослойный персептрон, алгоритм BackProp). Для этого в Deductor имеются соответствующие узлы-обработчики. Число примеров в нашем случае достаточное, так что для нейронной сети уравновесим их с помощью процедуры удаления примеров мажоритарного класса (undersampling). Отметим еще раз, что при проверке модели уравновешивать тестовое множество ни в коем случае нельзя, поскольку мы смотрим, как будет вести себя модель в реальных условиях.

Перед моделированием полезно оценить влияние входных переменных на выходную. Воспользуемся обработчиком **Корреляционный анализ** и откроем визуализатор **Матрица корреляции** (рис. 19.1). Там мы, в частности, видим, что количество покупок и потраченные суммы сильно влияют на отклик на рассылку и демонстрируют положительную связь.

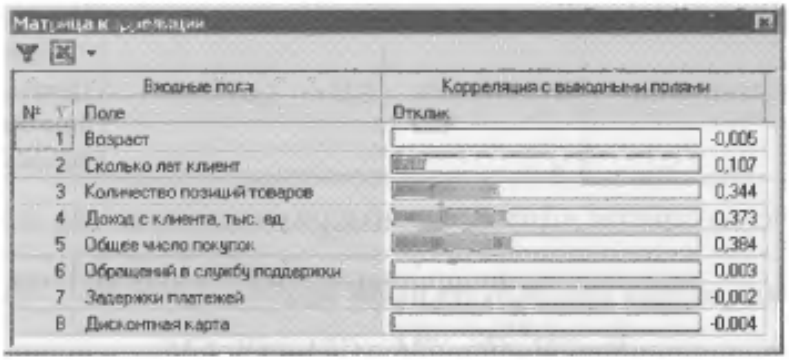

Рис. 19.1. Корреляция с полем «Отклик»

Приступим к построению моделей Data Mining. Импортируйте в сценарий текстовые файлы `responses1.txt` и `responses2.txt` и постройте модель логистической регрессии на данных первого файла, задав входные и выходное поля, как показано на рис. 19.2. Положительным исходом будет считаться наличие отклика.

Из табл. 19.1 следует, что отношение издержек обоих типов ошибок равно 10 : 1. Согласно правилу Байеса порог отсечения нужно установить равным 10 / 11 = 0,909.

На следующем шаге добавьте обработчик **Скрипт** к узлу импорта второго файла. Скрипт должен ссылаться на модель логистической регрессии. В результате вы получите вероятности отклика для каждой записи из тестового множества (рис. 19.3).

В задачах массовой рассылки не обойтись без анализа Lift-диаграммы (см. раздел 12.4). Для ее построения необходимо отсортировать примеры в порядке уменьшения вероятности положительного исхода. Это легко сделать, так как в выходном наборе данных узла **Логистическая регрессия** присутствует поле **Отклик Рейтинг**. Для других алгоритмов машинного обучения вопрос, по какому полю сортировать, решается каждый раз отдельно. Построим Lift-кривую, когда по оси *oy* откладывается кумулятивный процент откликов. Фрагмент сценария, реализующий формирование лифт-таблицы, состоит из пяти узлов (с учетом сортировки) и представлен на рис. 19.3.

