



В.Домрачев

Технології аналізу даних. (ч.2, приклад)

Теорія та практика побудови скорингових карт у системі SAS enterprise miner



Київ - 2019



Содержание:

- **Что такое скоринг**
- **Необходимость применения скоринговых моделей**
- **Практическое применение системы скоринга для оценки кредитных рисков**
- **Предложения**





Історія

Первая система кредитного scoringа появилась в американских банках во время Второй мировой войны. Тогда практически все кредитные аналитики были призваны на фронт. Чтобы компенсировать потерю уходящих специалистов, многие кредитные организации попросили их разработать общие правила принятия решений о выдаче кредитов, которыми бы смогли руководствоваться рядовые сотрудники.

В 1956 году американский инженер Бил Файр и математик Эрл Айзек впервые предложили использовать систему числовых рейтингов, рассчитываемых на основе исторической информации о заемщике для предсказания его дефолта. Метод получил название scoring, и сегодня все розничные кредиты и значительная часть займов малому и среднему бизнесу проходят численное рейтингование на этапе рассмотрения заявки.

Сами изобретатели scoringа основали компанию Fair Isaac Company, которая в настоящее время носит название FICO.

Что такое Credit Scoring?

- Математическая модель
(статистические методы)
- Анализ кредитной истории «прошлых»
клиентов
- Определение вероятности неблагоприятных
событий
 - Например, вероятность того, что конкретный
потенциальный заемщик (или существующий
клиент) вернет кредит (вернет кредит в срок).
- Принятие обоснованных решений

- Скоринг – балльное оценивание
- Скоринговый алгоритм – способ расчета балла
- Кредитный скоринг – оценка кредитного рейтинга с помощью некоторой шкалы или алгоритма
- Кредитно-скориговая карта – таблица, по которой рассчитывается кредитный рейтинг
- Балл отсечения – пороговое значение скоринг-балла



- ‘What is credit scoring?’ Simply stated, it is the use of statistical models to transform relevant data into numerical measures that guide credit decisions.

The Credit Scoring Toolkit
Theory and Practice for Retail Credit Risk
Management and Decision Automation

Raymond Anderson

Published in the United States
by Oxford University Press Inc., New York
© Raymond Anderson 2007



Скоринговая карта (CK)

- Credit scoring is the use of predictive models (algorithms), to rank cases by their probability of being ‘good’ or ‘bad’ at a future date, based upon lenders’ past experiences.
- Most people understand a scorecard as a piece of paper that allows a scorekeeper, spectator, or participant to keep track of competitors’ performance in a sporting activity.
- The final scorecard can be presented to the layman as a series of statements, as shown above.

*Alan Greenspan, U.S. Federal Reserve Chairman,
in an October 2002 to the American Bankers Association.*

- [The use of credit scoring technologies] has expanded well beyond their original purpose of assessing credit risk. Today they are used for assessing the risk-adjusted profitability of account relationships, for establishing the initial and ongoing credit limits available to borrowers, and for assisting in a range of activities in loan servicing, including fraud detection, delinquency intervention, and loss mitigation. These diverse applications have played a major role in promoting the efficiency and expanding the scope of our credit delivery systems and allowing lenders to broaden the populations they are willing and able to serve profitably.

- «Скоринговые системы дают количественную оценку будущего состояния или доходности заявителя.
- Кредитная политика – это свод правил, которые могут использовать результаты скоринговых систем для принятия решения о выдаче кредита.
- Система андеррайтинга является комплексным понятием. С ее помощью можно определить, отвечает ли претендент требованиям программы кредитования ..., достаточна ли его документация и т.д.; она включает в себя скоринговую систему и кредитную политику в качестве модуля принятия решений – ее «мозга».

Руководство по кредитному scoring. Под ред. Элизабет Мэйз. – Минск, 2008. – 464 с.

Базель

- Internal ratings and default and loss estimates must play an essential role in the credit approval, risk management, internal capital allocation, and corporate governance functions of banks using the IRB approach. *Rating systems and estimates designed and implemented exclusively for the purpose of qualifying for the IRB approach and used only to provide IRB estimates are not acceptable.* It is recognised that banks will not necessarily be using exactly the same estimates for both IRB and internal purposes. For example, pricing models are likely to use PDs and LGDs relevant to the life of the asset. Where there are such differences, a bank must document them and demonstrate their reasonableness to their supervisor.
- The ‘use test’—Basel II Framework paragraph 444
- Credit scoring was initially used to provide risk rankings (power), and lenders are now also expected to provide reasonable default rate estimates (accuracy).
- Базель II предоставляет банкам возможность использовать **внутренние рейтинги** при расчете достаточности капитала для покрытия кредитных рисков.
- Кредитный скоринг позволяет оценить вероятность дефолта контрагента (PD), рассчитать ожидаемые (Loss Given Default) потери и выполнить требования Базель II: провести четкое разграничение рисков и их **точное количественное выражение.**



Скоринговые модели

- Сегодня банки оценивают потенциальных заемщиков с помощью скоринговых моделей. Это своеобразные весы, которые взвешивают математически рассчитанную способность клиента вовремя расплатиться с кредитором. Поскольку скоринговые модели базируются на статистических законах, при оценке потенциальных заемщиков возможны ошибки.

Скоринговые модели

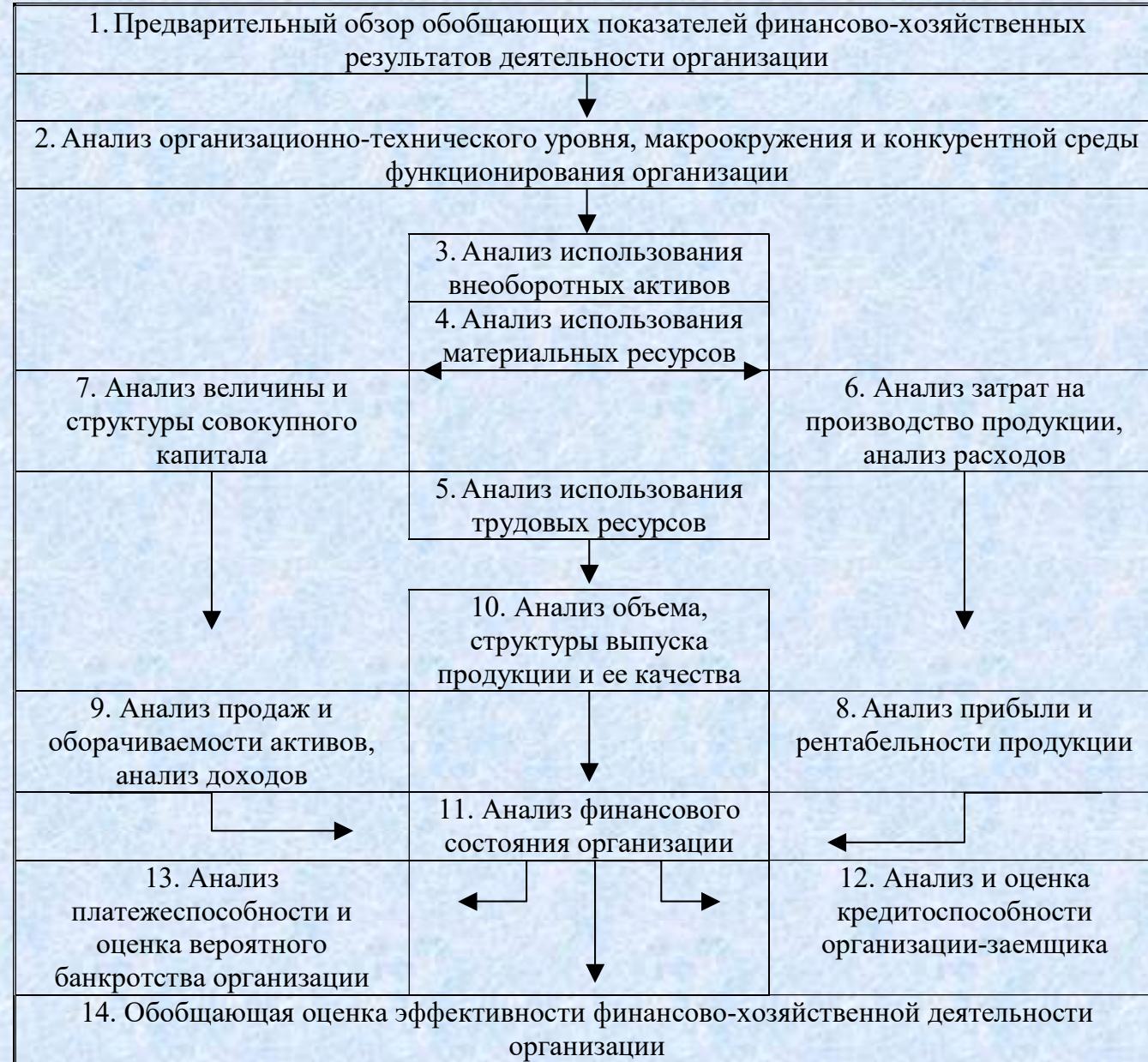
- В общепринятой практике кредитный scoring определяется двумя задачами, каждая из которых имеет свои характерные аспекты и особенности.
 1. Создание скоринговых моделей – моделей оценки кредитоспособности
 2. Построение скоринговой инфраструктуры

В январе 1941 года Национальное бюро экономических исследований США опубликовало исследование Дэвида Дюрана «Элементы риска потребительского кредитования в рассрочку» (Risk Elements in Consumer Installment Financing). Д. Дюран определил основные группы факторов, максимально влияющие на степень кредитного риска, и коэффициенты влияния каждого из них.

- Предлагалось использовать следующие факторы и правила их учета.
- возраст - 0,1 балла за каждый год свыше 20 лет (максимум - 0,30);
- пол - женский (0,40), мужской (0);
- срок проживания - 0,042 за каждый год в данной местности;
- профессия - 0,55 - за профессию с низким риском, 0 - за профессию с высоким риском, 0,16 - другие профессии;
- работа - 0,21 - предприятия в общественной отрасли, 0 - другие;
- занятость - 0,059 - за каждый год работы на данном предприятии;
- финансовые показатели - наличие банковского счета - 0,45, наличие недвижимости - 0,35, наличие полиса по страхованию - 0,19.
- Если набранная сумма баллов не превышает 1,25, то заемщик считается неплатежеспособным, в противном случае - кредитоспособным.

Юрики (пример)

- Заемщик 1:
 - - Тип – Производственная; Торговая ...
 - - Количество сотрудников – до 1000 человек;
 - - Сектор рынка – Нефтяная промышленность;
 - - Количество лет на рынке – свыше 20 лет;
 - - Доля собственных средств предприятия – до 60%;
 - - Доля заемных средств предприятия – до 40%;
 - - Обороты по счету – до 500 млн. грн.;
 - - Кредитная история – Неблагоприятная;
 - - Владение недвижимостью – Владеет на сумму до 10 млн. грн.



- Операционная рентабельность (%) ,
- Чистая рентабельность (%) ,
- Рентабельность капитала и резервов (%) ,
- Общая ликвидность ,
- Абсолютная ликвидность ,
- Скорость оборота активов (%) ,
- Скорость оборота дебиторов (дни),
- Скорость оборота запасов (дни) ,
- Скорость оборота задолженности поставщикам (дни),
- Дефицит финансирования (дни) ,
- Займы в кредитных учреждениях /активы,
- Платежеспособность (собственный капитал/активы)

Виды скоринга



Кредитный (либо анкетный) скоринг

- Кредитный (либо анкетный) скоринг (англоязычный эквивалент – application scoring) — получение показателя кредитоспособности потенциального заемщика на основе некоторых его характеристик, прежде всего содержащихся в анкете заемщика. Внедрение данного вида скоринга позволяет банку
 - повысить точность оценки заемщика,
 - ускорить саму процедуру оценки,
 - минимизировать человеческий фактор в принятии решения,
 - создать централизованное накопление данных о заемщиках,
 - снизить формируемые резервы на возможные потери по кредитным обязательствам.

Поведенческий скоринг (behaviourscoring)

- Поведенческий скоринг (behaviour scoring) — динамическая оценка ожидаемого поведения клиента по погашению кредита, основанная на данных об истории трансакций по его счетам и используемая, в частности, для предупреждения возникновения задолженности.

Коллекторский скоринг

- Коллекторский скоринг (скоринг взыскания, collection scoring) — определение приоритетных направлений работы в отношении заемщиков с кредитным счетом, состояние которого классифицировано как «неудовлетворительное».





Скоринг мошенничества (**fraud scoring**)

- Скоринг мошенничества (**fraud scoring**) — скоринг, направленный на выявление возможных мошенников среди лиц, претендующих на получение кредита или уже существующих клиентов-заемщиков. Этот тип скоринга, как правило, используется в связке с **application-** и **behavioral-** скорингом для более детального анализа заемщиков.



