

Лабораторная работа №5

Скоринговые модели для оценки кредитоспособности заемщиков

Описание бизнес-задачи

Технологии кредитного скоринга — автоматической оценке кредитоспособности физического лица — в банковской среде традиционно уделяется повышенное внимание. Сегодня можно сказать, что экспертные методы уходят в прошлое, и все чаще при разработке скоринговых моделей обращаются к алгоритмам Data Mining. Классическую скоринговую карту можно построить при помощи логистической регрессии на основе накопленной кредитной истории, применив к ней ROC-анализ для управления рисками. Кроме того, хорошие и легко интерпретируемые модели можно получить, используя деревья решений.

Постановка задачи. В коммерческом банке имеется продукт «Нецелевой потребительский кредит»: кредиты предоставляются на любые цели с принятием решения в течение нескольких часов. За это время проверяются минимальные сведения о клиенте, в основном такие, как отсутствие криминального прошлого и кредитная история в других банках.

В банке накоплена статистическая информация о заемщиках и качестве обслуживания ими долга за несколько месяцев. Руководство банка, понимая, что отсутствие адекватных математических инструментов, позволяющих оптимизировать риски, не способствует расширению розничного бизнеса в области потребительского кредитования, поставило перед отделом розничных рисков задачу разработать скоринговые модели с различными стратегиями кредитования, которые позволили бы управлять рисками, настраивая уровень одобрений, и минимизировать число «безнадежных» заемщиков.

Исходные данные. Вообще говоря, информация о заемщиках — физических лицах и кредитных договорах хранится в банковской информационной системе. Там же содержатся графики и даты погашений кредита, сведения о просрочках, об их суммах, о процентах и т. д. Получить для построения скоринговой модели таблицу с параметрами заемщиков и информацию о наличии просрочек — отдельная задача. Будем считать, что она уже выполнена и результат представлен в виде текстового файла.

Важным также является вопрос о том, что понимать под параметрами заемщика.

Здесь уместно обратиться к лекциям, где рассматривались методические аспекты

подготовки и сбора данных для анализа, и вспомнить, что на этом этапе требуется активное взаимодействие с экспертами: они с высоты своего опыта ограничат круг входных переменных, которые потенциально могут влиять на кредитоспособность будущего заемщика. Кроме того, следует учитывать аспекты бизнеса и технические вопросы (например, сложно проверить в короткий срок достоверность признака «Сфера деятельности компании», а потому полагаться на него не стоит).

Скоринговые модели часто строятся на категориальных переменных, и для этого непрерывные признаки квантуются при помощи ручного выбора точек разрыва.

Скажем, переменная *Стаж работы* разбивается на три категории: «до 1 года», «от 1 до 3 лет», «свыше 3 лет». Такую модель легче интерпретировать, но она менее гибкая при моделировании связей: горизонтальные «ступени» дают плохую аппроксимацию при наличии частых крутых «склонов».

В банковской практике перед скорингом заемщик, как правило, проходит процедуру андеррайтинга — проверку на удовлетворение жестким требованиям: соответствие возрасту, отсутствие криминального прошлого и, конечно, наличие определенного дохода. При этом выдвигаются требования к минимальному уровню дохода и рассчитывается возможный лимит кредита. При его расчете участвует один из двух коэффициентов — П/Д либо О/Д.

Коэффициент «Платеж/Доход» (П/Д) — отношение ежемесячных платежей по кредиту заемщика к его доходу за тот же период. Считается, что значительная величина этого коэффициента (свыше 40 %) свидетельствует о повышенном риске как для кредитора, так и для заемщика.

Коэффициент «Обязательства/Доход» (О/Д) — отношение ежемесячных обязательств заемщика к его доходу за тот же период с учетом удержаний налогов. В обязательства включаются расходы, связанные с выплатой планируемого кредита, а также имеющиеся другие долгосрочные обязательства (выплаты по иным кредитам, на содержание иждивенцев, семьи, алиментов, обязательные налоговые платежи и пр.). Считается, что размер ежемесячных обязательств заемщика не должен превышать 50–60 % его совокупного чистого дохода.

Заявки клиентов, не прошедшие андеррайтинг, получают отказ и даже не попадут на скоринг. Поэтому на вход скоринговой процедуры выгоднее подавать не доход клиента, а отношение О/Д или П/Д.

В нашей задаче представлено 2709 кредитов (файл `loans.txt`) с известными исходами платежей на протяжении нескольких месяцев после выдачи кредита. Набор данных уже разбит на два множества — обучающее (80 %) и тестовое (20 %) — при помощи процедуры стратифицированного сэмплинга так, чтобы в каждом множестве доля плохих кредитов была примерно одинаковой. В табл. 17.1 отображены структура и описание полей текстового файла с кредитными историями.

Таблица 17.1. Данные по заемщикам и качеству обслуживания ими долга

№	Поле	Описание	Тип
1	Код	Служебный код заявки	Целый
2	Дата	Дата выдачи кредита	Дата/время
3	О/Д, %	Коэффициент О/Д («Обязательства/Доход») в %	Вещественный
4	Возраст	Возраст заемщика (полных лет) на момент принятия решения о выдаче кредита	Целый
5	Проживание	Основание для проживания: собственник; муниципальное жилье; аренда	Строковый
6	Срок проживания в регионе	Менее 1 года; от 1 года до 5 лет; свыше 5 лет	Строковый
7	Семейное положение	Холост/не замужем; женат/замужем; разведен (-а)/вдовство; другое	Строковый
8	Образование	Среднее; среднее специальное; высшее	Строковый
9	Стаж работы на последнем месте	Менее 1 года; от 1 года до 3 лет; свыше 3 лет	Строковый
10	Уровень должности	Сотрудник; руководитель среднего звена; руководитель высшего звена	Строковый
11	Кредитная история	Информация берется из бюро кредитных историй. Если имеется негативная информация о клиенте (просрочки по прошлым кредитам), то ему присваивается категория «отрицательная»	Строковый
12	Просрочки свыше 60 дней	Факт наличия просрочек свыше 60 дней: 0 — отсутствовали, 1 — имели место	Целый
13	Тестовое множество	Служебный признак, отвечающий за то, к какому множеству относится запись. TRUE соответствует тестовому множеству	Логический

Скоринговая карта на основе логистической регрессии

Базовым статистическим алгоритмом, который строит аналог традиционной балльной скоринговой карты, является логистическая регрессия.

Импортируйте файл с кредитными историями в Deductor. Скоринг представляет собой задачу бинарной классификации, которая относит заемщика к одному из двух классов — «плохой» или «хороший». Если заемщик «хороший» — кредит выдается, если «плохой» — выносится отрицательное решение. Разделение заемщиков на «плохих» и «хороших» осуществляется на основе качества обслуживания ими долга, проще говоря — наличия просрочек. В банковском деле существуют различные шкалы перехода от числа просрочек к классу заемщика, и это тема для отдельного обсуждения. Примем следующее правило: если у клиента была хотя бы одна просрочка свыше 60 дней, то он относится к классу неблагонадежных. Запустите Мастер обработки, в категории Прочие выберите Калькулятор, запишите это условие, в результате чего появится новое вычисляемое поле — *Класс заемщика* (рис. 17.1).

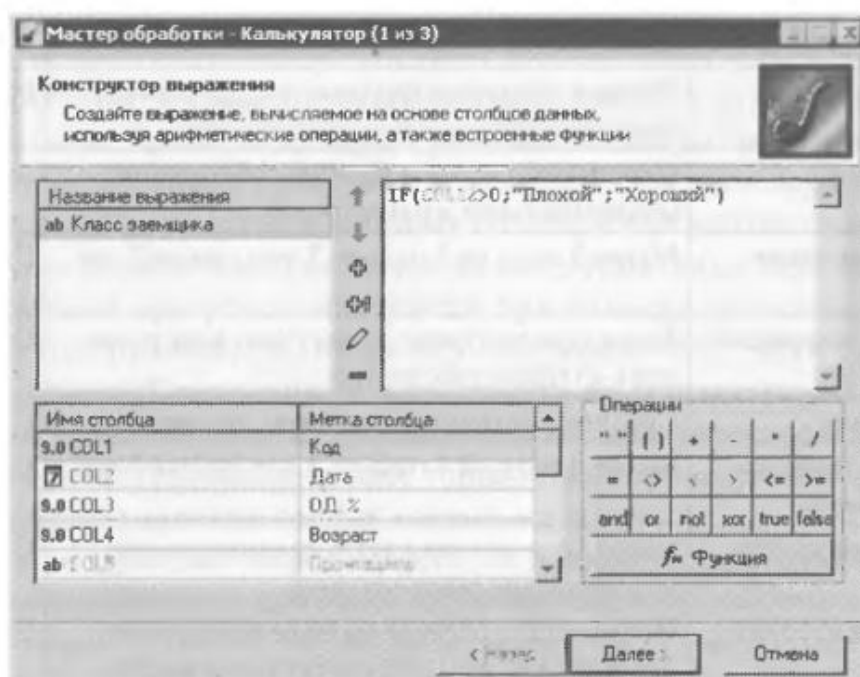


Рис. 17.1. Создание нового поля Класс заемщика

Далее с помощью визуализатора Статистика можно узнать, что имеется 500 записей с «плохими» кредитами, что составляет 18,5 % всех выданных кредитов. Это не так уж и мало: в практике кредитного скоринга число записей миноритарного класса может быть и меньше, вплоть до 1–3 %. Поэтому задача классификации заемщиков всегда решается в условиях сильной несбалансированности классов (см. раздел 12.6).