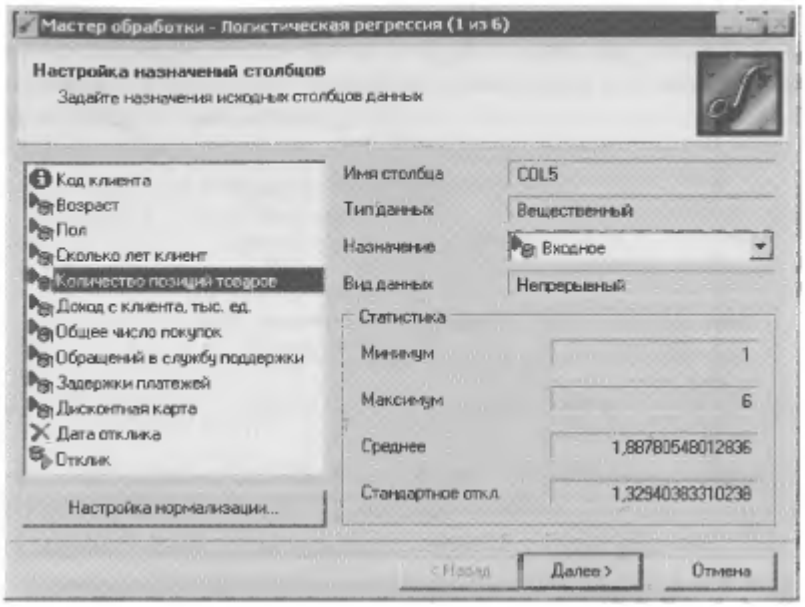

Рис. 19.2. Установка назначения столбцов

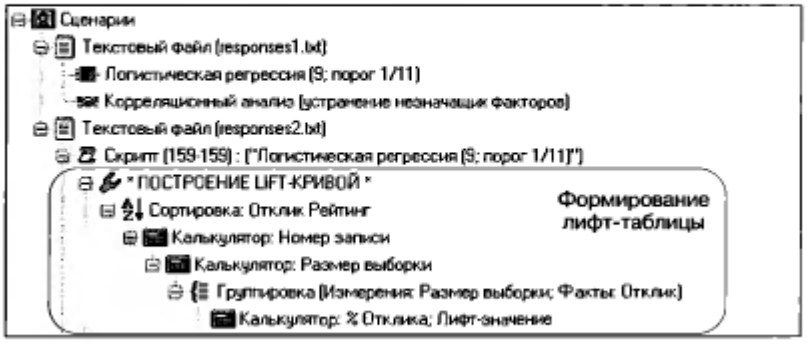

Рис. 19.3. Сценарий к задаче о массовой рассылке

На рис. 19.4 изображена получившаяся Lift-кривая. Диагональная линия отражает работу бесполезного классификатора, то есть ситуацию, когда списки получателей рассылки формируются случайным образом. График кривой, соответствующей нашей модели, проходит достаточно высоко, что говорит о хорошем качестве прогнозирования клиентского отклика. Видим, что при объеме рассылки, равном 25 % от тестового множества, мы получим около 75 % всех возможных откликов. Если бы мы проводили рассылку по принципу случайности, то для получения такого же отклика нам пришлось бы отправить письма 74 % клиентов. Разница в 49 % и есть эффект, который даст нам стратегия рассылки, с использованием модели логистической регрессии, когда в первую очередь письма будут отправляться респондентам с максимальными вероятностями отклика.

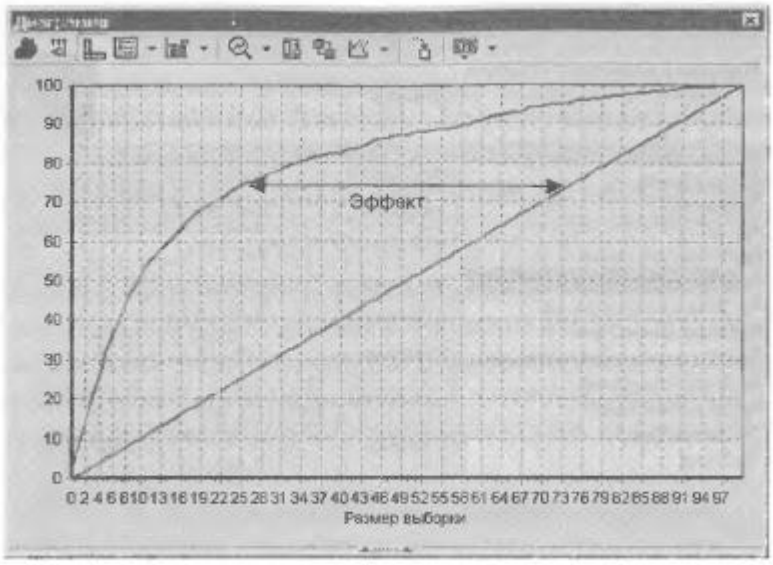

Рис. 19.4. Lift-кривая для модели отклика на основе логистической регрессии

Теперь представьте, что компания ранее использовала правило, согласно которому рассылка производилась в первую очередь тем клиентам, которые принесли наибольшие доходы. Отсортируем записи по убыванию значений поля *Доход с клиента*, повторим расчеты и построим Lift-кривую (рис. 19.5). Она окажется гораздо хуже Lift-кривой для логистической регрессии, но при малых объемах рассылки (до 6 %) не уступит ей в эффективности.