Формальные критерии:

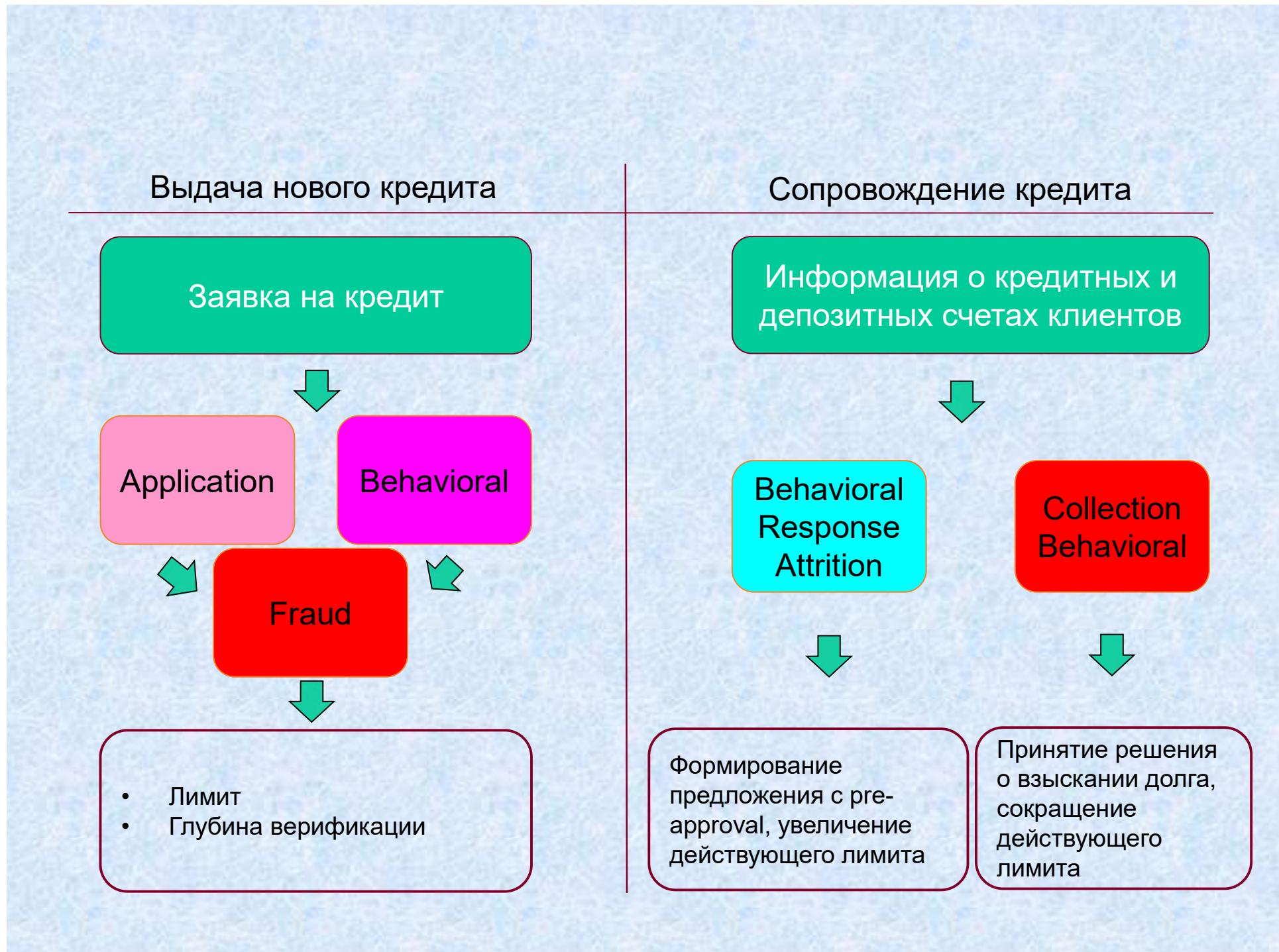
- ✓ Клиент не проходит по возрасту
- ✓ Паспорт клиента просрочен
- ✓ Проверка стажа работы
- ✓ Проверка срока постоянной регистрации
- ✓ Проверка необходимого количества стационарных телефонов
- ✓ Отсутствие необходимых поручителей
- ✓ Проверка наличия фотографии в заявке

Признаки мошенничества клиента:

- ✓ Телефон, указанный как домашний, фигурирует в базе как рабочий, либо наоборот
- ✓ Домашний телефон не соответствует фактическому адресу проживания
- ✓ Рабочий телефон фигурирует в базе как рабочий, но на другом месте работы

Признаки внутреннего мошенничества:

- ✓ Еженедельный (ежемесячный) лимит выдач на кредитного эксперта, точку, группу точек, город
- ✓ Кредитный эксперт выдал более N кредитов на одинаковую сумму в течение заданного срока



Этапы скоринга

- Процесс разработки скоринговой модели можно разделить на несколько уровней:
- Изучение предметной области
- Выборка возможных значимых факторов
- Препроцессинг (подготовка данных для алгоритма)
- Построение (тренировка) модели калибровка
- Применение полученной модели
- Оценка эффективности и интерпретация результатов

Данные для скоринга

- **Пользователь:** анкета, доп. документация
- **Внутренняя информация:** счета, кредиты, депозиты
- **Внешняя:** кредитные бюро (отличается от скорингового), другая инфа
- **Другая информация:** экономические показатели, рейтинг страны, отрасли, региона
- **Обычно:** От 40 до 200 параметров для выдачи и от 60 до 120 для выданных

Общепринятая информация

Категории	Параметры
1. Персональная информация	Возраст, Пол, Телефон, Гражданство
2. Информация о семье	Семейный статус, Количество содержальцев
3. Информация о проживании	Количество лет проживания по данному адресу, Вид жилья
4. Занятость	Количество лет работы на данном месте, Позиция (должность)
5. Финансовое состояние	Основные активы, Текущие кредиты, Состояние счета, Количество ранее возвращенных кредитов
6. Обеспечение (залог)	Стоймость залога, Гарант
7. Прочее	Цель кредита, Сумма кредита, Срок кредита

Пример анкеты

	Personal:	Work:	Financial:	Bank:
Contact	Name, Title, ID number	Employer name	Income:	Name and branch Account type Account number References Credit/store cards held
	Address/Previous Address		Own Spouse Other	
	Phone numbers: Home, Work, Cell, Fax		Expenses:	
	Email		Rent/Bond Motor vehicle Other credit	Request: Loan amount Repayment: Method Period Frequency Insurance Opt out clauses Other
Stability	Time @ address	Time @ employment	Balance sheet: Assets Liabilities	
	Time @ previous address	Time @ previous employment		
Demographics	Gender	Type of business/ Industry	Security: Goods: Age Type Surety Guarantor	
	Age/ Date of birth			References: Credit Personal
	Marital status/ #Dependants	Employment level		
	Accommodation type	Level of education	Signature:	Conditions:

Сопровождение кредита

Внутренняя информация

Static	Dynamic
Product type	Outstanding balance
Open date	Payment due
Market segment	Credits/repayments
Original loan/account limit	Debits/purchases
Loan term	Available credit
Cycle/billing date	Interest income
Interest rate	Fee income
Repayment method	Date last payment
Settlement value	Date last purchase
Date closed	Arrears amount
Date in recoveries	Arrears in months
Lost/stolen/fraud/deceased indicators	Times in arrears

Приклад карти

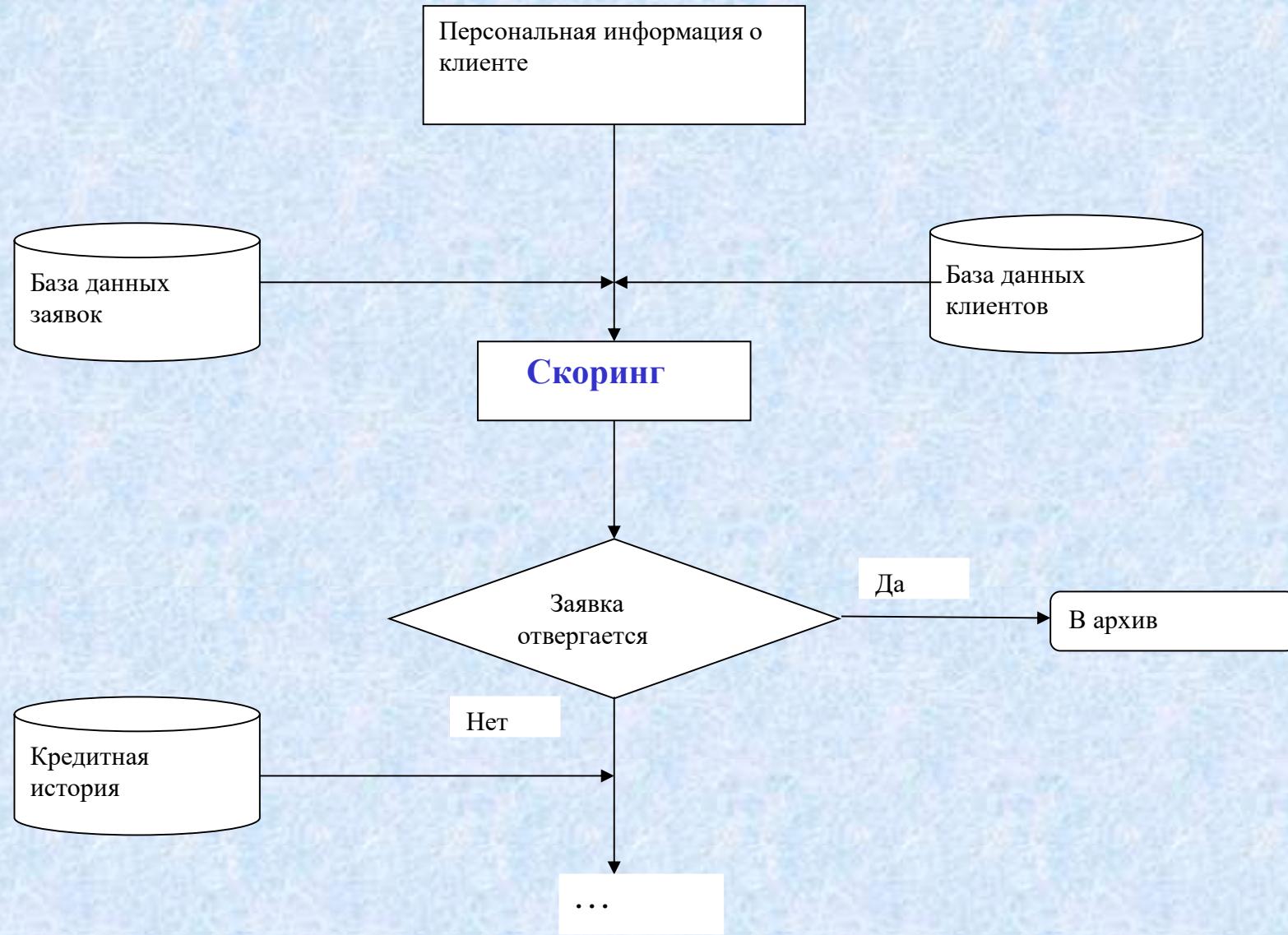
Скорингова карта

Показник	Значення	Балів	Бал
Вік	<20	15	0
	20-25	34	0
	25-30	55	0
	30-35	90	0
	35-50	114	0
	50-60	97	0
	>60	15	0
Чоловік / жінка		5	0
Сімейний стан	неодруж.	87	0
	шлюб: (заміжня)	115	0
	Женат (заміжня), мешкають окремо	30	0
	Розлучений	70	0
	Вдовий	65	0
Кількість дітей	0	87	0
	1	64	0
	2	52	0
	3	14	0
	більше 3	4	0
Діяльність	Держслужба	93	0
	Комерційна структура	124	0
	Пенсіонер	19	0
	Інше	47	0
Кваліфікація	Відсутня	3	0
	Обслуга	17	0
	Спеціаліст	72	0
	Службовець	83	0
	Керівник	122	0
Стаж роботи на останній	< року	6	0
	< двох років	28	0
	< 3 років	51	0
	< 5 років	62	0
	> 5 років	89	0

Методы скоринга

- В настоящее время в кредитном скоринге используются следующие методы, причем они могут применяться как отдельно друг от друга, так и в различных комбинациях:
- **методы статистики** (дискриминантный анализ, линейная регрессия, логистическая регрессия, деревья классификации); использование статистических методов сводится к построению правила классификации, основанного на линейной скоринговой функции;
- **методы исследования операций** (линейное программирование, нелинейная оптимизация);
- **методы искусственного интеллекта** (нейронные сети, экспертные системы, генетические алгоритмы, методы ближайших соседей, байесовские сети, логико-вероятностные методы).

Место скоринга в принятии решения о выдаче кредита



Мировая практика кредитования

- 35% история платежей
- 30% характеристики клиента
- 15% длина кредитной истории
- 10% новые кредиты
- 10% новые инструменты кредитования

