

Внешний анализ: сегментация клиентской базы

Деревья решений

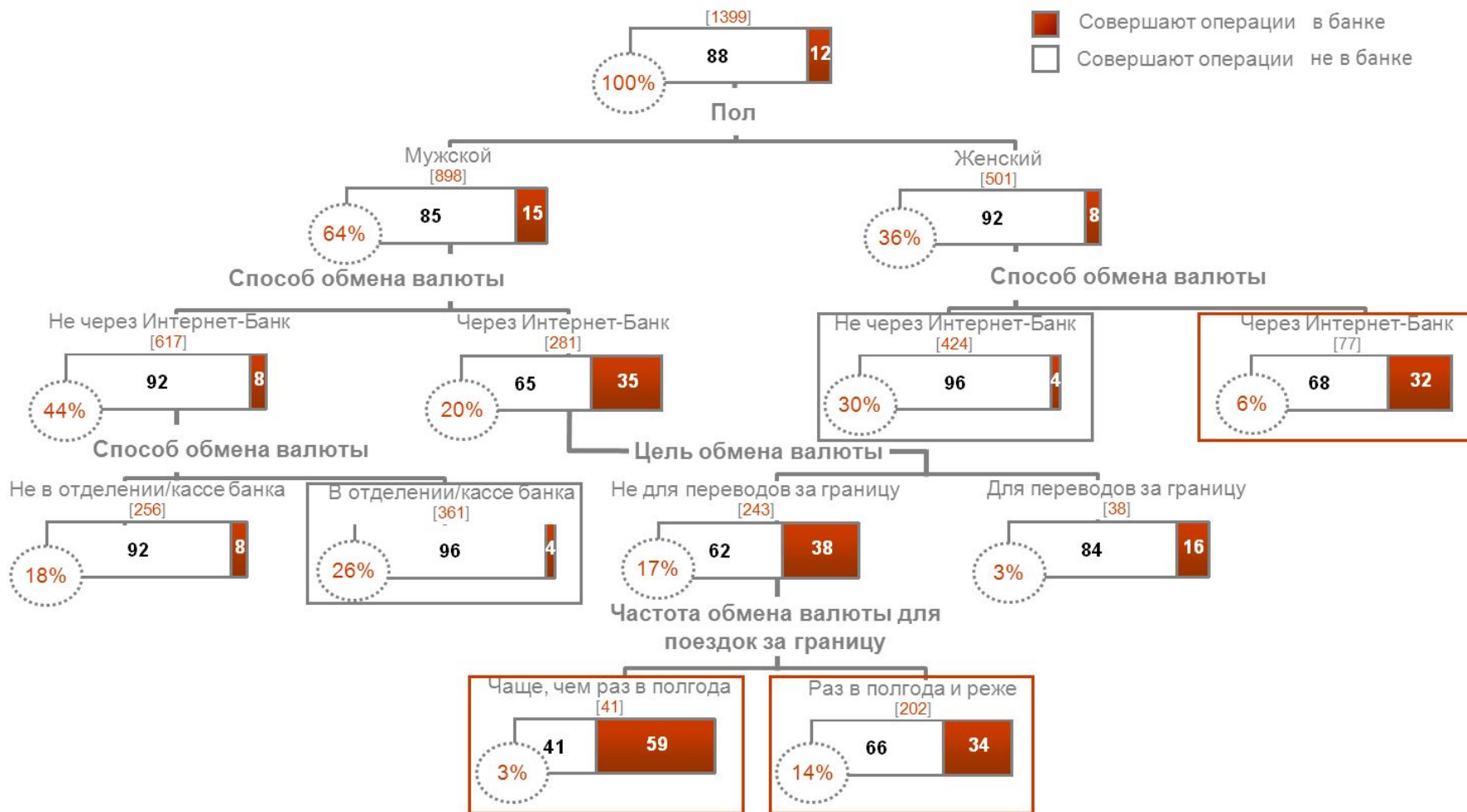
План

- Понятие дерева решений. Применение деревьев решений в задаче выявления рыночных сегментов.
- Алгоритмы построения дерева решений.
- Реализация сегментации на основе деревьев решений в SPSS, Deductor и др. программах.

Дерево решений для сегментации заемщиков банка



Дерево решения для сегментации обменивающих



Совершают операции в банке
 Совершают операции не в банке

5% Доля целевой группы во всей выборке [1041] Численность целевой группы
 Группы, в которых процент обменивающих валюту в банке максимальный
 Группы, в которых процент обменивающих валюту НЕ в банке максимальный

Понятие дерева решений

- **Дерево решений (классификации)** – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение.
- **Дерево классификации** – набор последовательно выделенных сегментов с наибольшими различиями целевой переменной (например, группы с максимальным и минимальным процентом заинтересованных в услуге).
- Это позволяет найти, сочетание каких признаков сильнее всего влияет на целевую переменную, а также определить наиболее перспективные целевые группы.