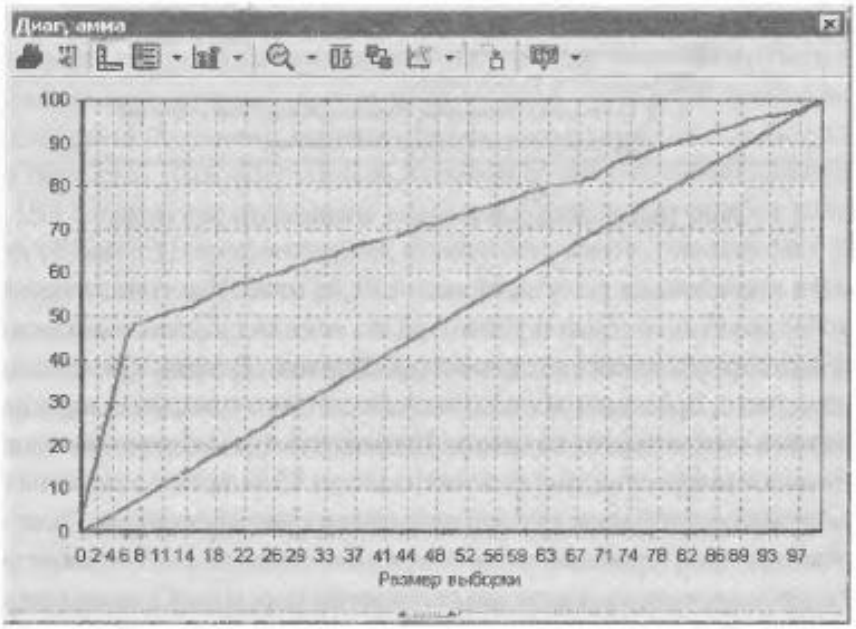

Рис. 19.5. Lift-кривая для случая рассылки по убыванию значений поля «Доход с клиента»

Теперь построим модель нейронной сети. Процедура undersampling предполагает, что придется пожертвовать примерно 9/10 примеров с клиентами, от которых не было отклика. В сценарии Deductor это реализуется с помощью обработчиков **Фильтр**, **Калькулятор** и **Слияние** (рис. 19.6).

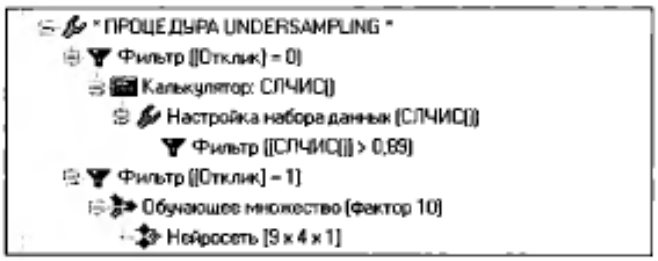

Рис. 19.6. Фрагмент сценария, выполняющий процедуру undersampling

При настройке обработчика **Нейросеть** первые два шага аналогичны шагам узла **Логистическая регрессия**. Третий шаг мастера отвечает за архитектуру многослойного персептрона и параметры активационной функции (рис. 19.7).