КРЕДИТЫ, ВЫДАННЫЕ В УКРАИНЕ ФИЗЛИЦАМ ПО ЦЕЛЕВОМУ НАЗНАЧЕНИЮ, МЛРД ГРН*



**Потребительские кредиты
(включая кредиты на покупку автомобилей) – 131,5**



**На приобретение, строительство и
реконструкцию недвижимости – 54,7**



Другие кредиты – 4,5

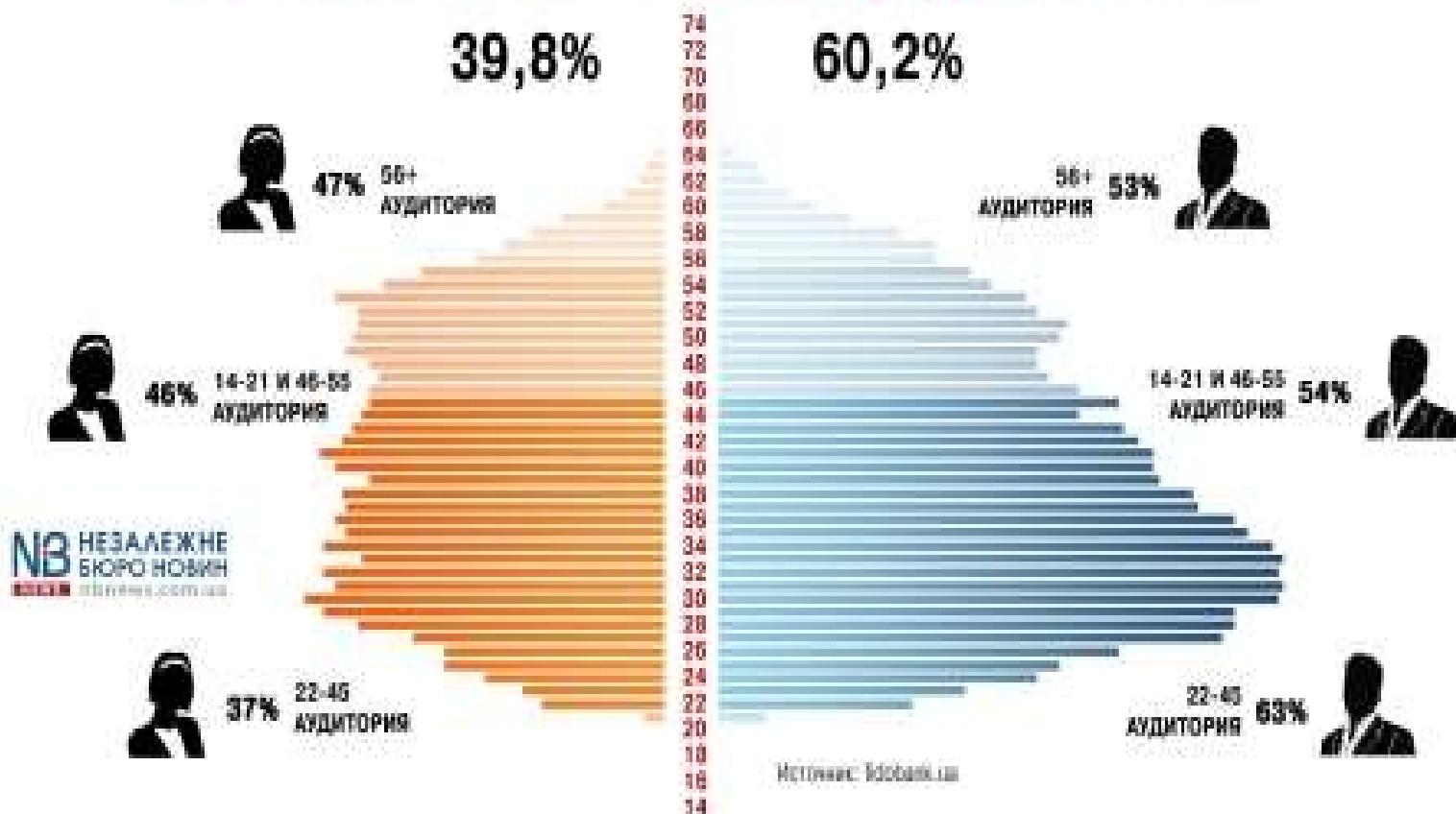
ВСЕГО - 190,8

*Данные НБУ на 1 августа 2013 года

NB НЕЗАЛЕЖНЕ
БЮРО НОВИН
NEWS nbnews.com.ua

<http://nbnews.com.ua/ru/tema/109315/>

ПОРТРЕТ УКРАИНСКОГО ДОЛЖНИКА



Кто строит скоринговые карты

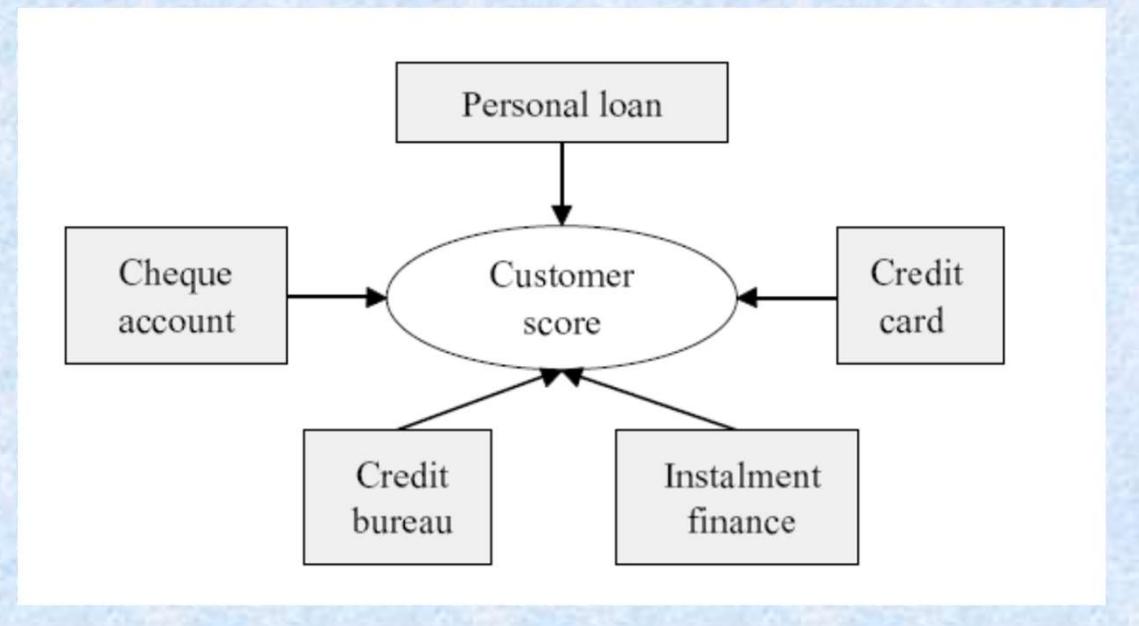
- Банки сами
 - Кредитные бюро
 - Скоринговые бюро
 - Специальные предприятия
 - Рейтинговые компании
-
- Как часто
 - Цена
 - Управление

Отклонение заявки на кредит

- Информация по несоставшимся кредитам не может быть использована в качестве обучающей выборки.
- Некоторые из несоставшихся кредитов могли бы быть выданы и скоринговые расчеты были бы тогда другими.
- Но даже если бы все отклоненные попали в плохие скоринговые расчеты и тогда были бы другими.
- Т.е. в оценке кредитоспособности новых претендентов содержится некая систематическая ошибка.
- Смещение результатов scoringа происходит из-за того, что в обучающей выборке содержатся только состоявшиеся заемщики.
- Степень этой ошибки можно снизить.

Виды скоринговых карт

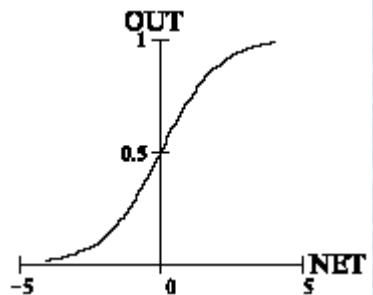
- Универсальные
- По сегментам рынка
- По регионам
- По банкам



Виды скоринговых алгоритмов

- Логистическая регрессия – наиболее часто используемый метод построения скоринговых карт
- Деревья решений
- Нейронные сети



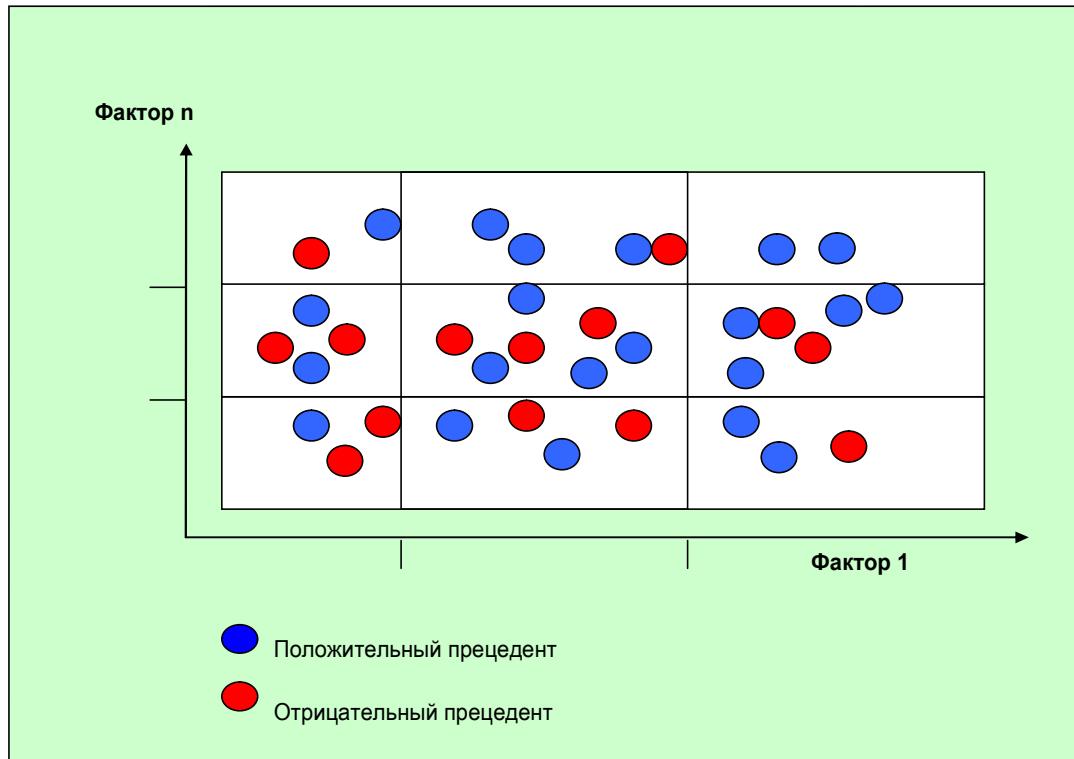


Логистическая регрессия

- Логистическая регрессия применяется для предсказания вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится так называемая зависимая переменная, принимающая лишь одно из двух значений — в случае скоринговой модели, значения 1 в случае «хорошего» клиента и 0 в случае «плохого», и множество независимых переменных (также называемых признаками, или регрессорами) — вещественных ..., на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной.

Логистическая регрессия

<i>Показатель</i>	<i>Значение</i>	<i>Баллы</i>
...
Возраст	20-25	100
	25-30	107
	30-40	123
Доход	1000-3000	130
	3001-5000	145
	5001-6000	160
Стаж работы

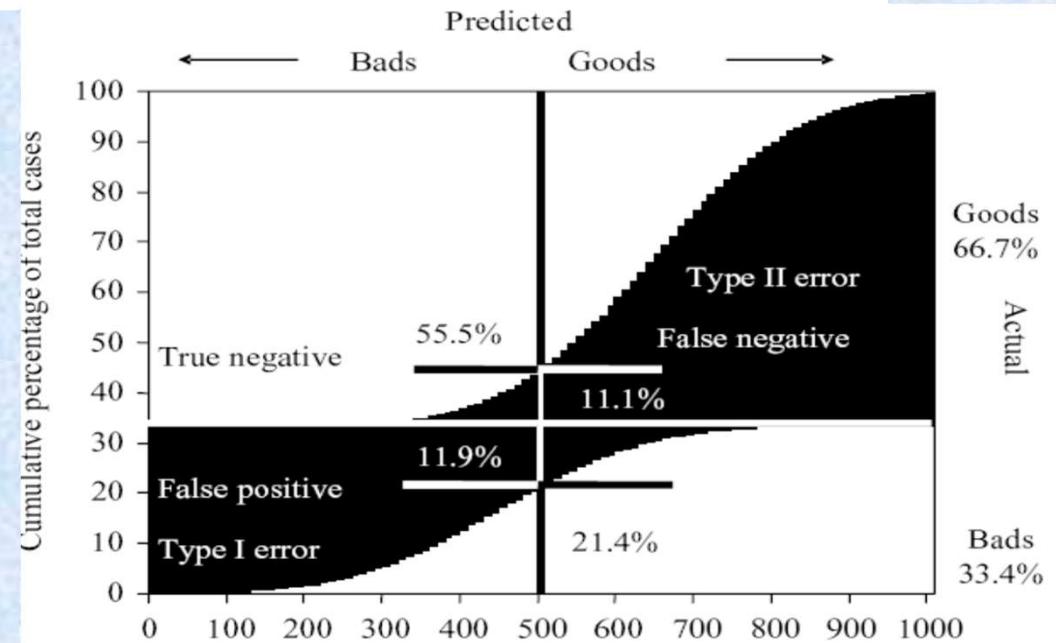
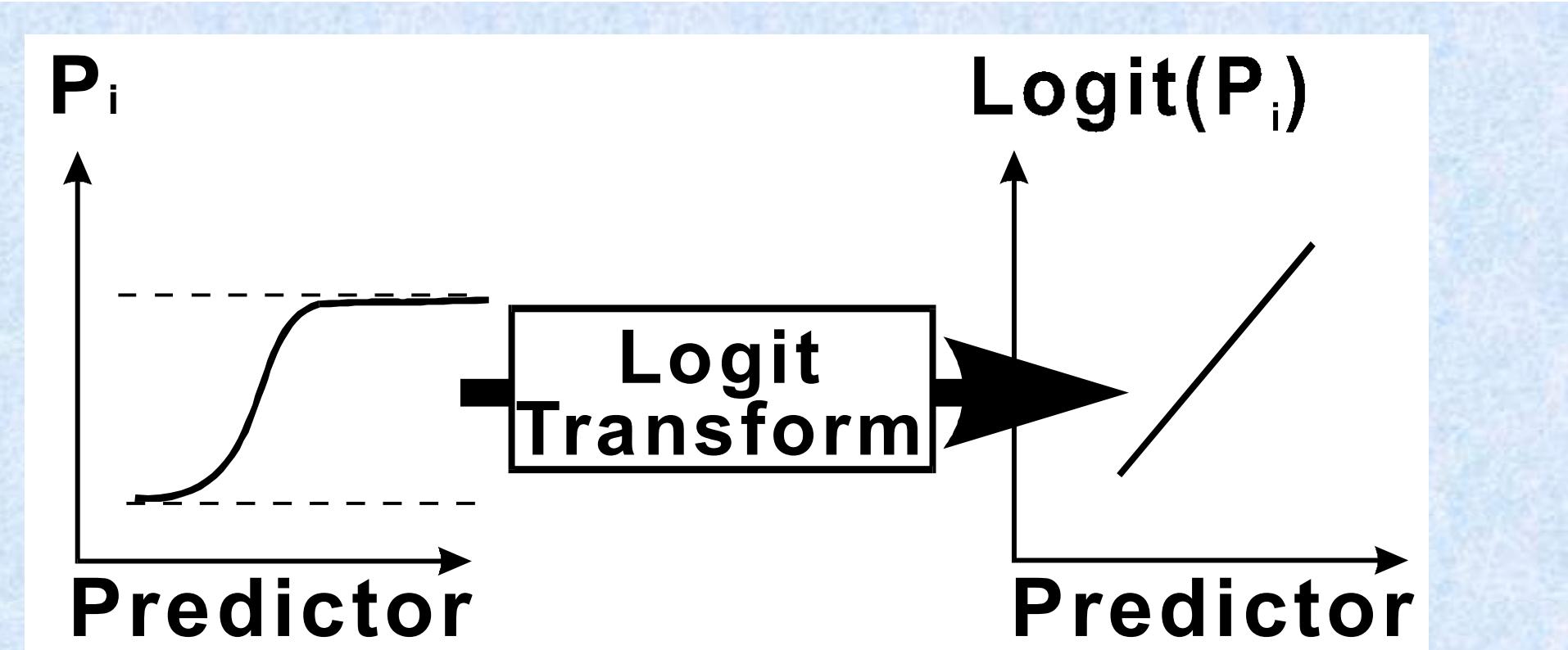


- Каждая ячейка сетки объединяет прецеденты из выборки с одинаковой вероятностью исхода. Координаты узлов этой сетки рассчитываются на основании статистических критериев, исходя из принципа максимальности различия между вероятностями исходов кредитных сделок для смежных сегментов прецедентов. Распределение + и – прецедентов по ячейкам будет определять весовые коэффициенты при регрессивных членах, которые в свою очередь позволяют рассчитать систему баллов для скоринговой карты.