Таким образом, выходная бинарная переменная — *Класс заемщика* — у нас уже имеется. В качестве входных имеет смысл оставить все, кроме *Код* и *Дата*: очевидно, что они никак не влияют на кредитоспособность.

Поля *Возраст* и *О/Д, %* оставьте непрерывными.

Построим модель логистической регрессии, которая рассчитает соответствующие коэффициенты регрессии. Для этого в сценарий после узла квантования добавьте обработчик Логистическая регрессия. Установите входные и выходные поля, как это показано на рис. 17.2.

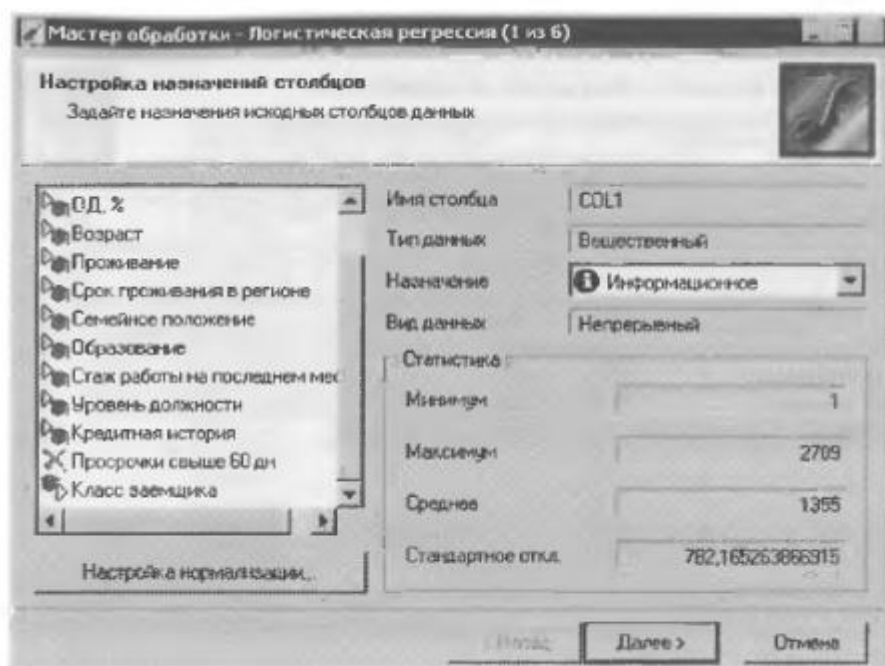


Рис. 17.2. Задание входных и выходных полей

В этом же окне нажмите кнопку **Настройка нормализации**. Для выходного поля **Класс заемщика** порядок сортировки уникальных значений (которых в логистической регрессии всегда два) определяется типом события: первое — отрицательное, второе — положительное (рис. 17.3). В скоринге принято, что чем выше рейтинг заемщика, тем выше кредитоспособность, поэтому значение «хороший» будет положительным исходом события (второе по счету), а «плохой» — отрицательным (первое по счету).

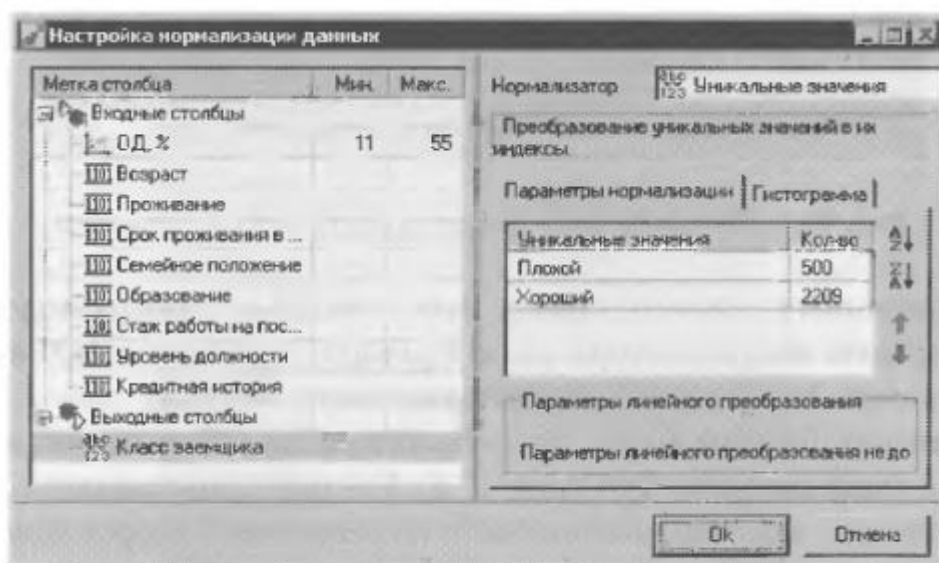


Рис. 17.3. Задание типов событий выходного поля

В следующем окне мастера будет предложено настроить обучающие и тестовые множества. Поскольку у нас есть специальное поле, в котором хранится информация о разбиении на множества, укажем его, установив соответствующие настройки (рис. 17.4).

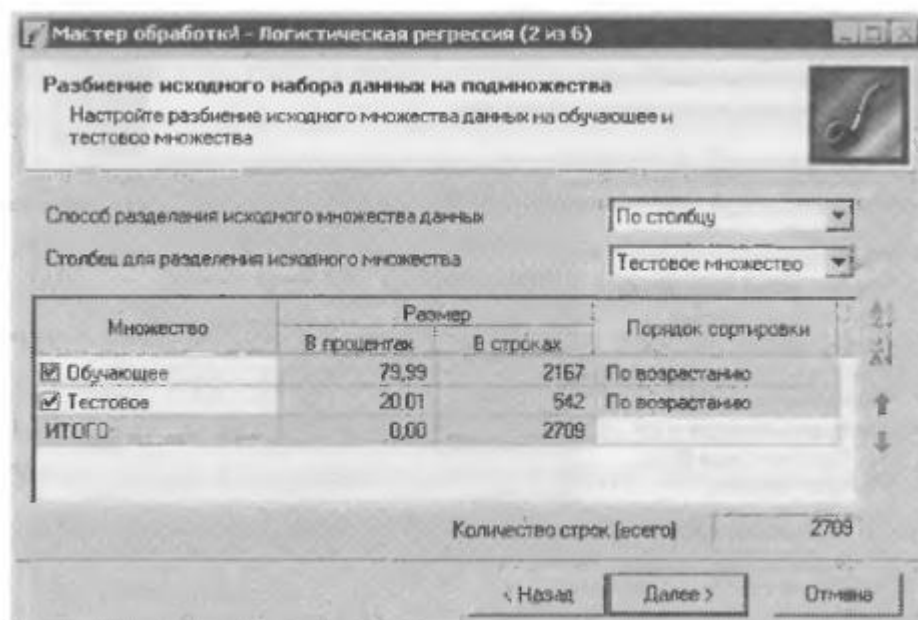


Рис. 17.4. Настройка разбиения набора данных

На третьем шаге мастера предлагается изменить параметры алгоритма логистической регрессии (рис. 17.5). По умолчанию порог классификации равен 0,5. Пока оставьте его без изменений.

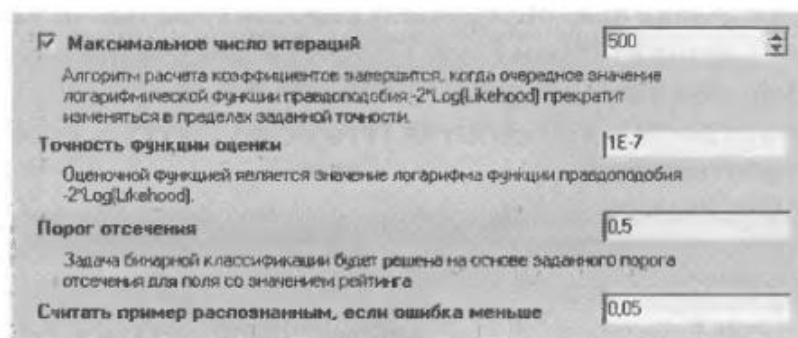


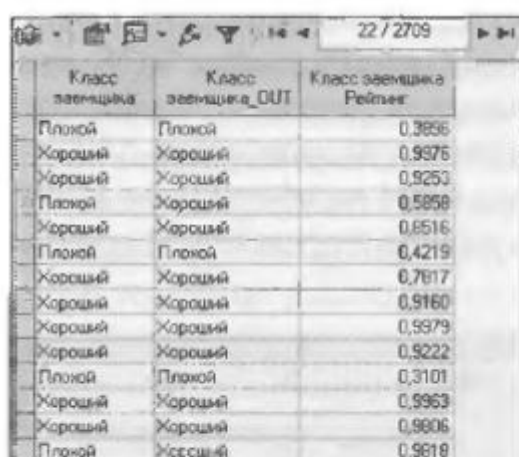
Рис. 17.5. Настройки алгоритма логистической регрессии

На последнем шаге нажмите кнопку **Пуск** — будет построена модель и мастер предложит выбрать визуализаторы узла. Укажите следующие: ROC-анализ, Коэффициенты регрессии, Что-если, Таблица сопряженности, Таблица.

В визуализаторе Таблица видно, что добавились две новые колонки: **Класс заемщика Рейтинг** и **Класс заемщика_OUT** (рис. 17.6). Рейтинг представляет собой рассчитанное значение y по уравнению логистической регрессии, а второе поле определяет принадлежность к тому или иному классу в зависимости от порога округления.