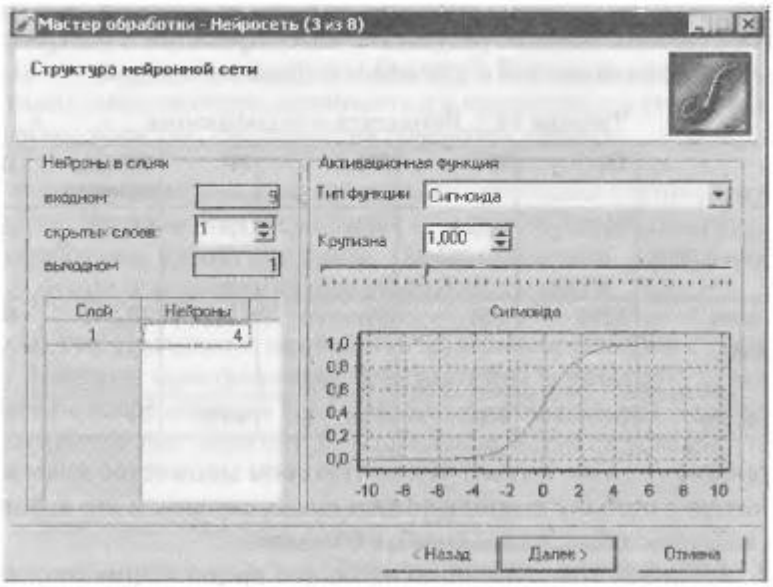

Рис. 19.7. Настройка структуры нейронной сети

Для нашей задачи достаточно одного скрытого слоя с четырьмя нейронами. Для обучения выберем алгоритм BackProp, после обучения — визуализатор **Граф нейросети**. Он позволяет графически представить нейронную сеть со всеми ее нейронами и синаптическими связями. При этом можно увидеть не только структуру многослойного персептрона, но и значения весов, которые принимают те или иные нейроны. В зависимости от веса нейрона он отображается определенным цветом, а соответствующее значение можно определить по цветовой шкале, расположенной внизу окна (рис. 19.8).

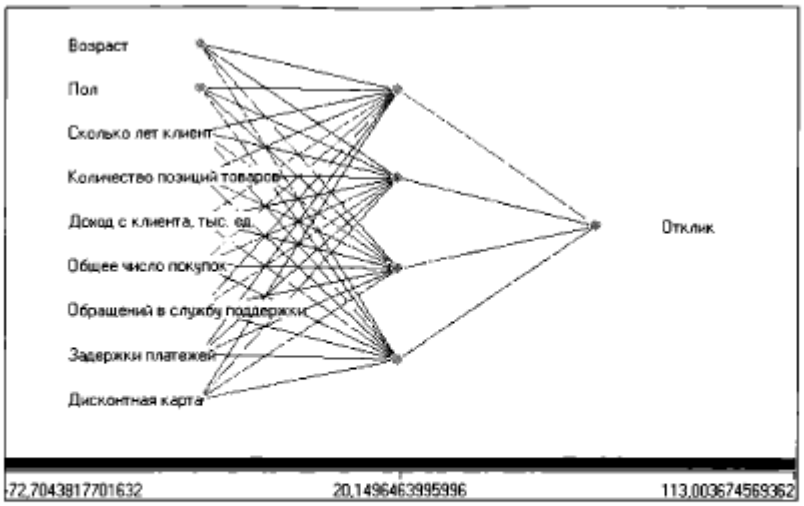

Рис. 19.8. Граф нейросети

В табл. 19.2 сведены воедино результаты классификации и издержки для двух моделей на тестовом множестве и для модели «разослать всем».

В табл. 19.2 сведены воедино результаты классификации и издержки для двух моделей на тестовом множестве и для модели «разослать всем».

Таблица 19.2. Результаты классификации

Модель	TN	TP	FN	FP	Общая ошибка	Доход, ед.
Логистическая регрессия (порог 0,909)	3119	643	138	1502	30,3 %	4928
Нейронная сеть (undersampling, 10 : 1)	3295	744	37	1326	25,2 %	6114
«Разослать всем»	0	781	0	4621	—	3189

Наши усилия привели к тому, что на тестовом множестве классификаторы чаще ошибаются в сторону ложноположительных случаев, и это хорошо, так как отношение издержек обоих типов ошибок большое.

Из табл. 19.2 видно, что лучшей моделью для предсказания отклика клиента становится нейронная сеть, дающая доход 6114 ед. С ее помощью эффективность рассылки увеличивается в 6114 / 3189 = 1,92 раза.

ЗАДАНИЕ.

Реализовать рассмотренный проект (см. постановку задачи). В отчет по лабораторной работе включить принт-скрины хода выполнения работы, сформировать и предоставить файлы проекта.