Математическая основа логит регрессии

- В логит регрессионной модели предсказанные значения зависимой переменной или переменной отклика не могут быть меньше (или равными) 0, или больше (или равными) 1, не зависимо от значений независимых переменных; поэтому, эта модель часто используется для анализа бинарных зависимых переменных или переменных отклика. При этом используется следующее уравнение регрессии (термин логит был впервые использован Berkson, 1944):
 - $y = \exp(b_0 + b_1 * x_1 + \dots + b_n * x_n) / [1 + \exp(b_0 + b_1 * x_1 + \dots + b_n * x_n)]$
 - Легко увидеть, что независимо от регрессионных коэффициентов или величин x , предсказанные значения (y) в этой модели всегда будут лежать в диапазоне от 0 до 1.
 - $p' = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$
 $p' = \ln \{p/(1-p)\}$
p – вероятность того, что событие (дефолт) произойдет
p/(1-p) – шанс того, что событие произойдет
для решения задач логит регрессии используется только метод максимального правдоподобия.

Скоринговый бал = $(b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n) (20 / \ln(2))$



Логистическая регрессия

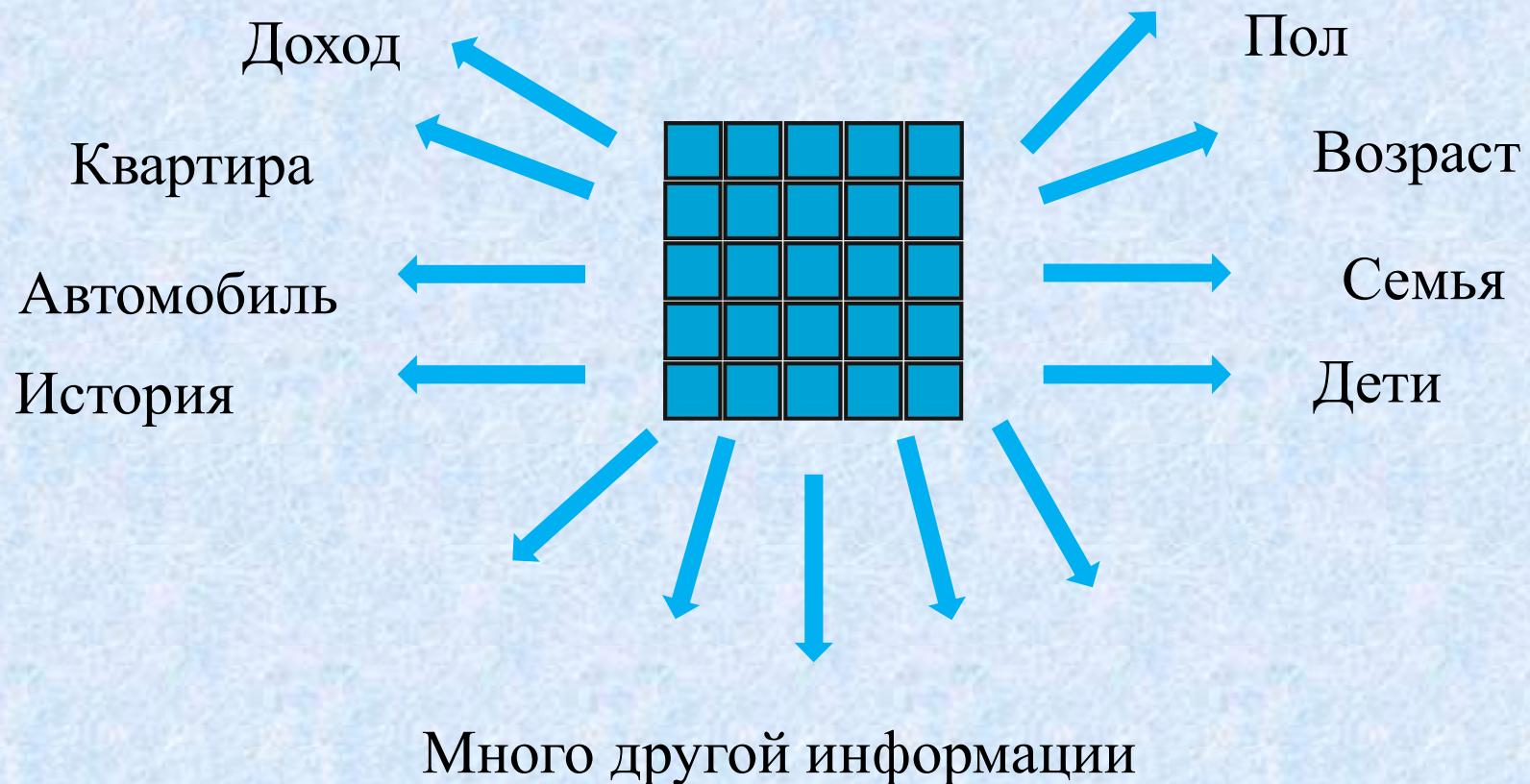
- Logistic regression requires the following assumptions:
 - (i) categorical target variable;
 - (ii) linear relationship, but this time with the log odds function;
 - (iii) independent error terms;
 - (iv) uncorrelated predictors; and
 - (v) use of relevant variables.

Логистическая регрессия

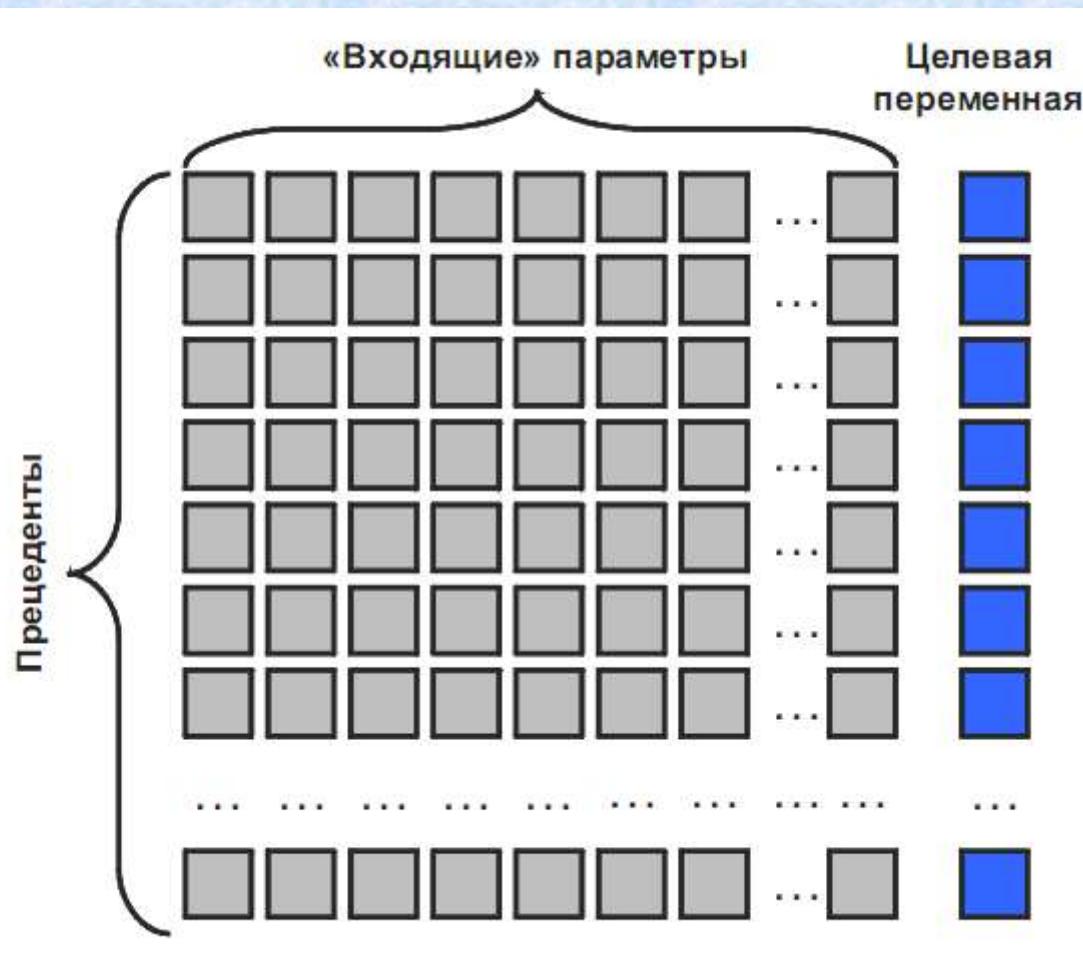
- While used primarily for binary target variables, it is also possible to use ‘ordered logistic regression’ for ordinal outcomes, such as subjective risk grades and survey responses.

Пример

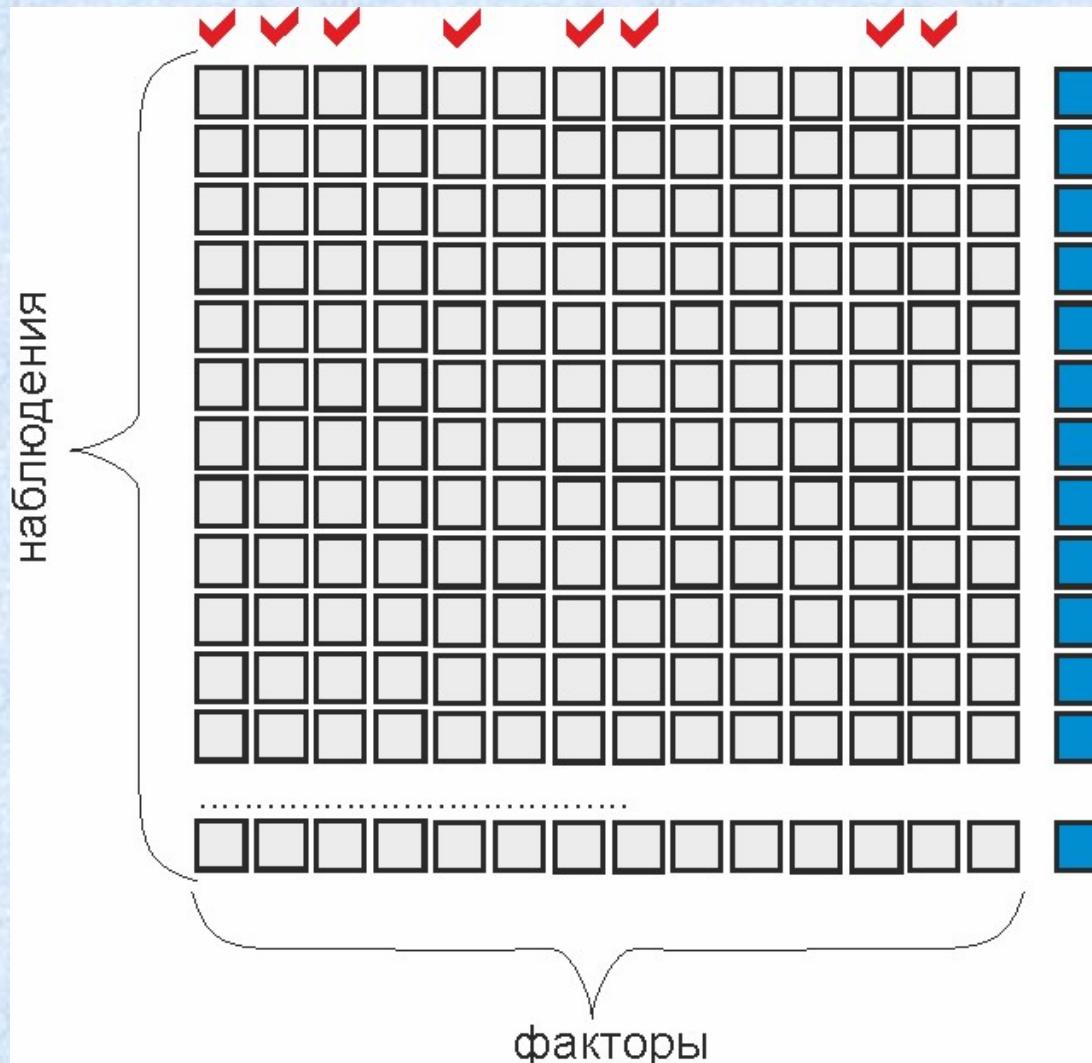
Профиль клиентов, пользующихся
услугой «депозит»



Процесс скоринга

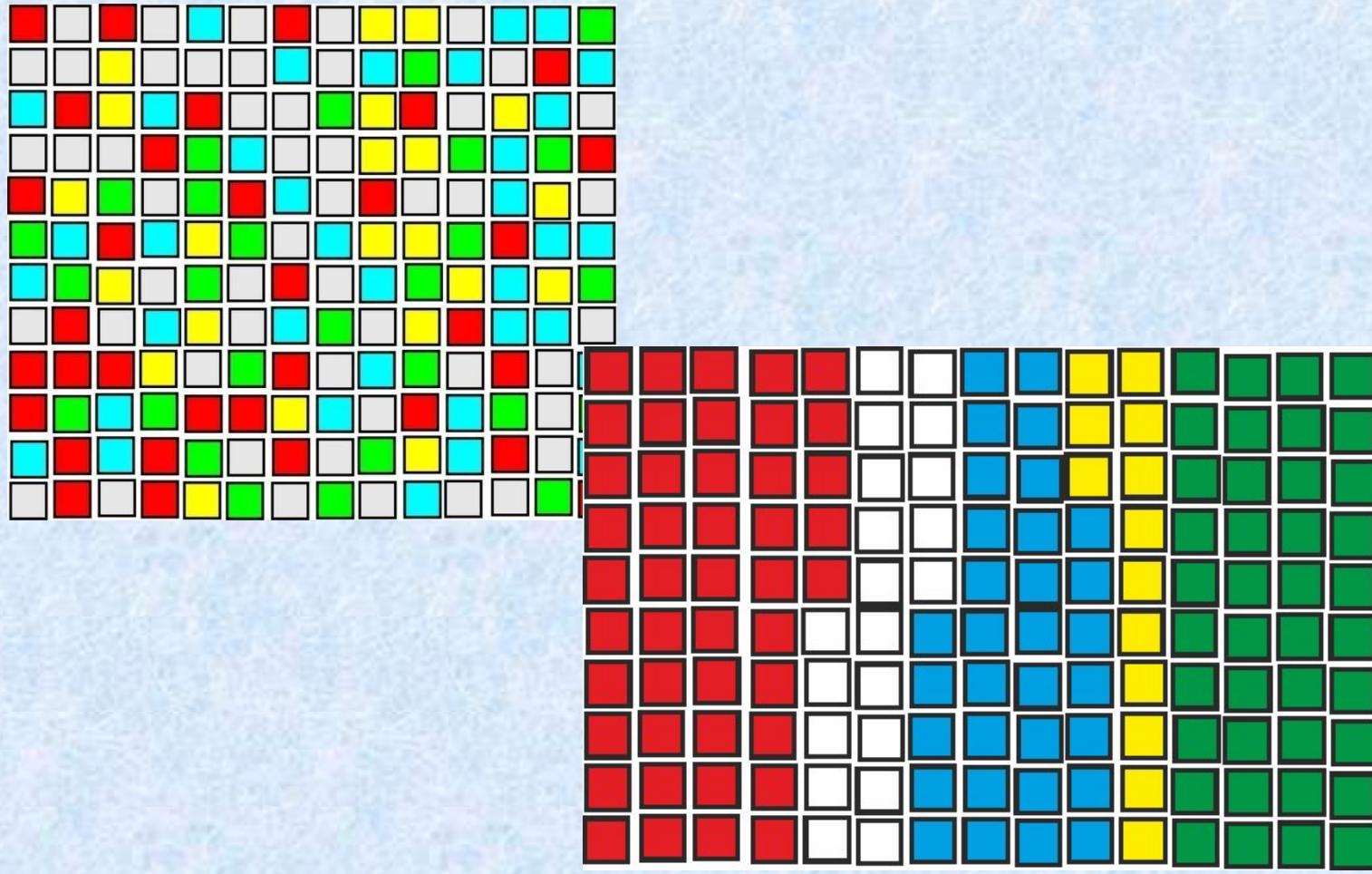


Процесс скоринга (значимые)



Значимость — в статистике: мера уверенности в не случайности полученной величины.