Достоинства деревьев решений

- быстрый процесс обучения
- генерация правил в областях, где эксперту трудно формализовать свои знания
- извлечение правил на естественном языке
- интуитивно понятная классификационная модель
- высокая точность прогноза
- построение непараметрических

Основные этапы алгоритмов конструирования деревьев

- **построение дерева (tree building)**
 - выбор атрибута для разбиения дерева
 - выбранный атрибут должен разбить множество так, чтобы получаемые в итоге подмножества состояли из объектов, принадлежащих к одному классу, или были максимально приближены к этому, т.е. количество объектов из других классов ("примесей") в каждом из этих множеств было как можно меньше
 - остановка
- **сокращение дерева (tree pruning)**
 - на основе анализа ошибок классификации

Алгоритмы построения деревьев решений

- CHAID, ECHAID (Exhaustive CHAID)
 - для получения оптимального разбиения используется критерий связи между категориальными переменными хи-квадрат (в случае, если целевая переменная является количественной, используется F-критерий). Исходно целевая переменная и переменные-предикторы могут быть как количественными, так и категориальными, однако количественные предикторы при построении дерева преобразуются в категориальные.
- ID3
- C.4.5
- CART (Classification And Regression Tree)
 - основан не на статистических критериях, а на уменьшении неоднородности сегментов (узлов) (индекс Gini). Хорошо работает в том случае, если все переменные в анализе являются количественными. В методе могут быть использованы как количественные, так и категориальные целевая переменная и переменные предикторы
- QUEST
 - В данном методе для выбора предикторов . применяются различные критерии, в зависимости от типа потенциального предиктора. Он позволяет избегать смещений, связанных с выбором предикторов с большим количеством категорий, но целевая переменная в данном случае должна быть категориальной. Предикторы могут быть как количественными, так и категориальными.

CHAID-анализ: основные идеи

- Метод основан на критерии хи-квадрат.
- На входе анализа – категориальная зависимая переменная (например, заинтересованность/незаинтересованность в услуге) и несколько независимых переменных (предикторов).
- Вначале ищется самый сильный фактор, который наилучшим образом объясняет различия между категориями зависимой переменной. Автоматически перебираются все предикторы, ищутся все комбинации значений и находится наилучшее решение, т. е. то, которое максимизирует различия (при котором наибольший хи-квадрат).
- Далее в каждой из полученных групп процесс повторяется заново: вновь перебираются все предикторы и находится оптимальное решение для второго уровня. То же – для следующих уровней. **В каждой из подгрупп процесс происходит независимо**, т.е. например, первым фактором оказался пол, а далее для женщин важен возраст, а для мужчин, скажем, семейное положение.

Пример: дерево решений в SPSS

- Целевая переменная
 - credit rating (кредитный рейтинг)
- Предикторы
 - Age (возраст)
 - Income level (уровень дохода)
 - Number of credit cards (количество кредиток)
 - Education (образование)
 - Car loans (количество автокредитов)

Шаг 1 – открытие данных

tree_credit.sav [Набор данных1] - Редактор данных IBM SPSS Statistics

Файл Правка Вид Данные Преобразование Анализ Прямой маркетинг Графика Сервис Окно Справка