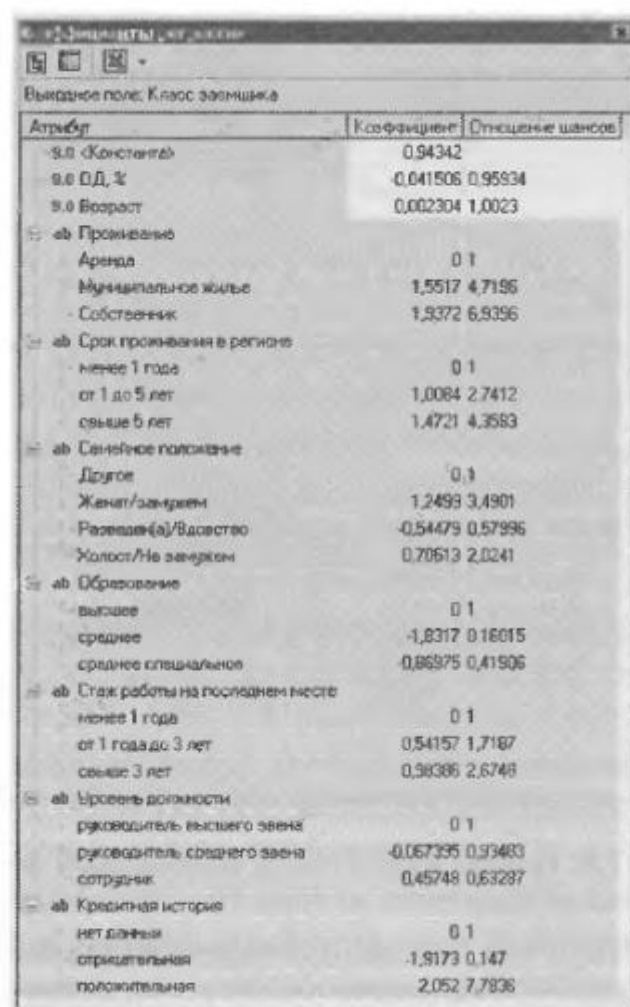
Визуализатор Коэффициенты регрессии наглядно показывает рассчитанные коэффициенты логистической регрессии, которые являются прототипом скоринговой карты, и соответствующие им отношения шансов (рис. 17.7).

Проинтерпретируем отношение шансов для признака *Стаж работы*. Если стаж работы на последнем месте от 1 до 3 лет, то шансы стать благонадежным заемщиком при фиксированных значениях других переменных в $OR = 1,71$ раза выше по сравнению с тем, у кого стаж менее 1 года. А если стаж свыше 3 лет, то шансы увеличиваются в 2,67 раза.



Класс заемщика	Класс заемщика_OUT	Класс заемщика Рейтинг
Плохой	Плохой	0,3856
Хороший	Хороший	0,9975
Хороший	Хороший	0,9253
Плохой	Хороший	0,5858
Хороший	Хороший	0,6516
Плохой	Плохой	0,4219
Хороший	Хороший	0,7617
Хороший	Хороший	0,9160
Хороший	Хороший	0,9979
Хороший	Хороший	0,9222
Плохой	Плохой	0,3101
Хороший	Хороший	0,9963
Хороший	Хороший	0,9806
Плохой	Хороший	0,9818

Рис. 17.6. Выходные поля



Выходное поле: Класс заемщика

Атрибут	Коэффициент	Отношение шансов
3.0 «Константа»	0,94342	
9.0 ОД, %	-0,041506	0,95934
9.0 Возраст	0,002304	1,0023
4.0 Проживание		
Аренда	0	1
Муниципальное жилье	1,5517	4,7196
Собственник	1,9372	6,9396
4.0 Срок проживания в регионе		
менее 1 года	0	1
от 1 до 5 лет	1,0084	2,7412
свыше 5 лет	1,4721	4,3583
4.0 Семейное положение		
Другое	0	1
Женат/замужем	1,2499	3,4901
Разведен(а)/Вдовство	-0,54479	0,57996
Холост/Не замужем	0,70613	2,0241
4.0 Образование		
высшее	0	1
среднее	-1,8317	0,16015
среднее специальное	0,86975	0,41906
4.0 Стаж работы на последнем месте		
менее 1 года	0	1
от 1 года до 3 лет	0,54157	1,7187
свыше 3 лет	0,98386	2,6748
4.0 Уровень должности		
руководитель высшего звена	0	1
руководитель среднего звена	-0,067395	0,93403
сотрудник	0,45748	0,63287
4.0 Кредитная история		
нет данных	0	1
отрицательная	-1,9173	0,147
положительная	2,052	7,7836

Рис. 17.7. Коэффициенты логистической регрессии

Проинтерпретируем теперь отношение шансов для поля *ОД, %*: $OR = 0,96$. Оно меньше единицы, значит, увеличение кредитной нагрузки снижает итоговый скоринговый балл клиента. Рассмотрим потенциального заемщика А, у которого доля выплат по кредиту в структуре дохода на 10 % меньше, чем у заемщика Б.

Тогда можно сказать, что снижение ежемесячных выплат на 10 % приводит к тому, что вероятность стать хорошим заемщиком вырастает в $\exp(10 \cdot 0,041506) = 1,52$ раза.

Визуализатор ROC-кривая выводит график ROC-кривой, на котором по умолчанию отображаются положение текущего порога отсечения, а также значения чувствительности и специфичности, показатель AUC и типы событий (рис. 17.8). Площадь под кривой равна 0,894 на обучающем множестве и 0,905 — на тестовом, что говорит об очень хорошей предсказательной способности построенной модели.

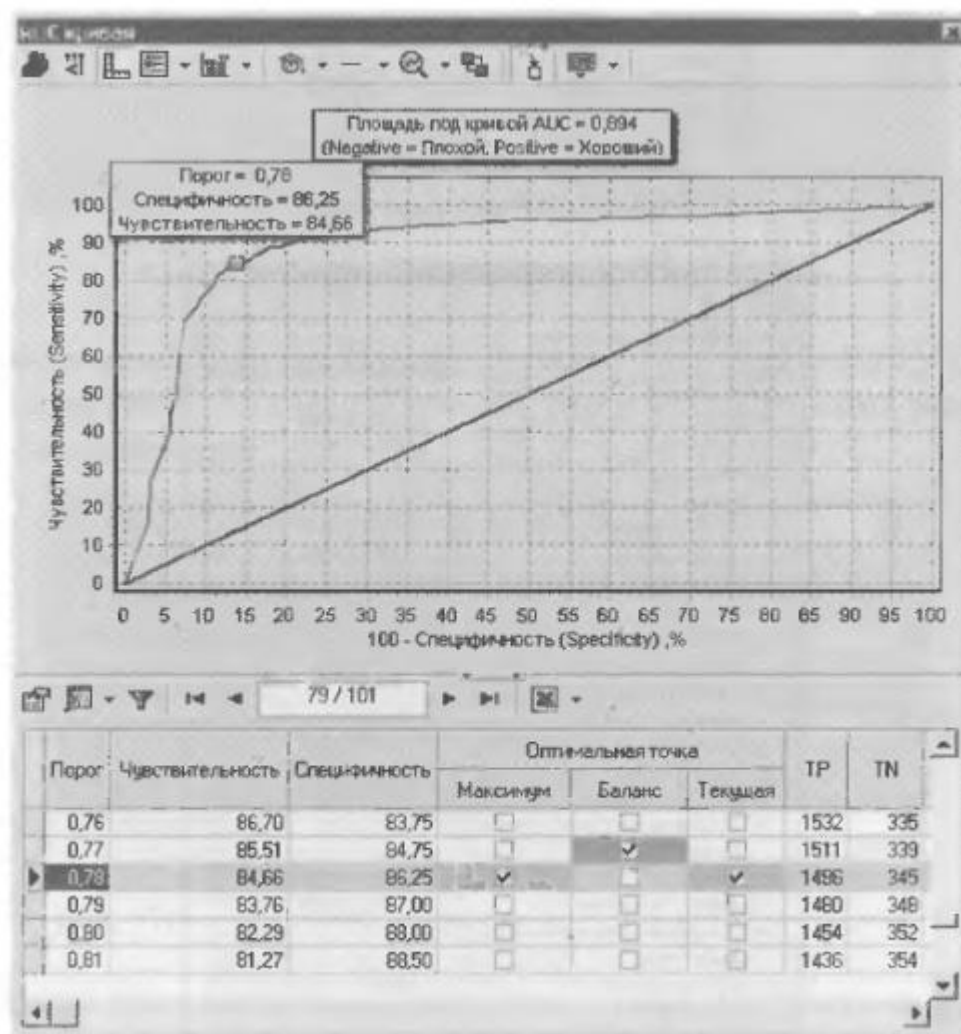


Рис. 17.8. График ROC-кривой скоринговой модели

Однако оптимальная точка для данной модели не равна 0,5. Максимальная суммарная чувствительность и специфичность достигается в точке 0,78 (для расчета и отображения оптимальной точки необходимо в меню кнопки Тип оптимальной точки выбрать пункт Максимум). Для установки нового порога отсечения, равного 0,78, следует перенастроить узел-обработчик логистической регрессии. В этой точке $Se = 85\%$, $Sr = 86\%$, что означает: 85 % благонадежных заемщиков будут выявлены классификатором, а $100 - 86 = 14\%$ недобросовестных заемщиков

получат кредит. На тестовом множестве наблюдается похожая картина: $Se = 84 \%$, $Sp = 86 \%$.