Пример scoringа (веса)



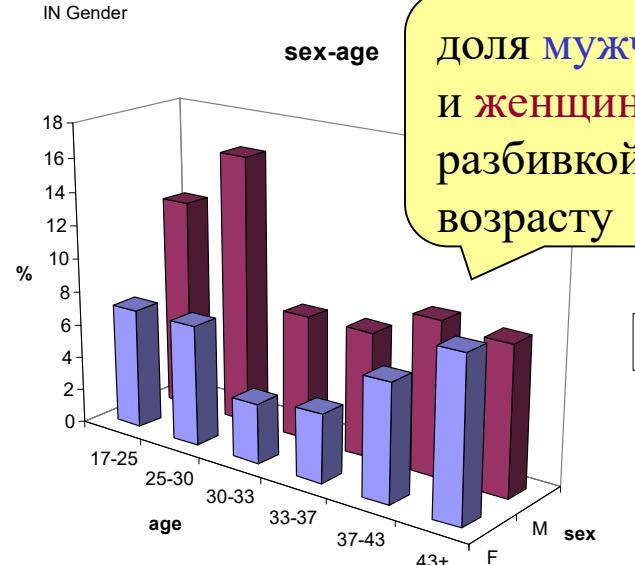
Разработка скоринга

Процесс

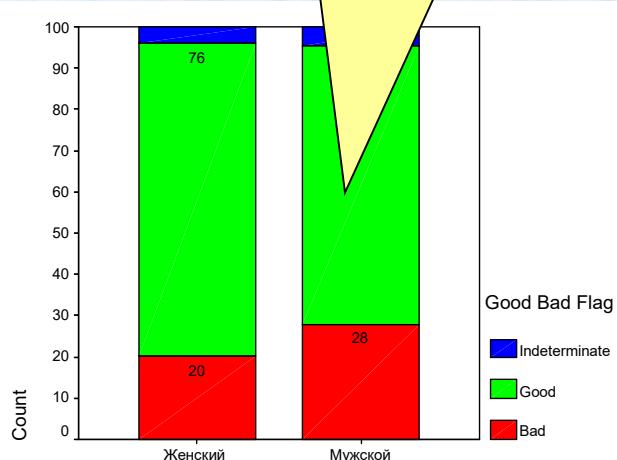
женщин
меньше...



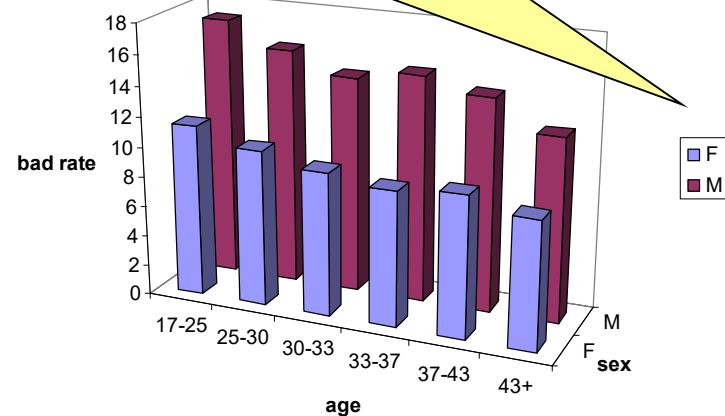
доля мужчин
и женщин с
разбивкой по
возрасту



... но они лучше
выплачивают
кредит



доля плохих заемщиков
среди мужчин и женщин



Важно

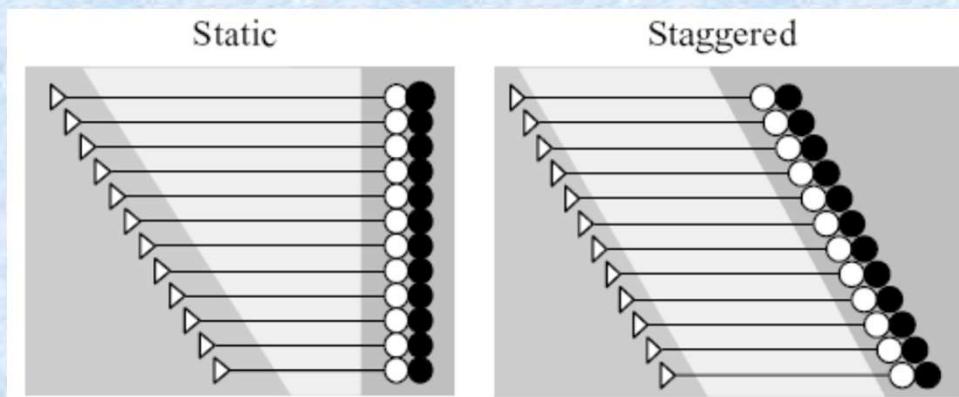
- Выбор правильных шкал параметров (с возможностью менять шкалу).
- Метод выбора (группировки) параметров:
 - **Метод последовательного отбора (Monotonic Event Rate),**
 - Метод лучших подмножеств (Optimal criterion),
 - Метод обратного исключения,
 - Метод прямого отбора,
 - Constrained Optimal,
 - Quantile.
- Статистика Wald это отношение коэффициентов b_i к их стандартному отклонению (ошибке). Тест Wald используется для установления статистической значимости каждого из коэффициентов модели.

Методы группировки

- Метод прямого отбора: пустая модель, вычисляем корреляцию выходной переменной с каждой входной, берем ту где корреляция больше, вычисляем ее значимость, и в случае значимой добавляем ее в модель... Пока все значимые переменные не войдут в модель.

Методы группировки

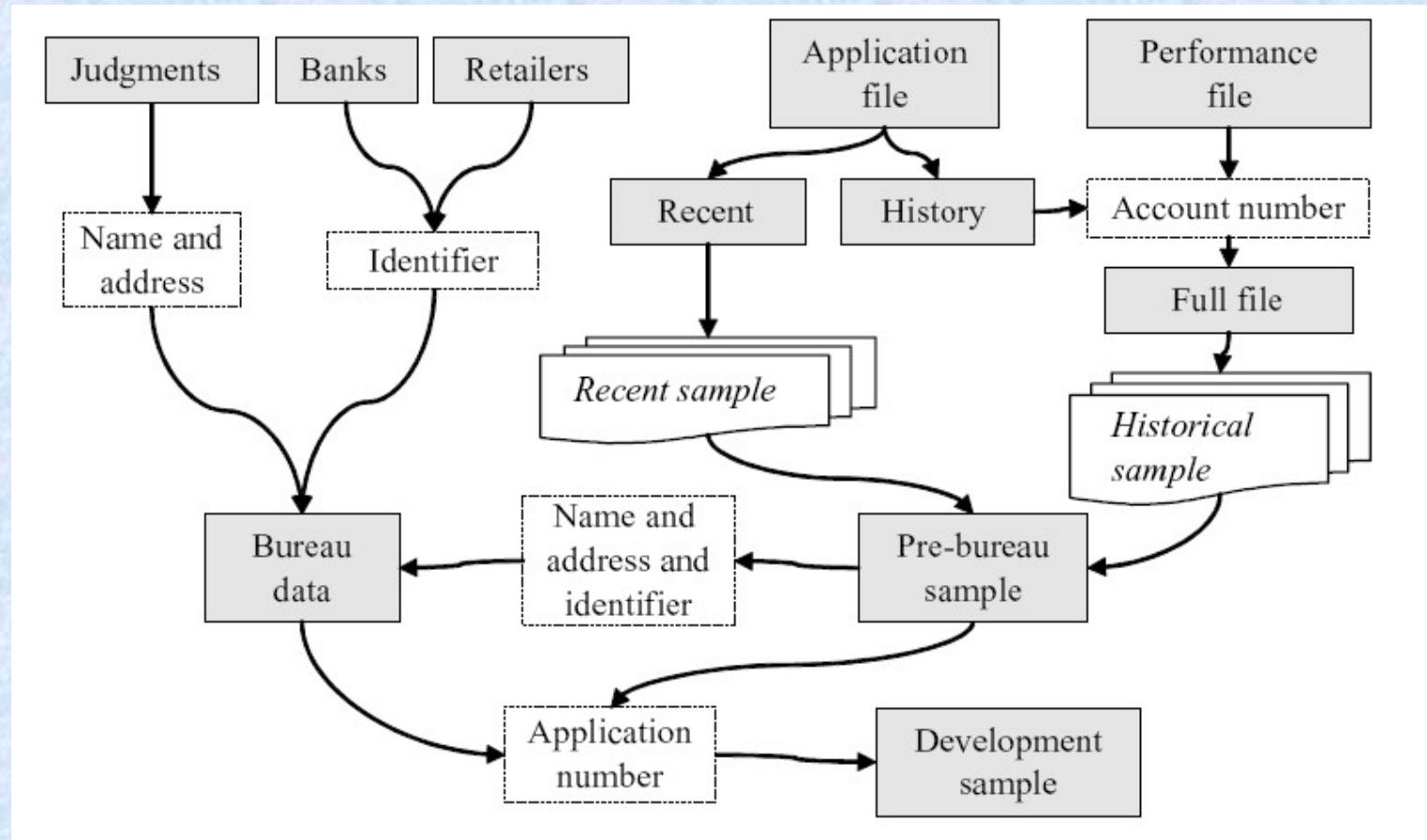
- Метод обратного исключения: процесс начинается с полной модели и последовательно исключаются не значащие переменные.



Методы группировки

- **Метод последовательного отбора:**
развитие Метода прямого отбора.
Переменные введенные ранее после
добавления новых могут стать
незначащими и их выбрасываем.

Процесс подготовки данных scoring



Проблема экспертная: хороший / плохой

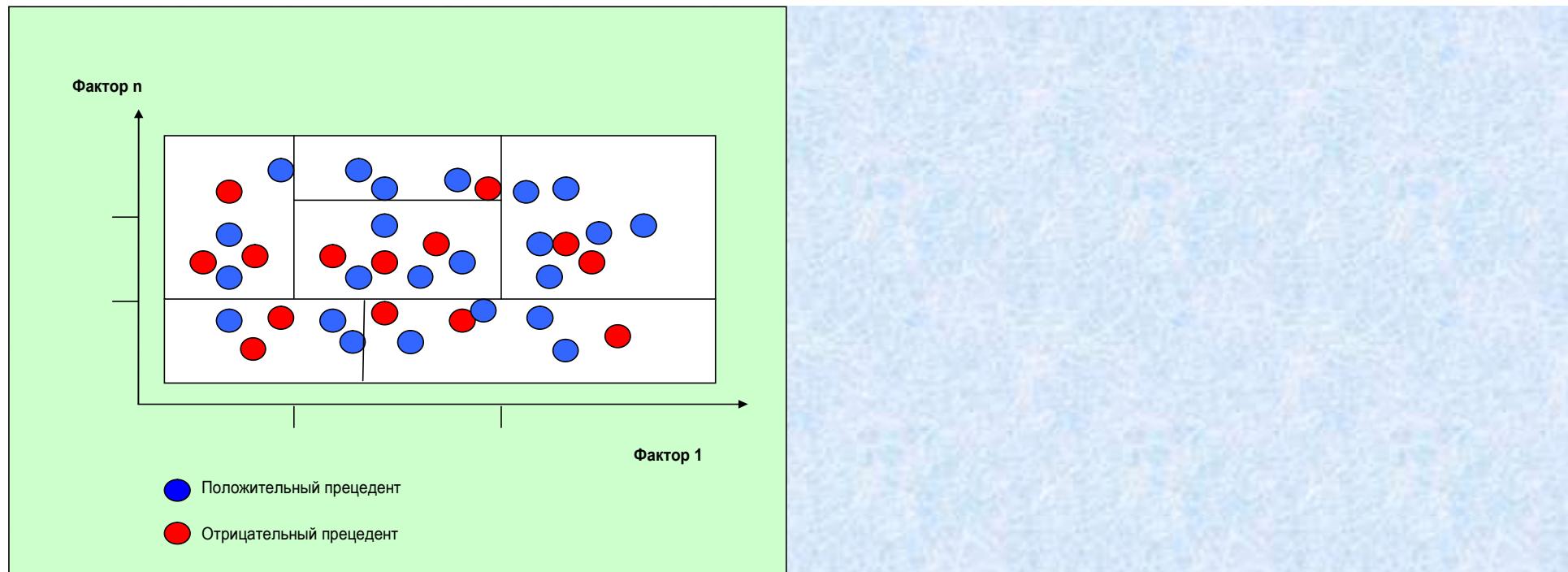
- Классификация плохих заемщиков (активов)
- Хороший – желанный заемщик
- Плохой – избегаем (уже рисковый)
- Неопределенный - ?
- Исключенный – (точно плохой) не рассматриваемый
- Временной фактор (все женихи разобраны)