Показано 6 переменных из 6

	Credit_rating	Age	Income	Credit_cards	Education	Car_loans	пер	пер	пер	пер	пер	пер	пер	пер	пер	пер	пер	пер	пер
1	,00	36,22	2,00	2,00	2,00	2,00													
2	,00	21,99	2,00	2,00	2,00	2,00													
3	,00	29,17	1,00	2,00	1,00	2,00													
4	,00	32,75	1,00	2,00	2,00	1,00													
5	,00	36,77	2,00	2,00	2,00	2,00													
6	,00	39,32	2,00	2,00	2,00	2,00													
7	,00	31,70	2,00	2,00	2,00	2,00													
8	,00	34,72	1,00	2,00	1,00	2,00													
9	,00	31,53	1,00	2,00	1,00	2,00													
10	,00	24,78	2,00	2,00	2,00	2,00													
11	,00	22,76	1,00	2,00	2,00	2,00													
12	,00	45,97	1,00	2,00	1,00	2,00													
13	,00	29,39	2,00	2,00	1,00	2,00													
14	,00	29,21	1,00	2,00	2,00	1,00													
15	,00	39,60	1,00	2,00	1,00	2,00													
16	,00	39,46	1,00	2,00	2,00	2,00													
17	,00	34,13	1,00	2,00	2,00	2,00													
18	,00	35,82	2,00	2,00	2,00	2,00													
19	,00	35,97	2,00	2,00	2,00	2,00													
20	,00	26,26	3,00	2,00	2,00	2,00													
21	,00	21,52	1,00	2,00	2,00	2,00													
22	,00	29,23	2,00	2,00	1,00	2,00													
23	,00	22,94	1,00	2,00	2,00	2,00													
24	,00	43,42	2,00	2,00	2,00	2,00													
25	,00	20,16	2,00	2,00	1,00	2,00													
26	,00	27,98	2,00	2,00	1,00	2,00													
27	,00	29,49	2,00	2,00	2,00	2,00													
28	,00	30,12	2,00	2,00	1,00	2,00													
29	,00	25,43	2,00	2,00	2,00	2,00													
30	,00	40,19	1,00	2,00	2,00	1,00													
31	,00	20,52	1,00	2,00	2,00	2,00													
32	,00	22,76	1,00	2,00	1,00	2,00													
33	,00	32,90	2,00	2,00	1,00	2,00													
34	,00	33,21	2,00	2,00	1,00	2,00													
35	,00	27,98	3,00	2,00	1,00	1,00													
36	,00	20,99	2,00	2,00	2,00	2,00													
37	,00	28,61	2,00	2,00	2,00	2,00													

Представление Данные Представление Переменные

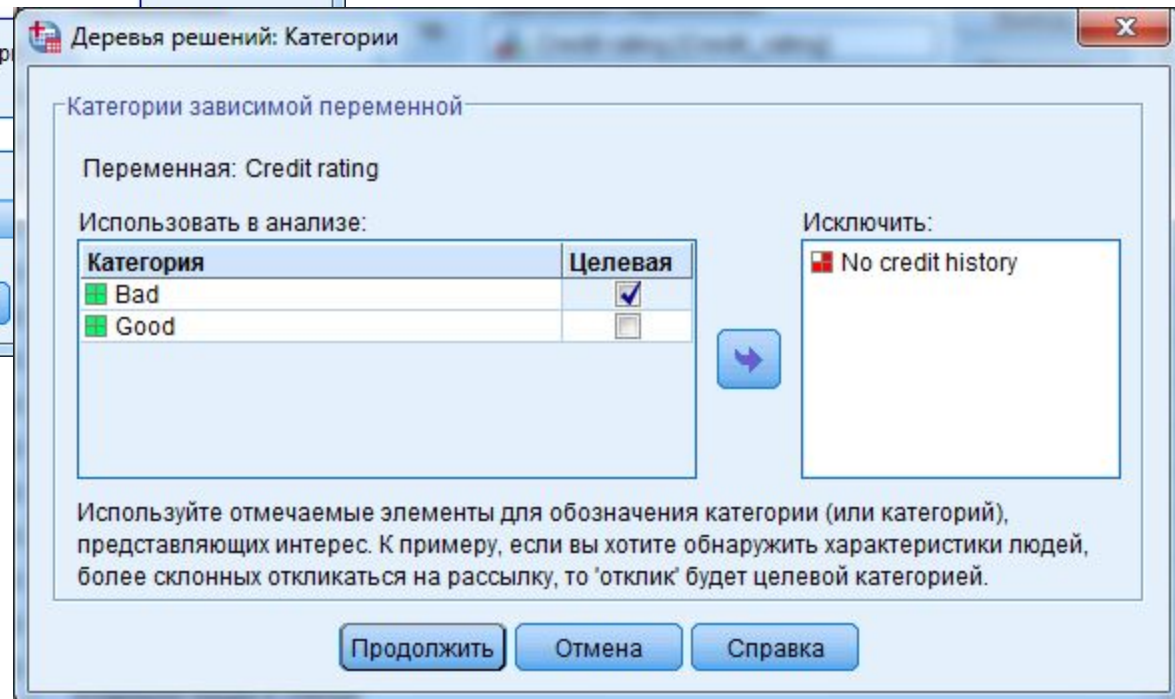
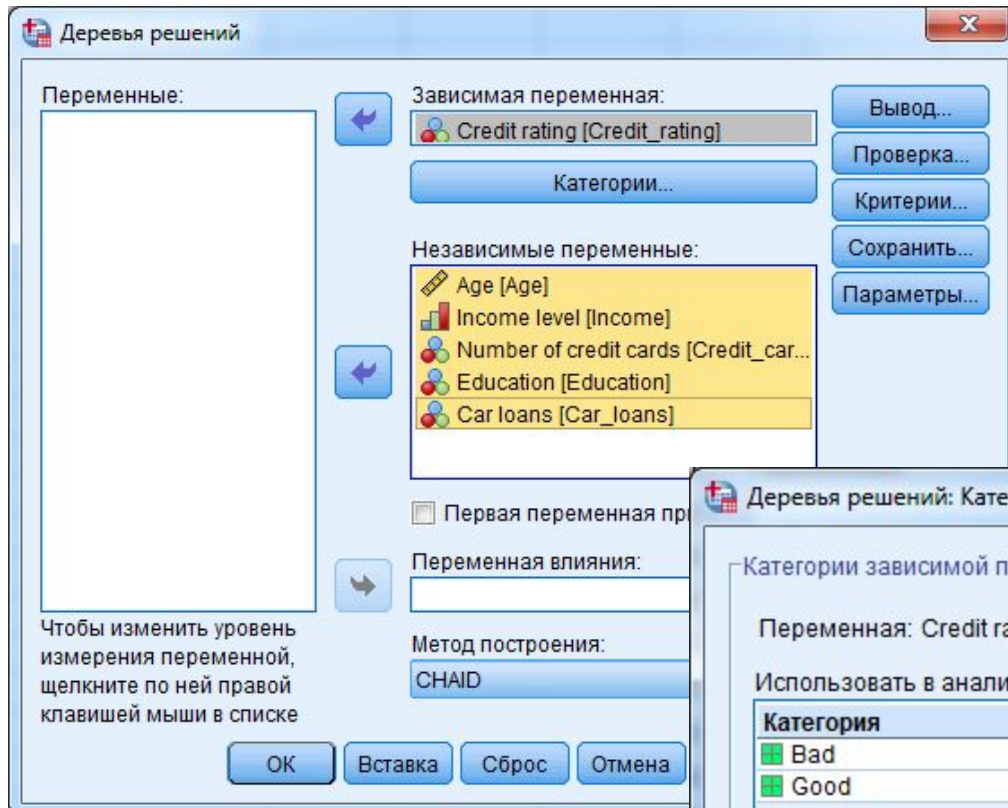
Процессор IBM SPSS Statistics rotos Unicode:ON