В общем случае, проецируя определения чувствительности и специфичности на скоринг (и учитывая, что класс заемщика «хороший» соответствует положительному исходу), можно заключить, что скоринговая модель с высокой специфичностью соответствует *консервативной кредитной политике* (чаще происходит отказ в выдаче кредита), а с высокой чувствительностью — *политике рискованных кредитов*. В первом случае минимизируется кредитный риск, связанный с потерями ссуды и процентов и с дополнительными расходами на возвращение кредита, а во втором — коммерческий риск, связанный с упущенной выгодой. Это хорошо иллюстрирует визуализатор Таблица сопряженности (рис. 17.9), которая есть не что иное, как матрица классификации.

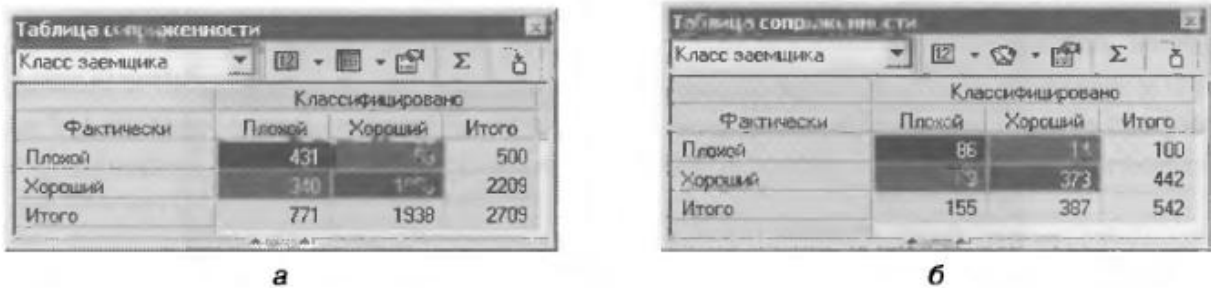


Рис. 17.9. Таблицы сопряженности: а — рабочая выборка, б — тестовая выборка

Из таблицы видно, что на обучающем множестве модель чаще отказывала в выдаче кредита «хорошим» заемщикам (271 ошибочный случай), чем выдавала кредит «плохим» (см. рис. 17.9, а). Точность классификации составила 85 %. На тестовом множестве наблюдается примерно та же картина (точность классификации 84,7 %), а уровень одобренных кредитов (Approval Rate, AR) здесь составляет $AR = (387 / 542) \cdot 100 \% = 72 \%$ при уровне дефолтных кредитов (Bad Rate, BR) равном $BR = (14 / 387) \cdot 100 \% = 3,62 \%$ (см. рис. 17.9, б). Если такая ситуация не устраивает, можно снизить порог отсеечения и добиться того, чтобы модель чаще выдавала положительное решение. Процент отказов уменьшится, но возрастет и кредитный риск. Поэтому выбор точки отсеечения зависит от поставленных целей — снизить долю «плохих» кредитов или увеличить кредитный портфель, чаще вынося положительное решение по клиенту.

Предположим, нам известны издержки ошибочной классификации (см. раздел 12.3): $C_{FN}/C_{FP} = 1/4$, то есть выдача кредита недобросовестному заемщику обходится в 4 раза дороже, чем отказ добросовестному. Тогда мы можем, используя правило Байеса, оценить оптимальный скоринговый балл P :

$$P > \frac{1}{1 + 1/4} = 0,80.$$

Визуализатор Что-если позволяет увидеть, как будет вести себя построенная модель при подаче на ее вход тех или иных данных. Иначе говоря, проводится

эксперимент, в котором, изменяя значения входных полей логистической регрессии, аналитик наблюдает за изменением значений на выходе. Возможность анализа по принципу «Что, если» особенно ценна, поскольку позволяет исследовать правильность работы системы, достоверность полученных результатов, а также ее устойчивость. Визуализатор Что-если включает табличное и графическое представления, которые формируются одновременно (рис. 17.10).

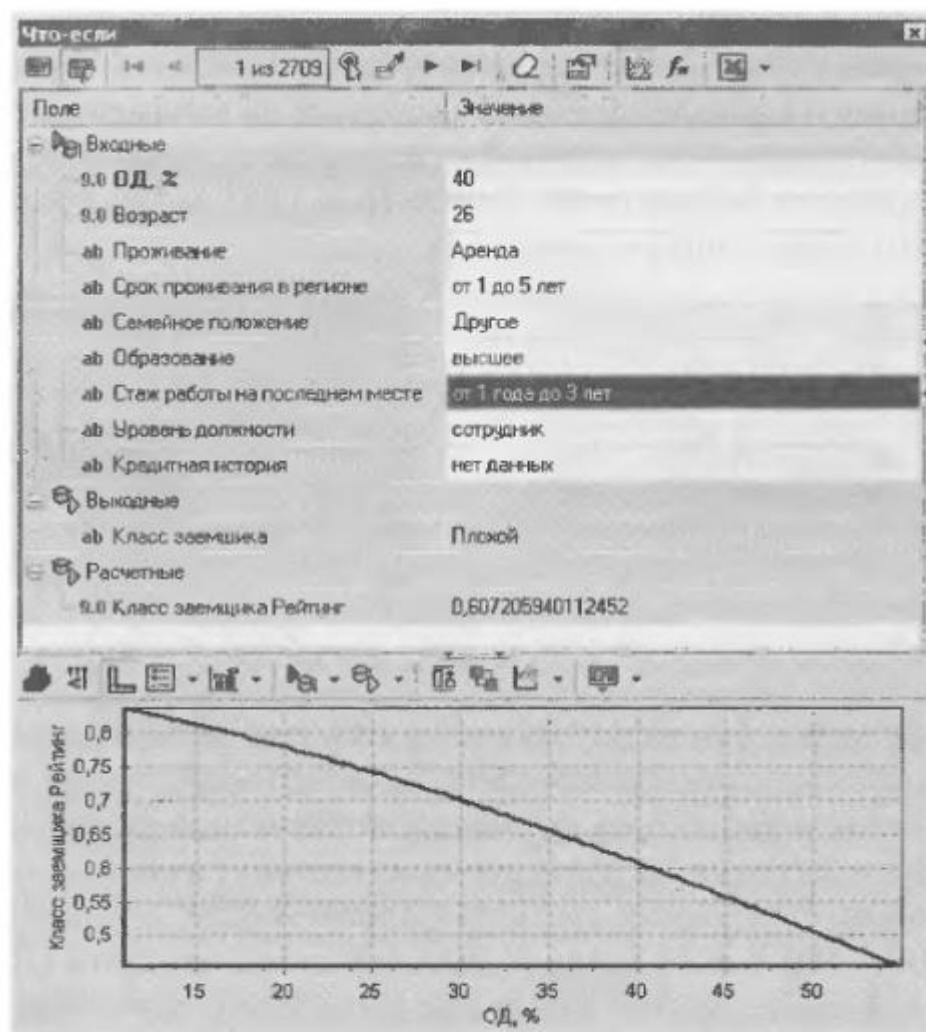


Рис. 17.10. Визуализатор «Что-если»

В верхней части табличного представления отображаются входные поля, а в нижней — выходные и расчетные. Изменяя значения входных полей, аналитик дает команду выполнить расчет и наблюдает рассчитанные значения выходов логистической регрессии.

В графическом представлении визуализатора **Что-если** по горизонтальной оси диаграммы откладывается весь диапазон значений текущего поля выборки, а по вертикальной — значения соответствующих выходов модели. На диаграмме **Что-если** видно, при каком значении входа изменяется значение на соответствующем выходе. Если, например, во всем диапазоне входных значений выходное значение для данного поля не изменялось, то диаграмма будет представлять собой горизонтальную прямую линию. В нашем случае установлена графическая зависимость изменения кредитного рейтинга конкретного клиента от коэффициента O/D (все остальные входы — константы). Видно, что с увеличением O/D рейтинг практически линейно падает.

При желании от модели логистической регрессии несложно перейти к скоринговой карте, для чего нужно перевести коэффициенты логистической регрессии в линейную шкалу.

Итак, подбирая порог отсеечения, мы можем установить желаемое соотношение уровня одобрений AR и ожидаемой величины просроченной задолженности BR .

Интересует вопрос, как сравнить несколько скоринговых карт между собой. Для этого строят различные отчеты и графики. Например, подвергают анализу кривые распределения кумулятивных процентов для хороших и плохих кредитов. По оси ox откладывают диапазоны скорингового балла, а по oy — накапливающуюся долю плохих (хороших) кредитов.

Сформируем такой отчет в Deductor. Нам понадобятся несколько узлов из группы *Трансформация данных*: Квантование, Группировка, Кросс-таблица, Замена данных и Калькулятор (преобразования, выполняемые ими, описывались в главе 3)¹, а также визуализатор **Диаграмма**. На рис. 17.11 можно наблюдать график, отражающий кумулятивный процент хороших и плохих кредитов нашей скоринговой карты.