Другие методы анализа риска



Дерево принятия решений

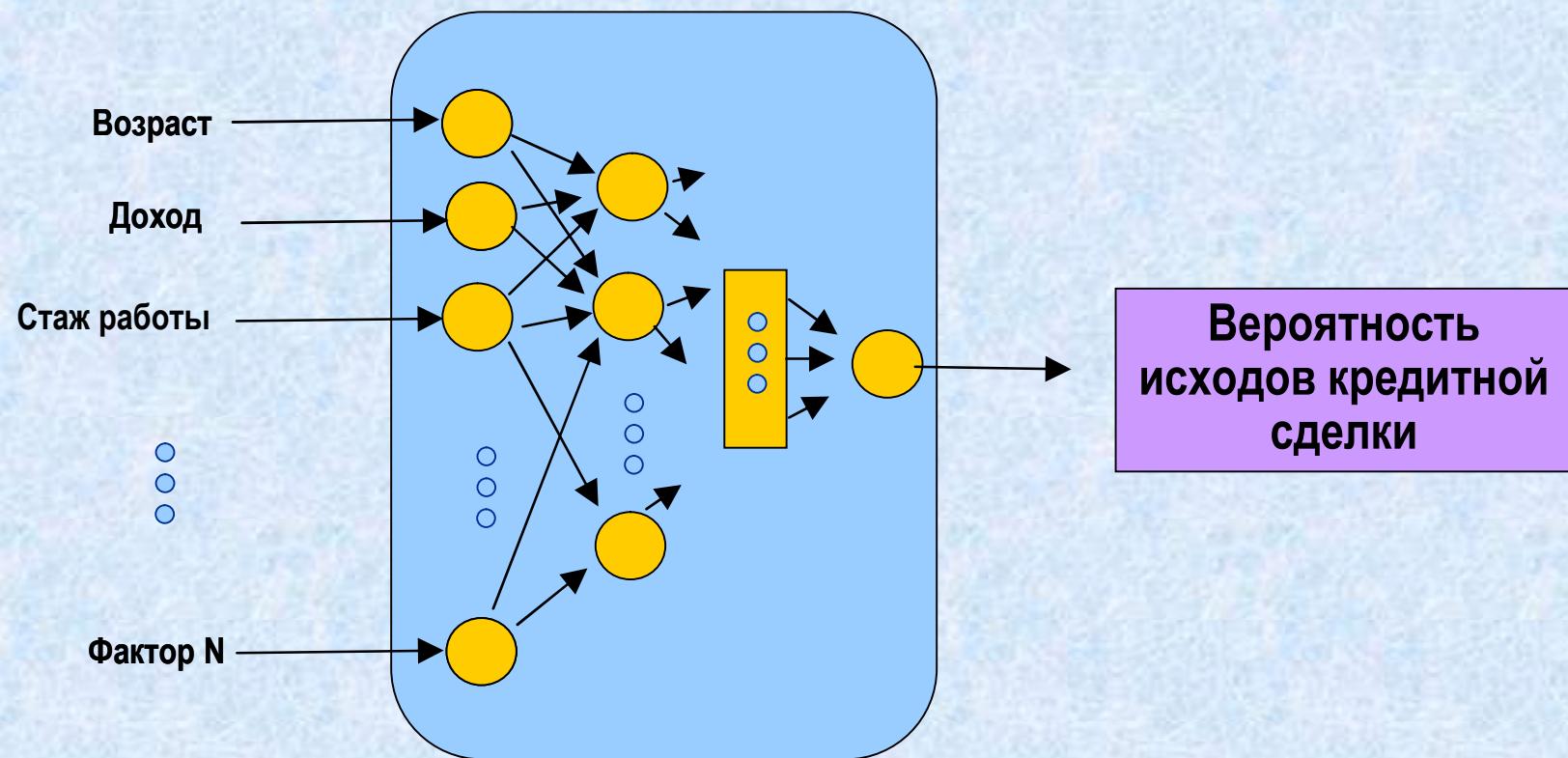
- Деревья классификации(рекурсивные алгоритмы разбиения), в отличие от предыдущих методов, не предназначены для построения скоринговой функции, они последовательно разделяют клиентов на группы по одной из переменных так, чтобы эти группы максимально возможно отличались по величине кредитного риска. Данный процесс продолжается до момента, пока оставшиеся группы не становятся настолько малы, что следующее разбиение не приведёт к статистически значимому различию в уровне риска. Дерево классификаций (дерево решений) является более общим алгоритмом сегментации обучающей выборки прецедентов, чем логистическая регрессия. В отличие от метода логистической регрессии в методе дерева классификации сегментация прецедентов задается не с помощью п-мерной сетки, а путем последовательного дробления факторного пространства на вложенные прямоугольные области (рис.).

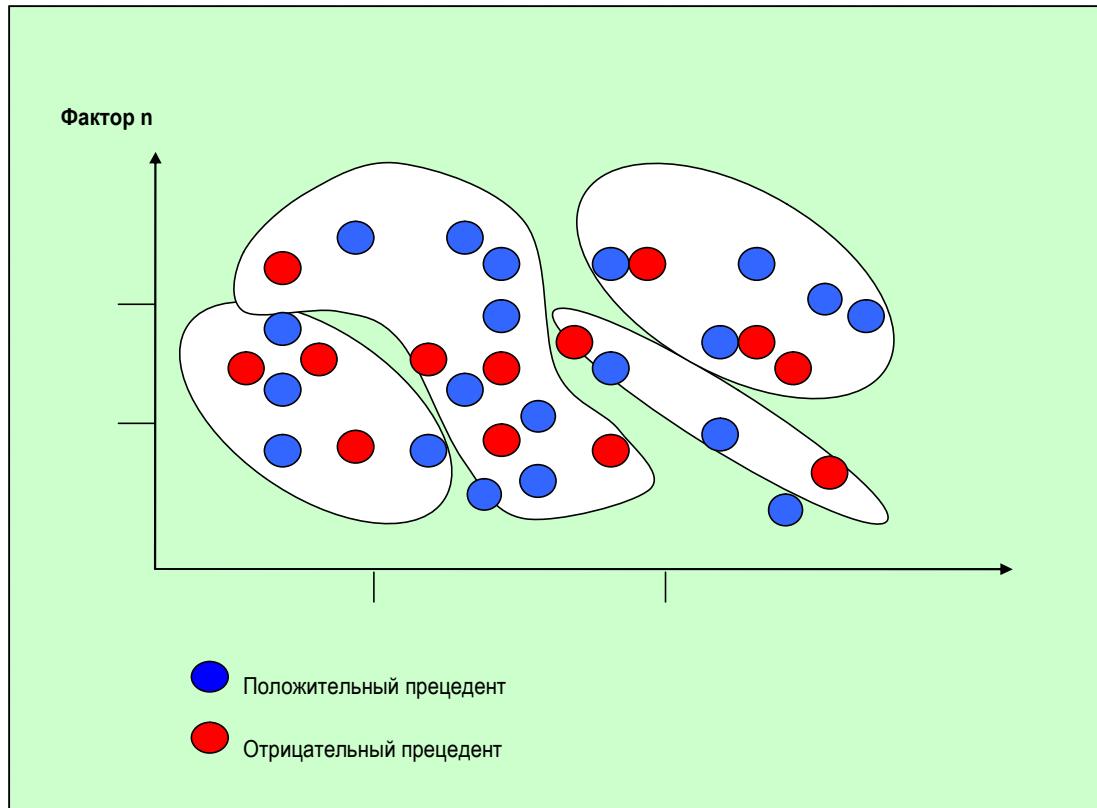


- Является более общим алгоритмом, чем обучающая регрессия. Сегментация происходит путем дробления пространства на прямоугольные области. По этой причине данный метод не приводит к построению классической скоринговой карты, но позволяет для каждого нового апликанта непосредственно рассчитать вероятность дефолта путем дробления сегментов на составные части. Критерием выбора границ сегмента является различие в соотношении + и – прецедентов в каждом вновь образуемом сегменте. Сегментация прекращается, если дальнейшее дробление не приводит к существенным различиям.

Метод нейронных сетей

- Нейронные сети — математические модели, а также их программные или аппаратные реализации, построенные по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток живого организма. Нейронные сети могут рассматриваться в качестве метода нелинейной регрессии. Однако они чаще применяются для скоринга юридических лиц, чем для скоринга частных лиц. Нейронная сеть позволяет обрабатывать прецеденты обучающей выборки с более сложным (чем прямоугольники) видом сегментов





- НС не приводит к построению классической скоринговой карты, но позволяет отнести каждого нового апликанта к сегменту схожих с ним по характеристикам прецедентов и использовать оценку вероятности дефолта для данного сегмента.
- Трудно определить, какой метод наилучший. Только сопоставление предикции и факта может дать оценку эффективности скоринговых моделей.

Метод ближайших соседей

- Метод ближайших соседей (англ. **k-nearestneighbor algorithm, kNN**) - метод автоматической классификации объектов. Основным принципом метода ближайших соседей является то, что объект присваивается тому классу, который является наиболее распространённым среди соседей данного элемента

Другие методы

- Генетические алгоритмы
- Линейное программирование
- ...



Основные источники погрешностей в скоринговых оценках

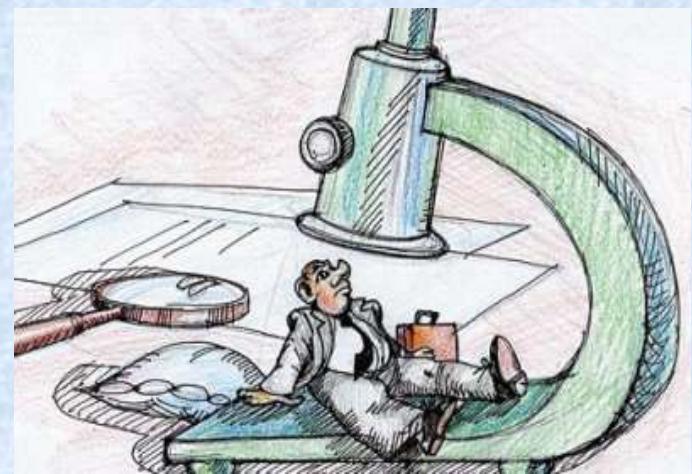
- Принципиально неустранимым источником погрешностей является базовое предположение scoringа об аналогичности поведения новых заемщиков поведению ранее кредитовавшихся клиентов, имеющих аналогичные признаки. Какой бы большой ни была положительная статистика результатов кредитования заемщиков с определенным набором признаков, это не является гарантией того, что обязательства по кредиту, выданному очередному заемщику с такими же признаками, будут им выполнены полностью и своевременно. При анализе результатов кредитования целесообразно выделять такие случаи. Если их число возрастает, это является сигналом о том, что применяемая scoringовая модель теряет актуальность и нуждается в обновлении либо замене.
- Вторым источником являются методические погрешности используемых моделей. Для параметрических методов это погрешности аппроксимации scoringовой функции, причем они могут быть весьма значительны.
- Третий источник погрешностей – это выдача scoringовыми моделями результатов классификации для тех наборов признаков заемщика, по которым в обучающей выборке нет исходных данных.

Основные требования к используемым методам

- **Данные**
 - обрезаются / отвергнутые
 - отсутствуют значения
- **Статистические предположения**
 - нормальность распределения, отсутствие мультиколлинеарности, линейность
- **Переменные**
 - регрессии – итеративные процедуры и результат зависит от используемого инструментария, от начальной модели, от критерия остановки итераций, ...

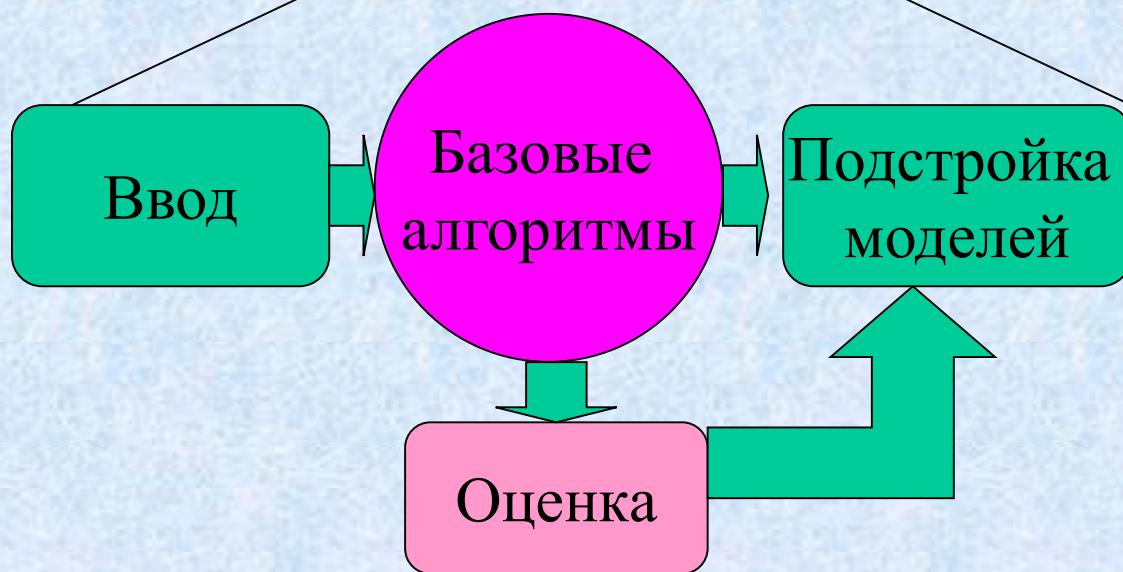
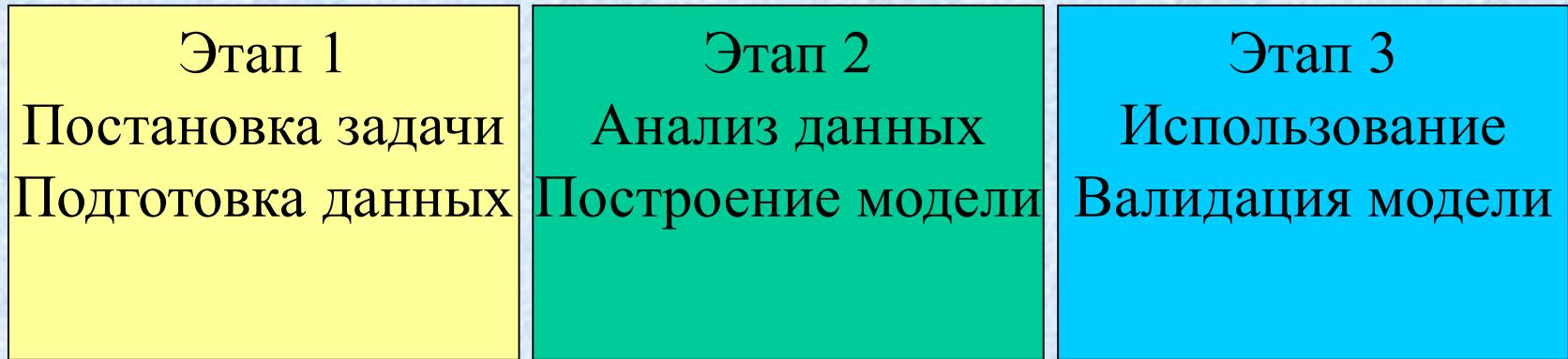
Задача