Шаг 2 – выбор метода

The screenshot displays the IBM SPSS Statistics interface. The 'Анализ' (Analyze) menu is open, and the 'Классификация' (Classification) option is selected. The submenu shows several classification methods, with 'Дереья классификации...' (Classification Trees...) highlighted. The background data editor shows a table with columns for 'Credit_rating', 'Age', and 'Income'.

	Credit_rating	Age	Income
1	.00	36,22	
2	.00	21,99	
3	.00	29,17	
4	.00	32,75	
5	.00	36,77	
6	.00	39,32	
7	.00	31,70	
8	.00	34,72	
9	.00	31,53	
10	.00	24,78	
11	.00	22,76	
12	.00	45,97	
13	.00	29,39	
14	.00	29,21	
15	.00	39,60	
16	.00	39,46	
17	.00	34,13	
18	.00	35,82	
19	.00	35,97	
20	.00	26,26	
21	.00	21,52	
22	.00	29,23	
23	.00	22,94	1,00
24	.00	43,42	2,00
25	.00	20,16	2,00
26	.00	27,98	2,00
27	.00	29,49	2,00
28	.00	30,12	2,00
29	.00	25,43	2,00
30	.00	40,19	1,00
31	.00	20,52	1,00
32	.00	22,76	1,00
33	.00	32,90	2,00
34	.00	33,21	2,00
35	.00	27,98	3,00
36	.00	20,99	2,00
37	.00	28,61	2,00

Шаг 3 – задание переменных



Шаг 4 - дополнительные настройки

Дерева решений: Критерии

Ограничения на размер дерева CHAID Интервалы

Максимальное количество уровней

- Автоматически
Максимальное количество уровней составляет 3 для CHAID; 5 для CRT
- Настраиваемая:
Значение:

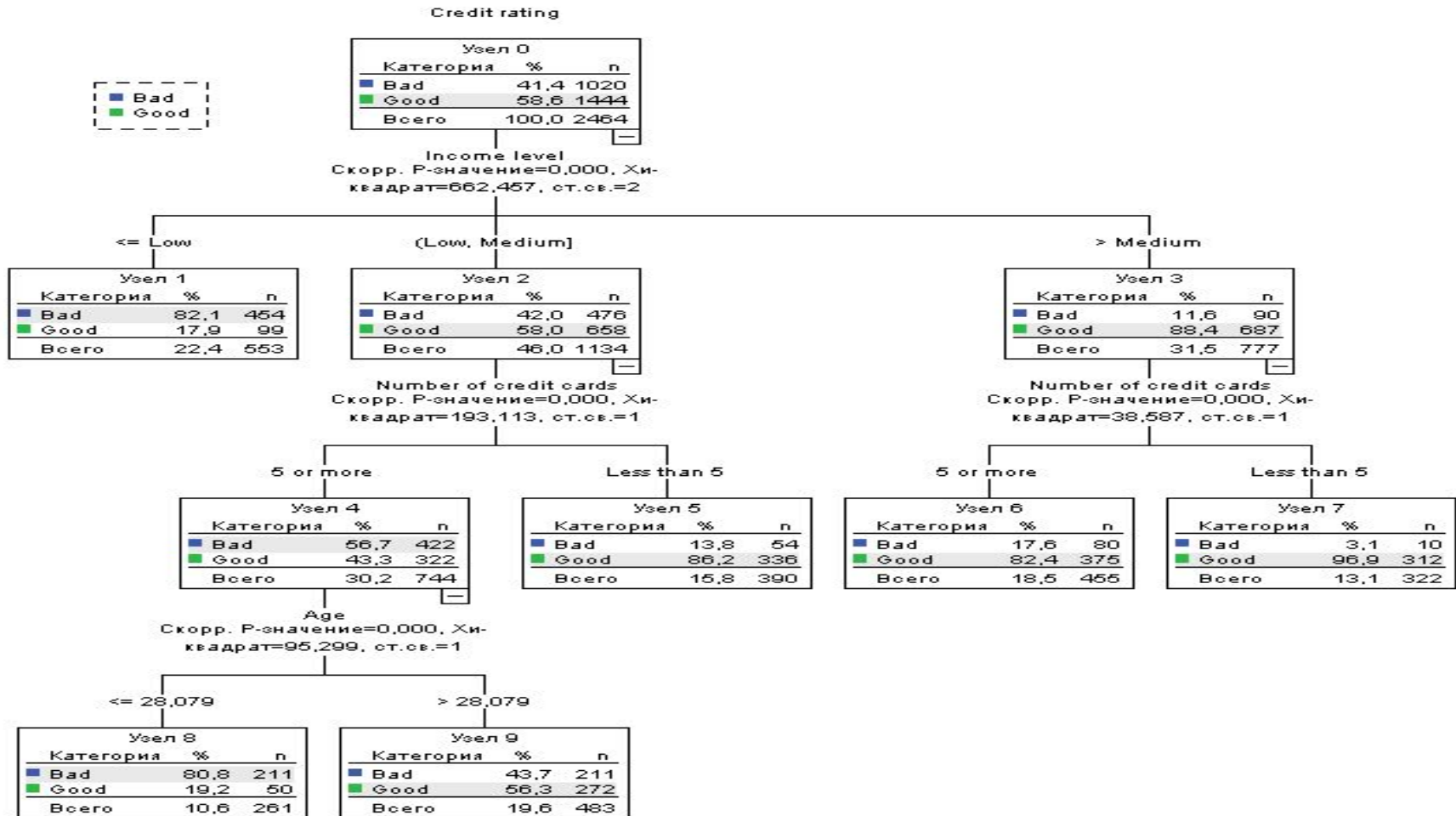
Минимум наблюдений в узле

Узел-отец:

Дочерний узел:

Продолжить Отмена Справка

Шаг 5 – анализ дерева



Шаг 5 – анализ дерева (продолжение)

Выигрыши для узлов

Узел	Узел		Выигрыш		Отклик	Индекс
	N	Проценты	N	Проценты		
1	553	22,0%	454	44,0%	82,1%	198,0%
8	261	10,0%	211	20,0%	80,0%	195,0%
9	483	19,0%	211	20,0%	43,0%	105,0%
6	455	18,0%	80	7,0%	17,0%	42,0%
5	390	15,0%	54	5,0%	13,0%	33,0%
7	322	13,1%	10	0,0%	3,0%	7,0%

Метод построения: CHAID

Зависимая переменная: Credit rating

Классификация

Наблюдаемые	Предсказанные		
	Bad	Good	Процент правильных
Bad	665	355	65,0%
Good	149	1295	89,0%
Общая процентная доля	33,0%	66,0%	79,0%

Метод построения: CHAID

Зависимая переменная: Credit rating

**Спасибо
за внимание!**