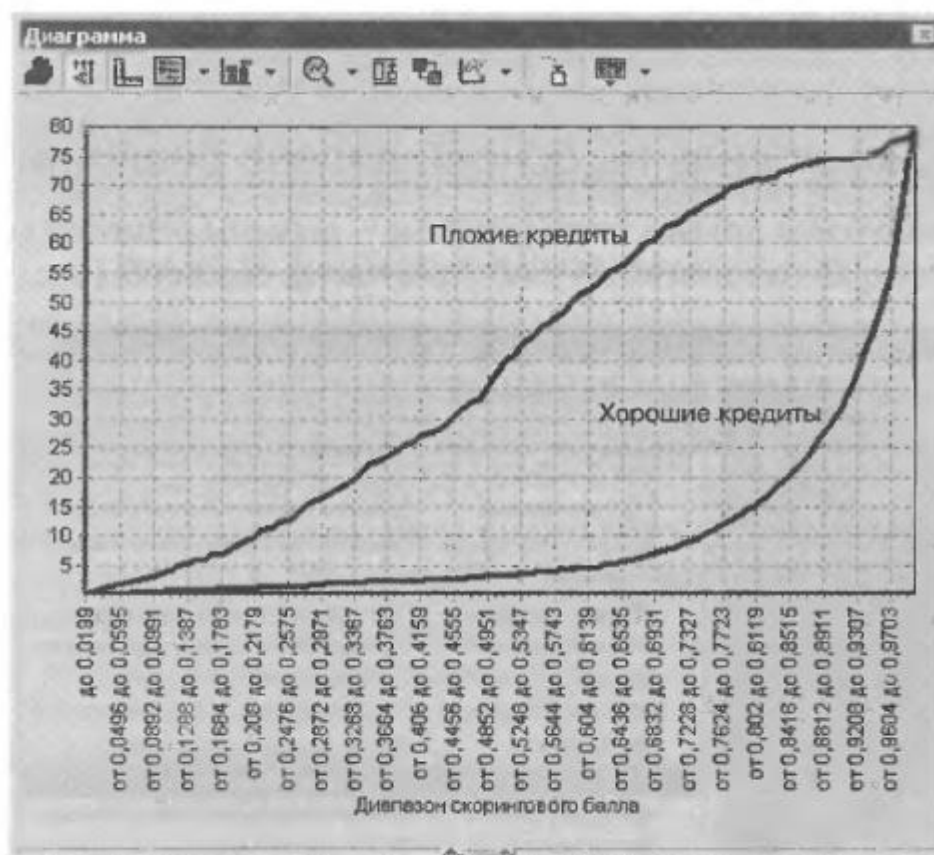


Рис. 17.11. Распределения кумулятивных процентов в скоринговой модели

Видно, что кривая для плохих кредитов проходит выше и обладает более крутым подъемом, чем кривая для хороших, поскольку в области низких значений сосредоточено больше плохих кредитов. Если скоринговая карта неэффективна, эти кривые будут похожи, а в пределе — налагаться друг на друга.

Окончательный сценарий будет иметь следующий вид (рис. 17.12).

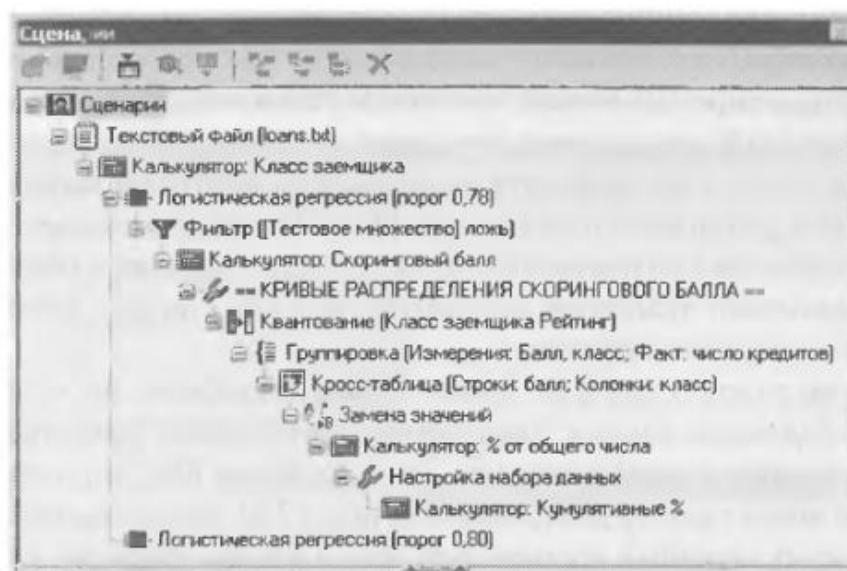


Рис. 17.12. Сценарий построения скоринговой модели на основе логистической регрессии

Скоринговая модель на основе дерева решений

Теперь воспользуемся другим инструментом — деревом решений

Добавьте в сценарий одноименный узел через Мастер обработки

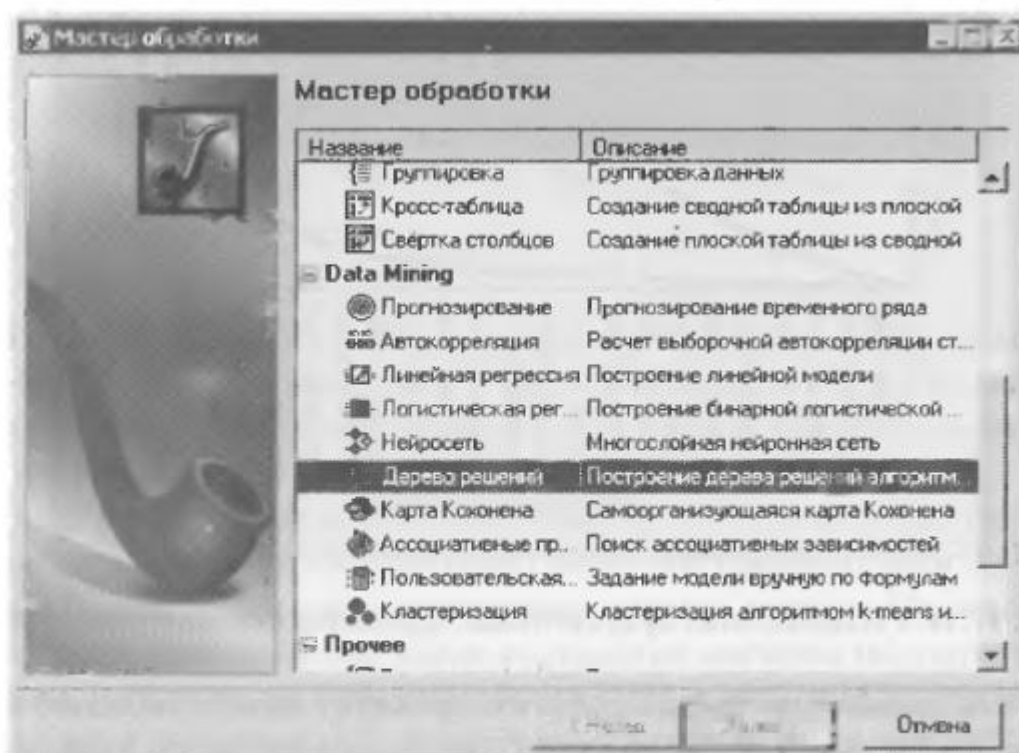


Рис. 17.13. Окно выбора нового узла обработки

Следующие два шага мастера аналогичны описанным ранее для узла Логистическая регрессия. На четвертом шаге откроется окно выбора параметров алгоритма C4.5 (рис. 17.14). Здесь не меняйте настройки, принятые по умолчанию, за исключением минимального количества примеров в узле, при котором будет создаваться новый. Задайте этот параметр равным примерно 1 % от объема обучающего множества; меньшее значение может привести к появлению недостоверных правил, большее — к почти полному отсутствию таковых.

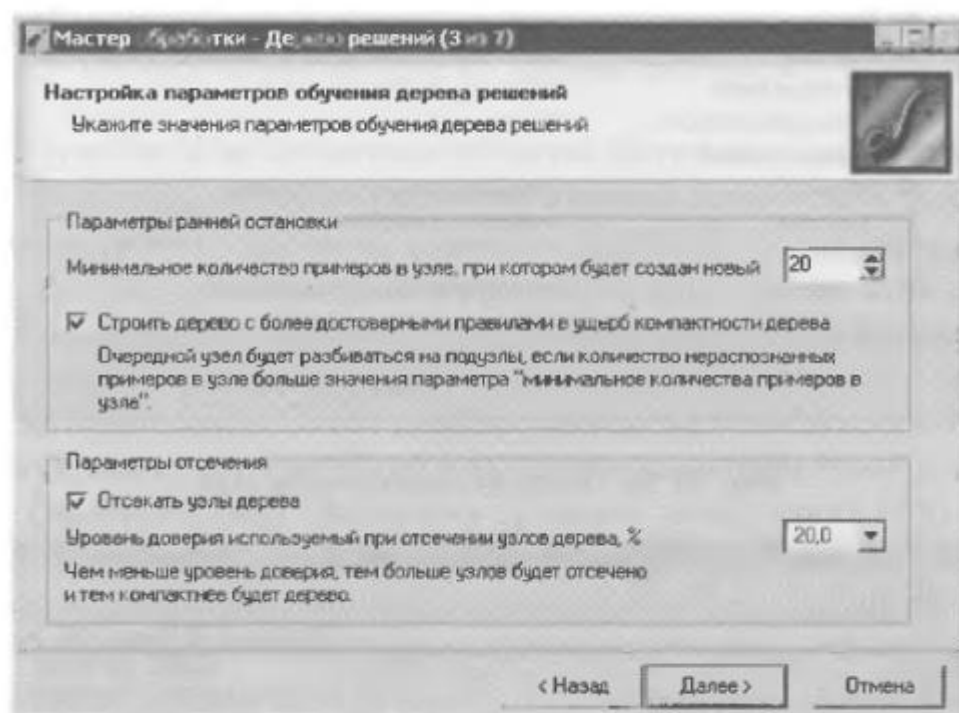


Рис. 17.14. Настройка алгоритма дерева решений

На следующем шаге в качестве желаемого способа построения дерева оставьте режим автоматического построения — это и есть алгоритм C4.5. Запустив его нажатием кнопки Пуск, пройдите по шагам мастера дальше и выберите нужные визуализаторы, как показано на рис. 17.15.