1. Собрать информацию о клиенте
2. Выявить факторы, влияющие на исход сделки
3. Разработать модель, рассчитывающую вероятность наступления целевого события
4. Определить точку отсечения

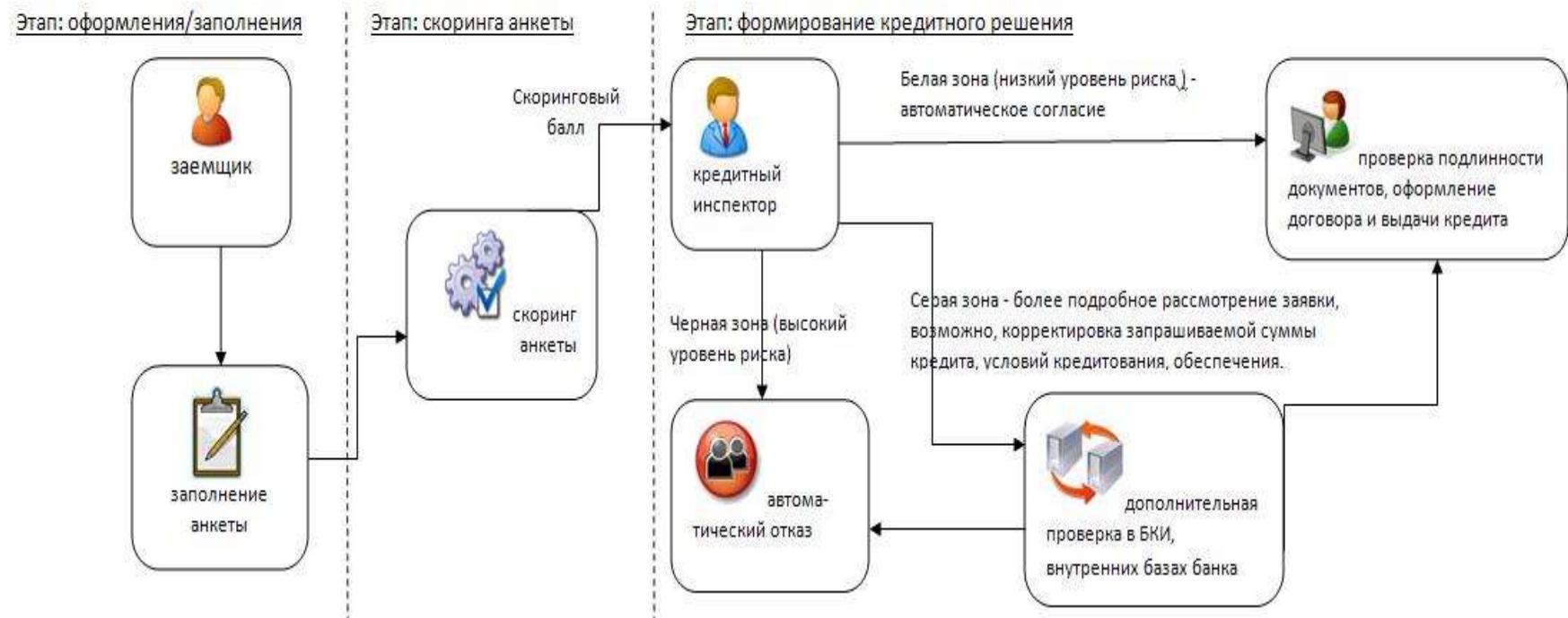


Этапы построения и внедрения скоринговой карты





Типовые бизнес-процессы
со скорингом



Инструменты

- Теоретически создать скоринговую систему сотрудники банка могут самостоятельно, используя «подручные» средства типа программ из набора Microsoft Office. «Построить скоринговую модель можно и с помощью Excel, что и делают специалисты в ряде банков, но процесс этот долгий и хлопотный, требует специальных знаний, не говоря уже о дальнейшем мониторинге таких моделей и их корректировке».
- Правда, в последние годы Excel используется крайне редко, обычно применяется специализированное или универсально-аналитическое (различные математические и статистические пакеты) ПО.

Преимущества скоринга в SAS

Не требует программирования и знания о том,
как производить расчеты

Визуализация процесса обработки информации

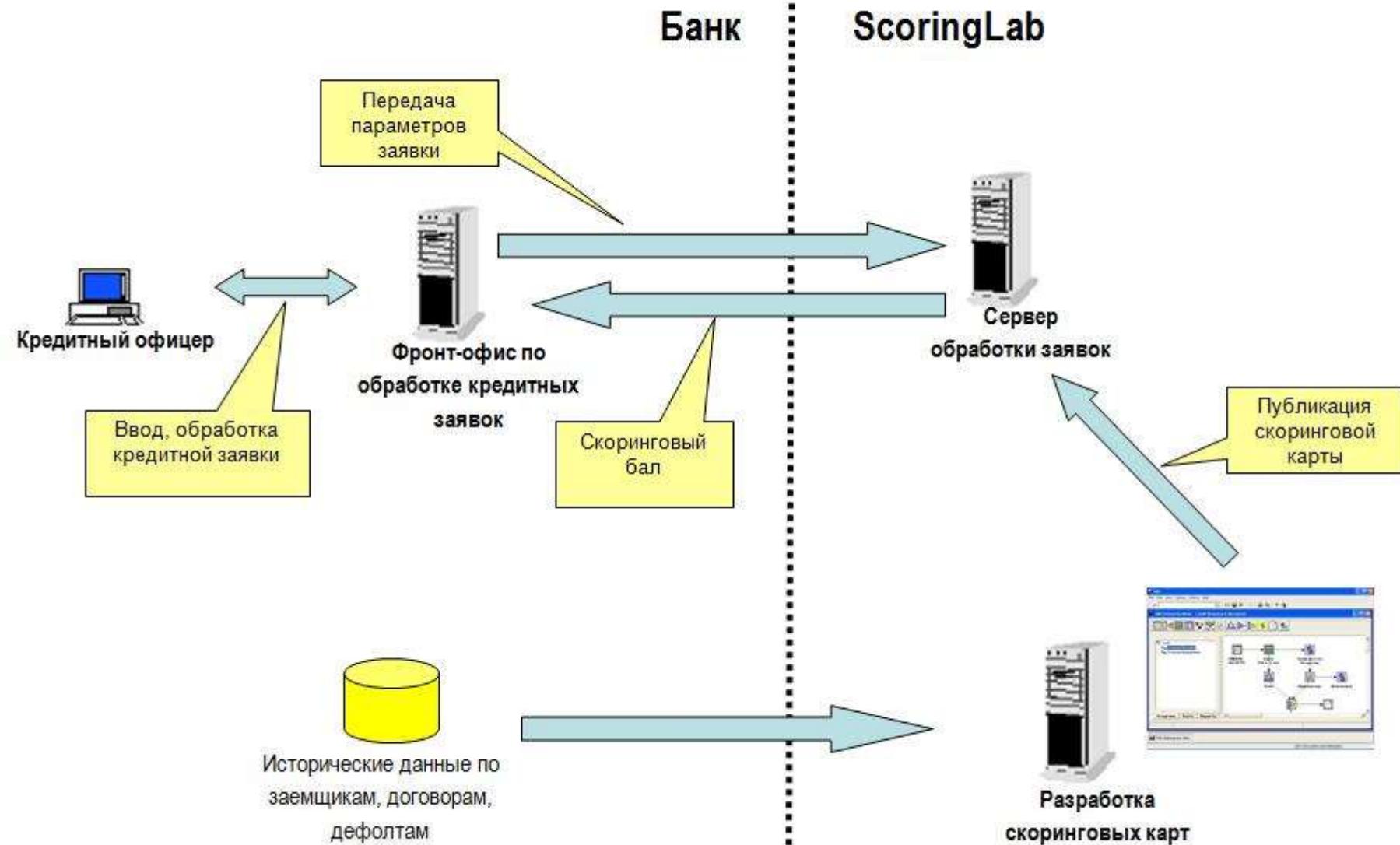
Визуализация результатов анализа

Решение SAS учитывает разнообразие:

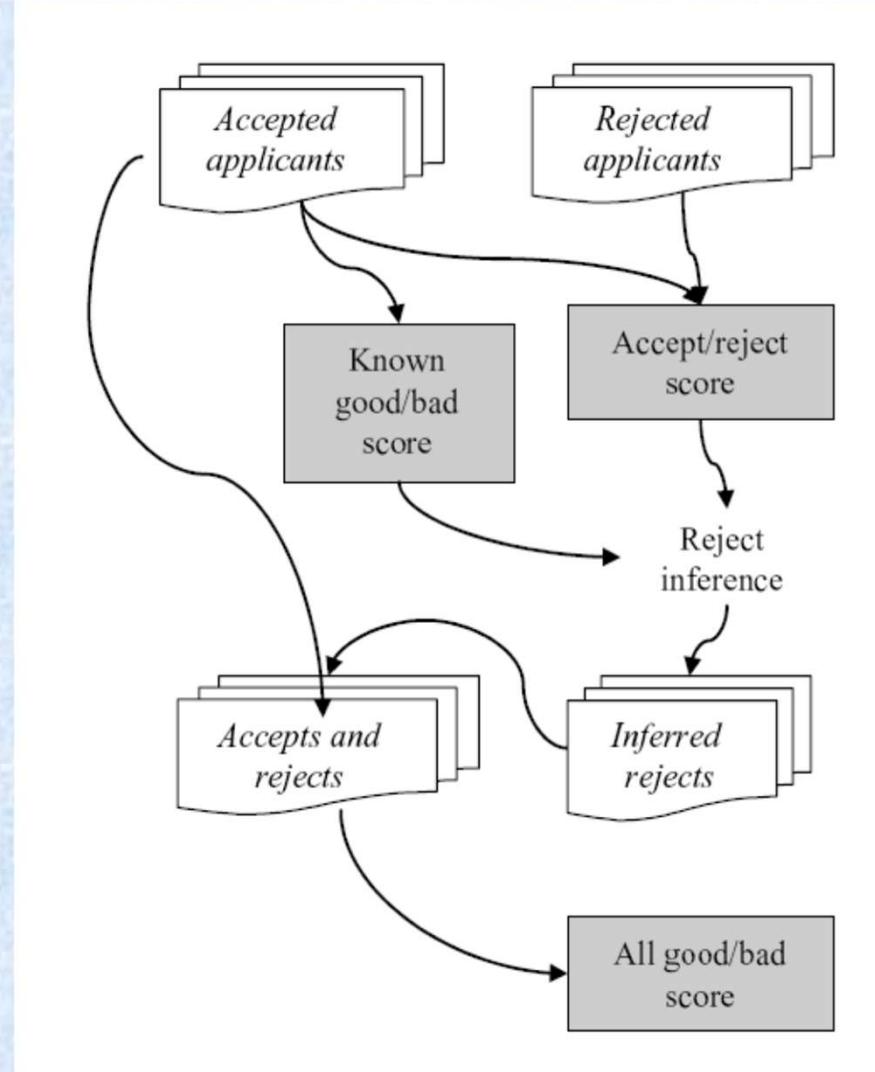
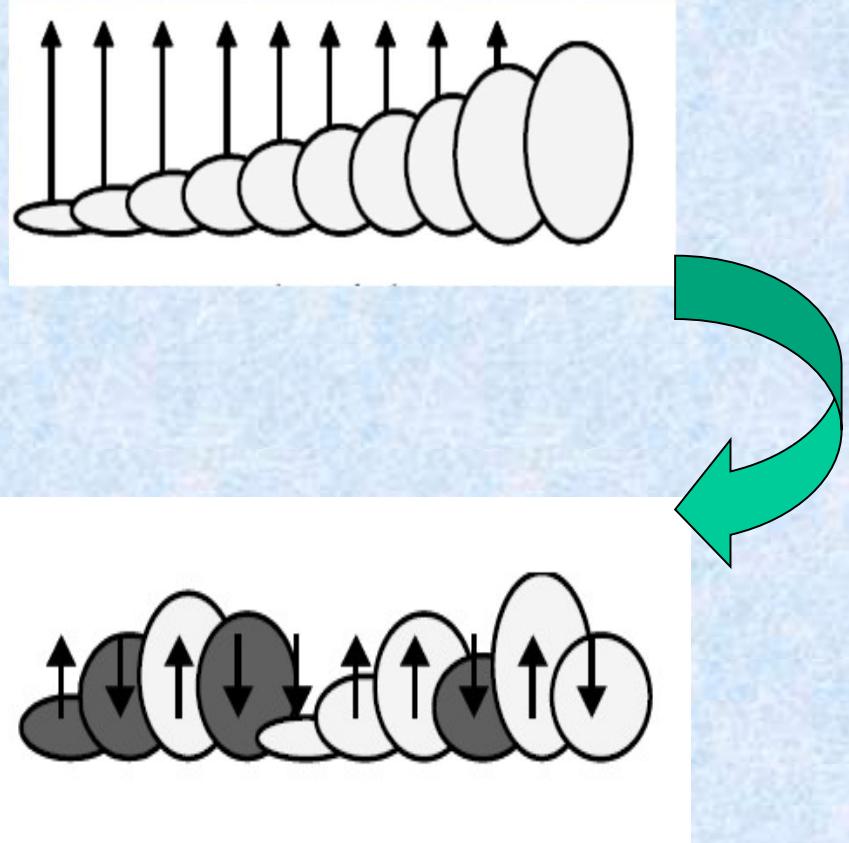
- **Кредитных продуктов**
 - *Ипотечное кредитование*
 - *Потребительский кредит*
 - *Кредитные банковские карточки*
 - ...
- **Задач кредитного scoringа**
 - *Application*
 - *Behavioral*
 - *Collection*
 - ...
- **Инструментов data-mining**
 - *Логистическая регрессия*
 - *Деревья решений*
 - *Нейронные сети*
 - ...

«Скоринга на аутсорсинге» — значительное сокращение временных затрат.

Вне зависимости от того, решит ли банк самостоятельно создавать скоринговую модель или поручит это специализированной компании



Влияние отвергнутых заявок



Калибровка скоринговой карты

- Выбор оптимального количества кластеров, которые дают лучший результат предсказания
- Calinski–Harabasz statistic
- Benchmark breakpoints
- Marginal risk boundaries

Calinski–Harabasz statistic

- Первый подход (Calinski–Harabasz) выбирает количество кластеров как значение аргумента, максимизирующую функцию CH(K),
$$CH(K) = (B(K)/(K - 1))/W(K)/(n - K),$$
где $B(K)$ и $W(K)$, соответственно, внешняя и внутренняя суммы квадратов элементов данных с K кластерами. Это один из самых первых предложенных методов. Он оказывается эффективным при данных небольших размерностей.