В результате работы алгоритма было выявлено 18 правил; точность классификации на обучающем множестве составила 85 %, на тестовом — 87 %. Визуализатор **Дерево решений** позволяет увидеть полученный набор правил в схематическом виде, а также выводит показатели достоверности и поддержки для каждого узла (рис. 17.16). Это и есть скоринговая модель. Она менее привычна, поскольку здесь не начисляются баллы за характеристики заемщика, но тоже объясняет результат классификации того или иного заемщика.

В принципе, достоверность каждого правила можно воспринимать как итоговый скоринговый балл с той оговоркой, что для плохих заемщиков он равен величине, полученной вычитанием из 100%-ного значения достоверности.

Теперь откройте таблицы сопряженности этого дерева решений (рис. 17.17).

Оказывается, в сравнении с моделью на основе логистической регрессии здесь совершенно другая ситуация. Дерево решений значительно чаще одобряет неблагонадежных заемщиков, потому что его построение идет в условиях несбалансированности классов. В результате доля дефолтных кредитов на тестовом множестве равна $BR = 51 / 475 \cdot 100 \% = 10,7 \%$, что в 3 раза выше этого же показателя в логистической модели (правда, уровень одобрений вырастает до 87,6 %). Что делать, если такая ситуация не устраивает? В логистической регрессии для решения этой проблемы мы варьировали порогом отсечения, а в дереве решений такой возможности нет.

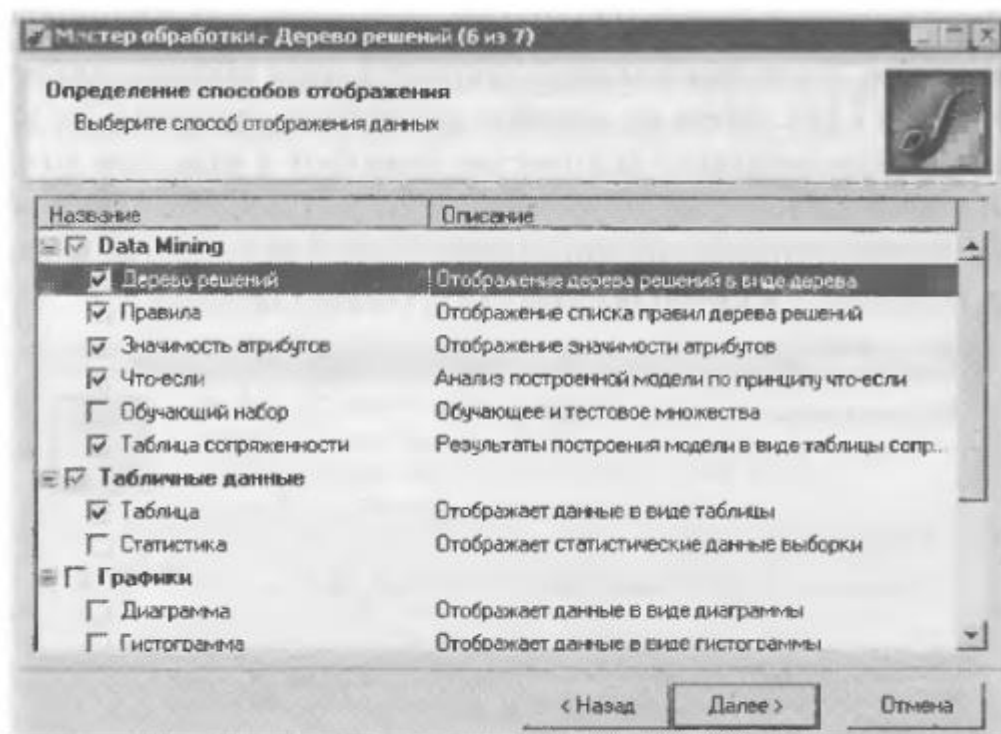


Рис. 17.15. Выбор визуализаторов узла

Нам помогут специальные стратегии сэмплинга для уравнивания обучающей выборки с дублированием

миноритарного класса (*oversampling*) и выборка с удалением примеров мажоритарного класса (*undersampling*). Поскольку примеров не так много (400 — с плохими клиентами и 1767 — с хорошими) и информация о каждом заемщике представляет ценность, имеет смысл использовать первый вариант — с дублированием. Пусть отношение издержек ошибочной классификации останется прежним: 1 : 4. Тогда, согласно правилу, к обучающей выборке нужно добавить $3 \cdot 400 = 1200$ примеров, и общее число записей составит 3367, а доля плохих увеличится до 47 %.

ЗАМЕЧАНИЕ

Процедуру дублирования записей, принадлежащих к миноритарному классу, нужно осуществлять только на обучающем множестве.

Для этой операции снова привлечем несколько узлов из группы *Трансформация данных*: *Фильтр* и *Слияние данных* (рис. 17.18, выделенный блок).

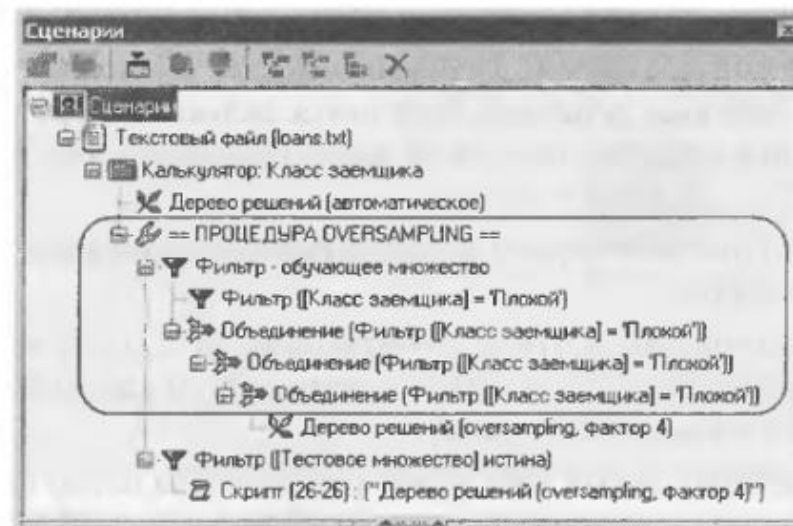


Рис. 17.18. Сценарий построения скоринговой модели на основе дерева решений

Фактически	Классифицировано		Итого
	Плохой	Хороший	
Плохой	78	22	100
Хороший	68	354	442
Итого	166	376	542

Рис. 17.19. Таблица сопряженности для тестовой выборки (процедура oversampling)

Построив дерево решений на сбалансированной выборке, убедитесь, что ситуация улучшилась: теперь на тестовом множестве модель чаще отказывает в выдаче хорошим заемщикам, нежели одобряет плохих (рис. 17.19). Эти результаты сравнимы с теми, которые выдает модель логистической регрессии.

Таким образом, мы получили несколько скоринговых моделей. Варьируя порогами отсека и применяя специальные

приемы борьбы с несбалансированностью классов, можно подобрать ту модель, которая отвечает заданным потребностям кредитного учреждения по уровню одобрений заявок и ожидаемой доле просроченной задолженности. Проверять новых клиентов можно при помощи обработчика Скрипт.

Интерактивное дерево решений

До этого мы получали дерево, которое строилось автоматическим способом, то есть алгоритм на каждом шаге выбирал атрибут для разбиения по заданному критерию. В главе 9 также говорилось о том, что алгоритмы построения деревьев «жадные», поэтому не факт, что итоговое дерево будет наилучшим. В то же время иногда имеются экспертные знания, которые позволяют «вмешаться» в процесс формирования дерева и выбора атрибутов, а также порогов для разбиения. Возможно, это и не повысит точность модели, но правила станут более логичными, с точки зрения экспертов.

Кредитный скоринг представляет собой тот самый случай, когда банковские аналитики имеют определенные знания и хотят, чтобы в модели ветвление по атрибутам осуществлялось в определенном порядке. Например, если имеются атрибуты *Наличие квартиры* и *Стоимость квартиры*, то разумно сразу после первого рассмотреть второй. Еще пример: после суммы кредита сразу желательно проанализировать первоначальный взнос.

В аналитической платформе Deductor имеется возможность построения интерактивных деревьев решений. Зададимся целью построить скоринговую модель на прежней выборке, приняв во внимание следующие пожелания экспертов.

- ☐ Первым атрибутом, по которому анализируют заемщика, должен быть атрибут *Кредитная история*.
- ☐ Далее необходимо рассмотреть коэффициент О/Д. Всех клиентов нужно разбить на три категории: заемщики с низким О/Д (до 20 %), с умеренным (от 20 до 40 %) и высоким (от 40 %).

Добавьте в сценарий новый узел дерева решений и на пятом шаге мастера поставьте переключатель в позицию **Интерактивный режим** (рис. 17.20).

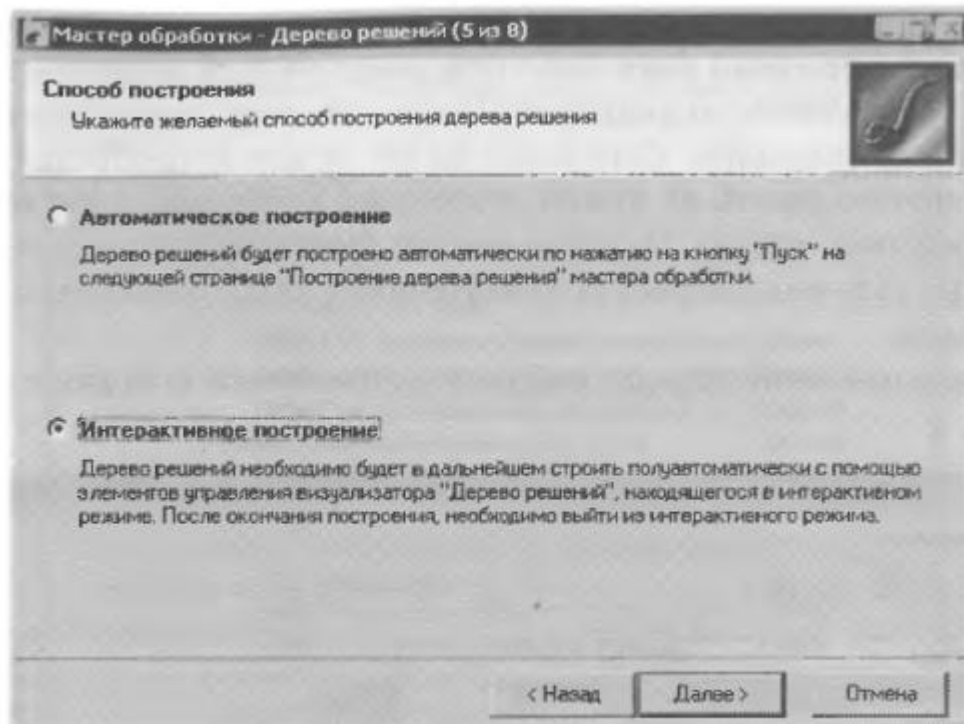


Рис. 17.20. Выбор способа построения дерева решений

В результате открывшийся визуализатор **Дерево решений** не будет содержать ни одного узла. На панели инструментов нажмите кнопку **Разбить текущий узел на подузлы...**, откроется соответствующее окно (рис. 17.21).

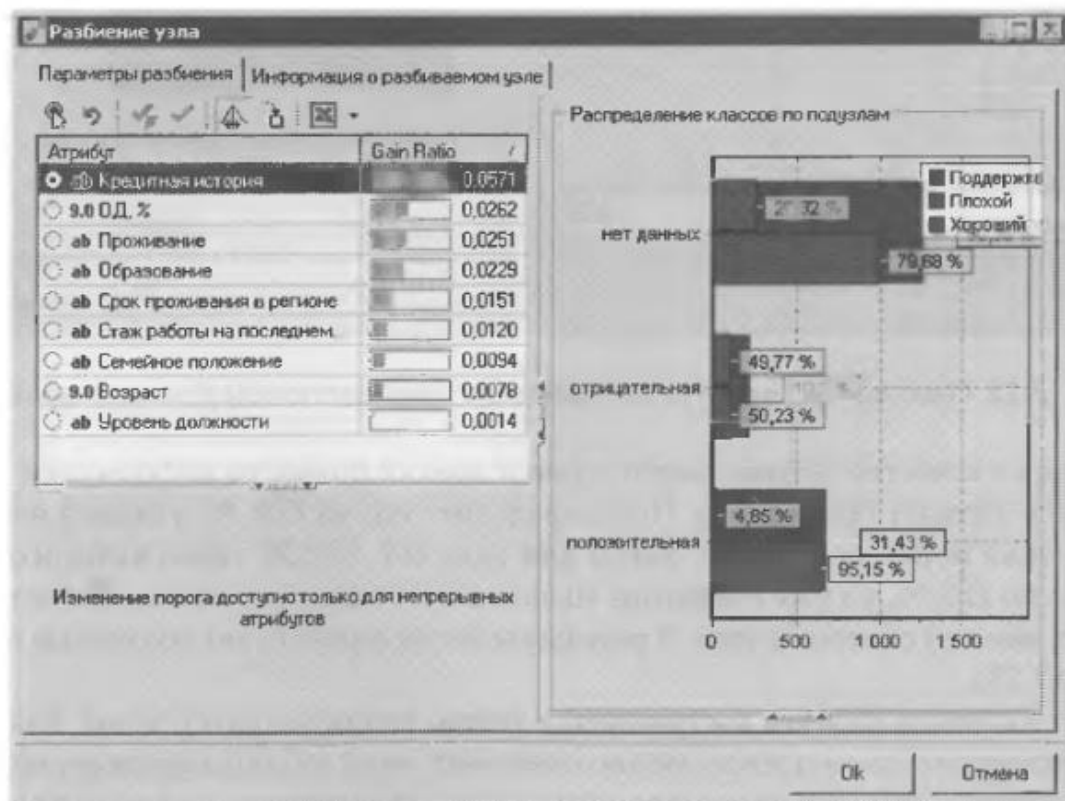


Рис. 17.21. Окно выбора атрибута для разбиения в интерактивном режиме: первый шаг

Слева в списке выводятся все атрибуты вместе с рассчитанными значениями прироста информации *Gain Ratio*, а справа — диаграммы распределения классов по подузлам. По умолчанию предлагается атрибут с максимальным значением *Gain Ratio*, но его можно переопределить. В данном случае ничего делать не нужно, поскольку разбиение и так начнется по атрибуту Кредитная история. Нажатие кнопки Ок приведет к тому, что в дерево добавится три узла этого атрибута со значениями *нет данных*, *отрицательная*, *положительная*.

Продолжим разбиение дальше, выбрав узел *Кредитная история = нет данных* (рис. 17.22).

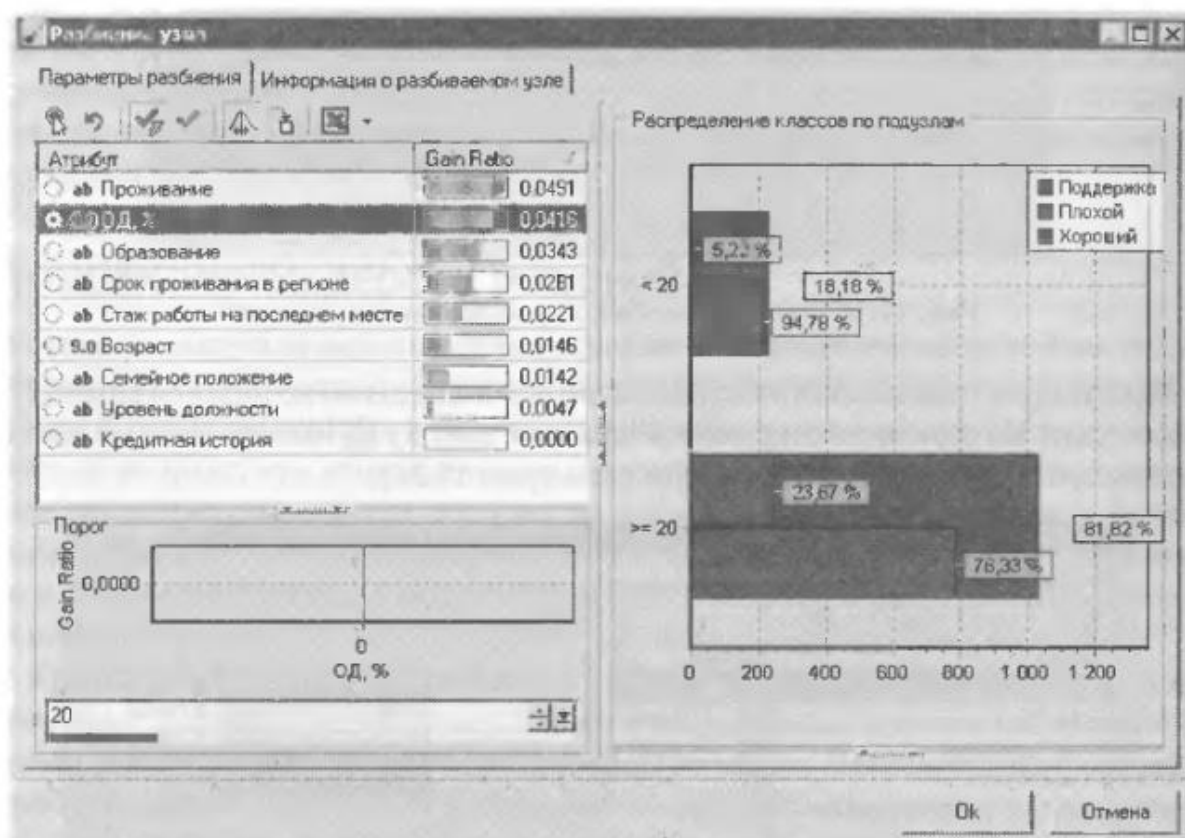


Рис. 17.22. Окно выбора атрибута для разбиения в интерактивном режиме: второй шаг

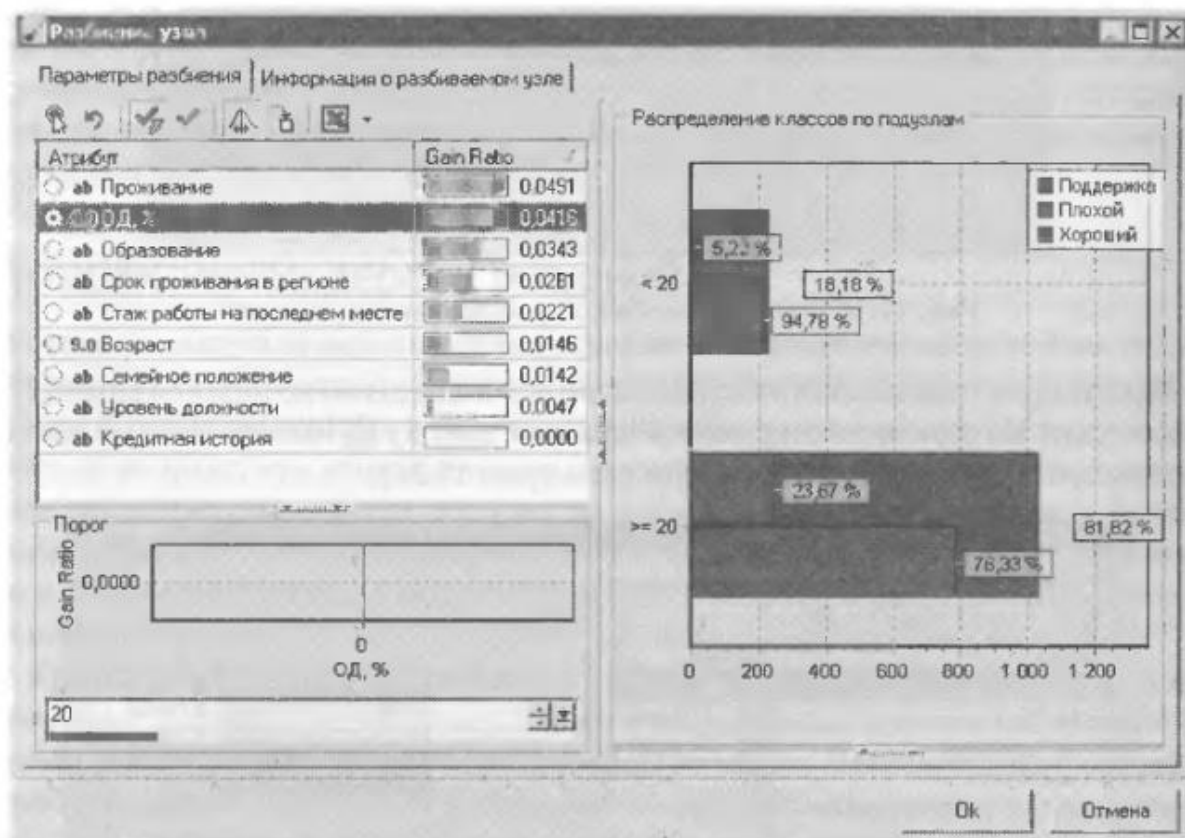


Рис. 17.22. Окно выбора атрибута для разбиения в интерактивном режиме: второй шаг

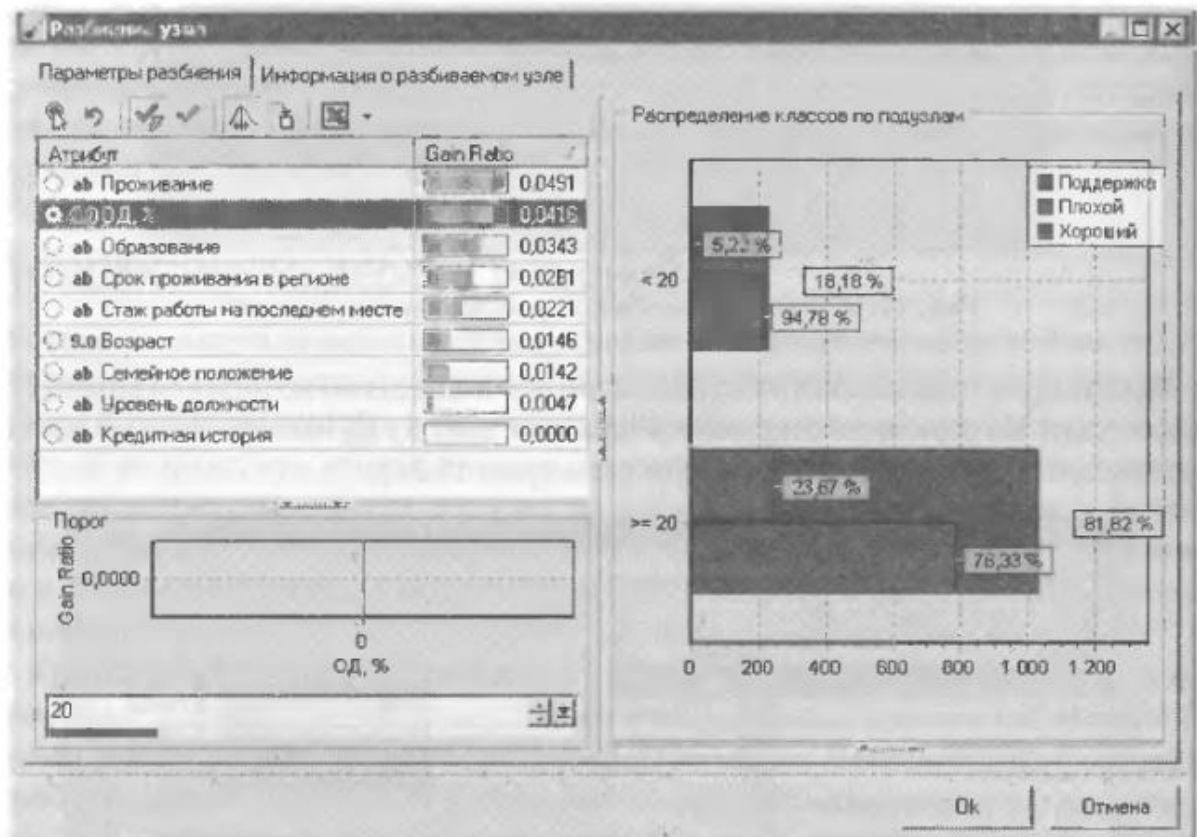


Рис. 17.22. Окно выбора атрибута для разбиения в интерактивном режиме: второй шаг

Здесь в качестве оптимального с точки зрения прироста информации предлагается атрибут **Проживание**. Переопределите его на **ОД, %**, указав в нижней части окна порог, равный 20. Затем для узла **ОД, % ≥ 20** снова выберите разбиение по **ОД, %**, но уже с порогом 40, после чего нажмите кнопку **Построить дерево**, начиная с текущего узла. В результате ветвь дерева будет полностью готова (рис. 17.23).

Аналогичным образом достраивается дерево для оставшихся узлов. Качество классификации, как и прежде, можно оценивать через таблицы сопряженности.

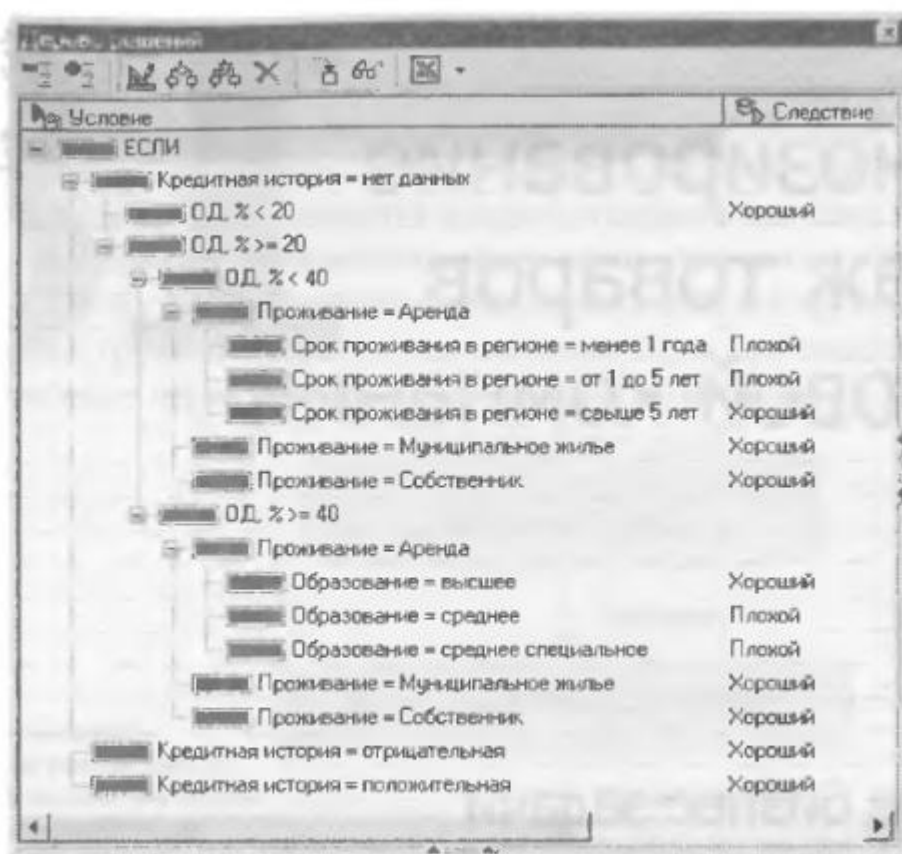


Рис. 17.23. Дерево решений, построенное в интерактивном режиме

ЗАДАНИЕ.

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