Разработка скоринговых карт

- Новая книга:
Сиддики Н. Скоринговые карты для оценки кредитных рисков / Наим Сиддики. – М.: Манн, Иванов и Фербер, 2014. – 268 с. Данная книга основана на материалах автора, который является идеологом построения системы скоринга в SAS: от разработки до внедрения. Содержит практические рекомендации по поэтапной разработке скоринговых карт от первого этапа запуска банком собственного проекта, сбора и очистки данных, настройки скоринговых моделей до рекомендаций по их внедрению, мониторингу и корректировке.
- Наим Сиддики — специалист по бизнес-решениям в области управления рисками, офис SAS® в Канаде. Более 10 лет специализируется на управлении кредитными рисками как в качестве консультанта, так и в качестве пользователя решений финансовых организаций. Наим Сиддики сыграл ключевую роль в разработке методологии SAS Credit Scoring и сегодня обеспечивает ее поддержку по всему миру.
С отличием закончил Имперский колледж науки, технологии и медицины Лондонского университета (имеет степень инженера-бакалавра), а также получил степень МВА в Йоркском университете в Торонто.



Этапы

- Этап 1. Подготовка и планирование
- Этап 2. Анализ данных и параметров проекта
- Этап 3. Создание базы данных для разработки модели
- Этап 4. Разработка скоринговой карты
- Этап 5. Управленческие отчеты по скоринговым картам
- Этап 6. Внедрение скоринговых карт

Этап 1. Подготовка и планирование

- Бизнес-план
- Определение орг целей и роли скоркарты
- Выбор между внутренней и внешней разработкой
- Определение типа скоркарты
- Определение риска проекта

Этап 2. Анализ данных и параметров проекта

- Анализ доступности и качества данных
- Определение параметров
- Определение периода и «окна выборки»
- Определение цели
- Исключения
- Сегментация
- Методология

Этап 3. Создание базы данных для разработки модели

- Формирование выборки
- Сбор и построение обучающей выборки
- Поправка на априорные вероятности

Этап 4. Разработка скоринговой карты

- Изучение данных
- Определение пропущенных значений и выбросов
- Выявление корреляции
- Анализ характеристик
- Создание предварительной скоринговой карты
- Анализ отклоненных заявок
- Создание финальной скоркарты
- Определение шкалы скоркарты
- Выбор скоринговой карты
- Контроль скоринговой карты

Этап 5. Управленческие отчеты по скоринговым картам

- Таблицы выигрыша (gains tables)
- Отчеты по характеристикам

Этап 6. Внедрение скоринговых карт

- Разработка стратегии
- Установка уровней отсечения
- Правила кредитных политик
- Управленческие отчеты

Скоринг в SAS enterprise miner



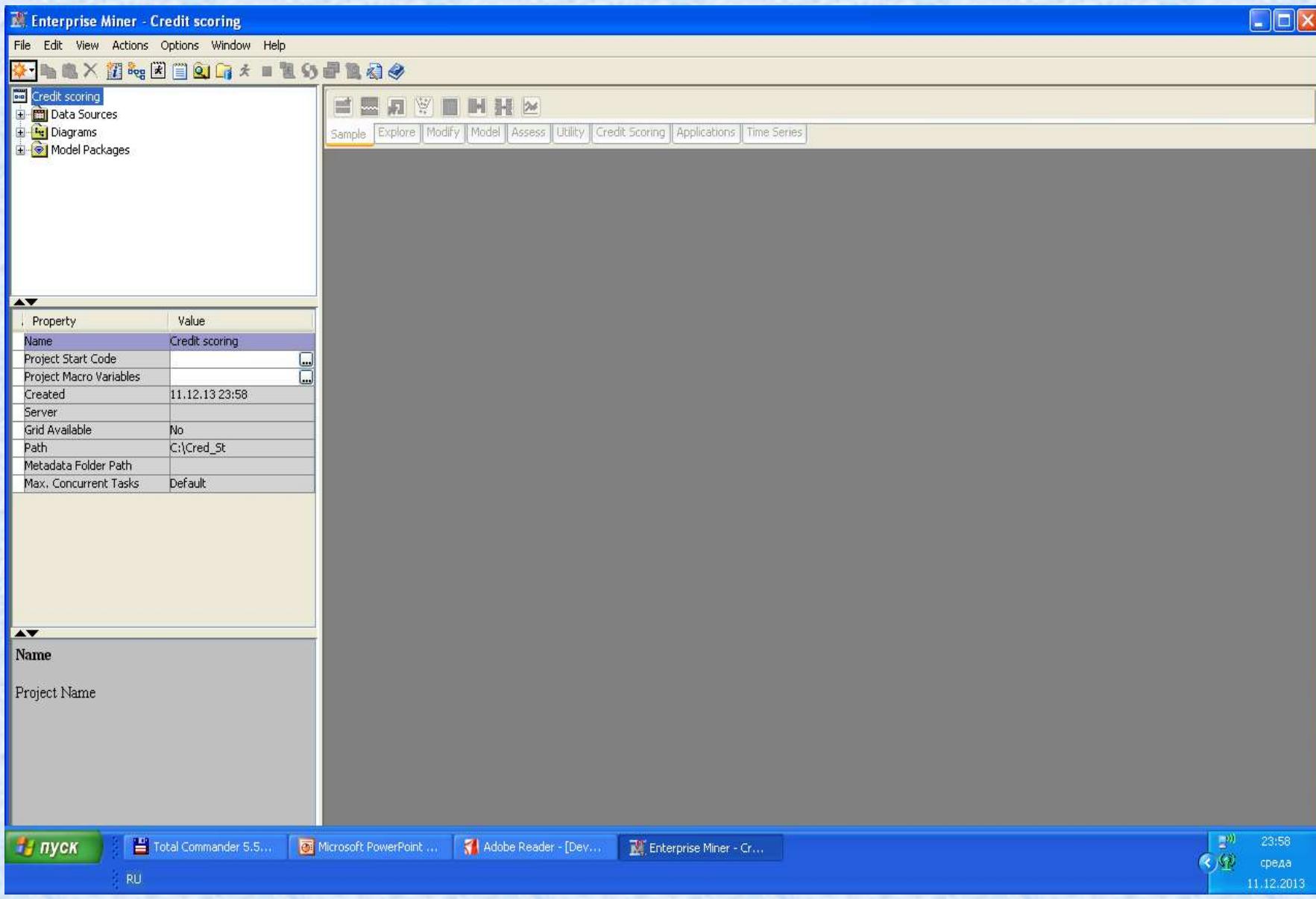
ОСНОВНЫЕ ЭТАПЫ

Требования к модели

- принятие решения о выдаче кредита
- определение уровня принятия решения
- расчет ожидаемых убытков
- ценообразование
- определение лимитов
- прогнозирование
- отчетность
- мониторинг
- индикаторы раннего предупреждения
- расчет резервов
- расчет капитала



Создание проекта + Ввод данных



Enterprise Miner - Credit scoring

File Edit View Actions Options Window Help

Credit scoring Data Sources Diagrams Model Packages

Sample Explore Modify Model Assess Utility Credit Scoring Applications Time Series

Select a SAS Table

SAS Libraries	Name	Engine	Path
Maps	Maps	V9	C:\Program Files\SASHome\...
Mapsgfk	Mapsgfk	V9	C:\Program Files\SASHome\...
Mapssas	Mapssas	V9	C:\Program Files\SASHome\...
Sampsio	Sampsio	V9	C:\Program Files\SASHome\...
Sashelp	Sashelp	V9	C:\Program Files\SASHome\...
Sasuser	Sasuser	V9	C:\Documents and Settings\...
Work	Work	V9	C:\DOCUME~1\161-AD~1\L...

Browse..

Get Details Refresh Properties... OK Cancel

Data Source W

Property Value

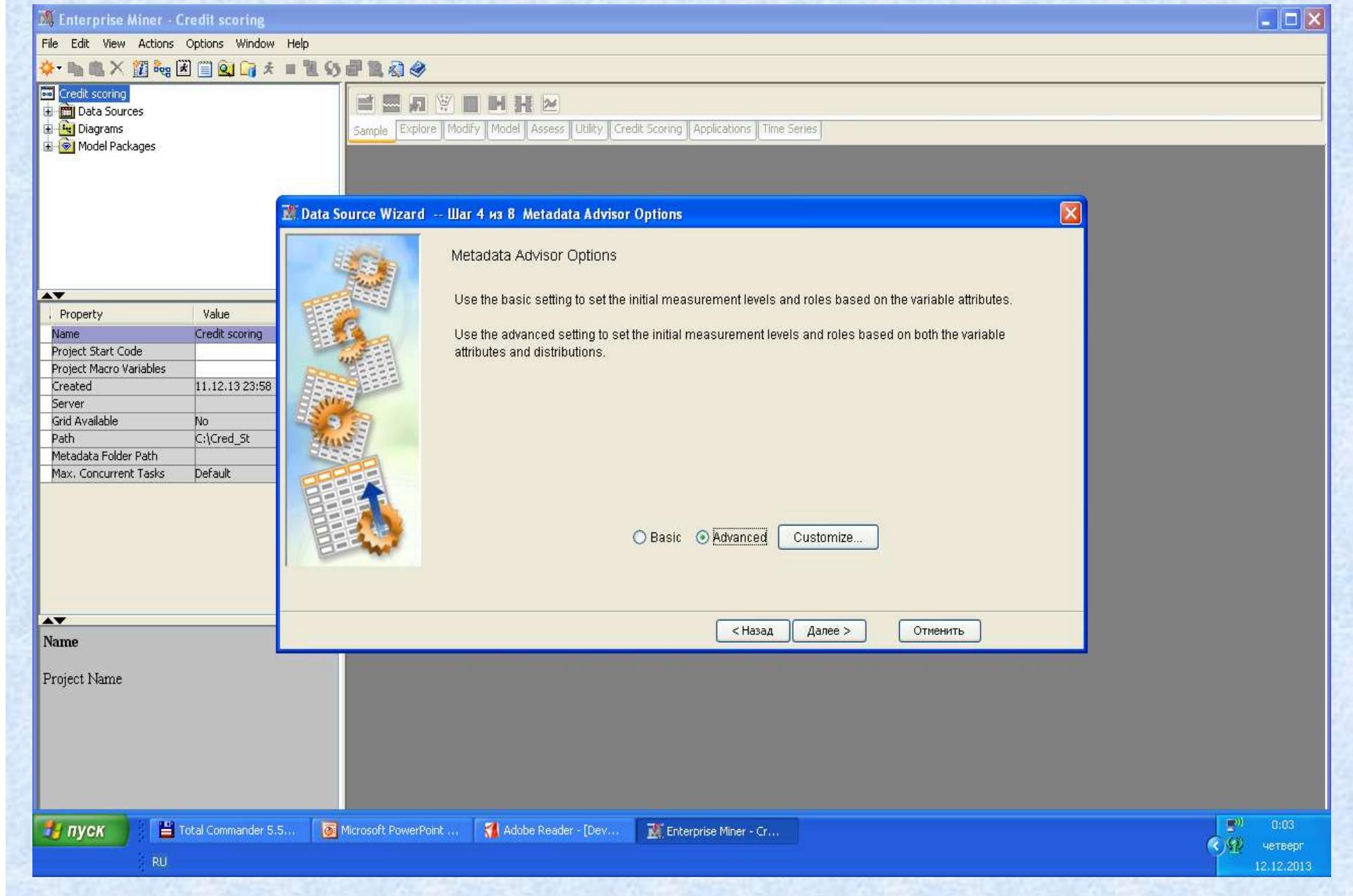
- Name Credit scoring
- Project Start Code
- Project Macro Variables
- Created 11.12.13 23:58
- Server
- Grid Available No
- Path C:\Cred_St
- Metadata Folder Path
- Max. Concurrent Tasks Default

Name

Project Name

0:01 четверг 12.12.2013

Назначение ролей переменным



Enterprise Miner - Credit scoring

File Edit View Actions Options Window Help

Credit scoring Data Sources Diagrams Model Packages

Sample Explore Modify Model Assess Utility Credit Scoring Applications Time Series

Data Source Wizard -- Шаг 5 из 8 Column Metadata

(нет) □ нет Равно ... Применить Сброс

Columns: □ Label □ Mining □ Basic □ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
AGE	Input	Interval	No		No	.	.
BUREAU	Input	Nominal	No		No	.	.
CAR	Input	Nominal	No		No	.	.
CARDS	Input	Nominal	No		No	.	.
CASH	Input	Interval	No		No	.	.
CHILDREN	Input	Nominal	No		No	.	.
DIV	Input	Binary	No		No	.	.
EC_CARD	Input	Binary	No		No	.	.
FINLOAN	Input	Binary	No		No	.	.
GB	Target	Binary	No		No	.	.
INC	Input	Nominal	No		No	.	.
INC1	Input	Nominal	No		No	.	.
INCOME	Input	Interval	No		No	.	.
LOANS	Input	Nominal	No		No	.	.
LOCATION	Input	Binary	No		No	.	.

Show code Explore Refresh Summary < Назад Далее > Отменить

Name
Project Name

пуск Total Commander 5.5... Microsoft PowerPoint... Adobe Reader - [Dev... Enterprise Miner - Cr... 0:06 четверг 12.12.2013 RU

Enterprise Miner - Credit scoring

File Edit View Actions Options Window Help

Credit scoring Data Sources Diagrams Model Packages

Sample Explore Modify Model Assess Utility Credit Scoring Applications Time Series

Data Source Wizard -- Шаг 6 из 10 Decision Configuration

Decision Processing

Do you want to build models based on the values of the decisions ?
If you answer yes, you may enter information about the cost or profit of each possible decision, prior probability and cost function. The data will be scanned for the distributions of the target variables.

No Yes

< Назад Далее > Отменить

Name
Project Name

пуск Total Commander 5.5... Microsoft PowerPoint ... Adobe Reader - [Dev... Enterprise Miner - Cr... 0:08 четверг RU 12.12.2013

Enterprise Miner - Credit scoring

File Edit View Actions Options Window Help

Credit scoring

Data Sources

CS_ACCEPTS

Diagrams

Model Packages

Sample Explore Modify Model Assess Utility Credit Scoring Applications Time Series

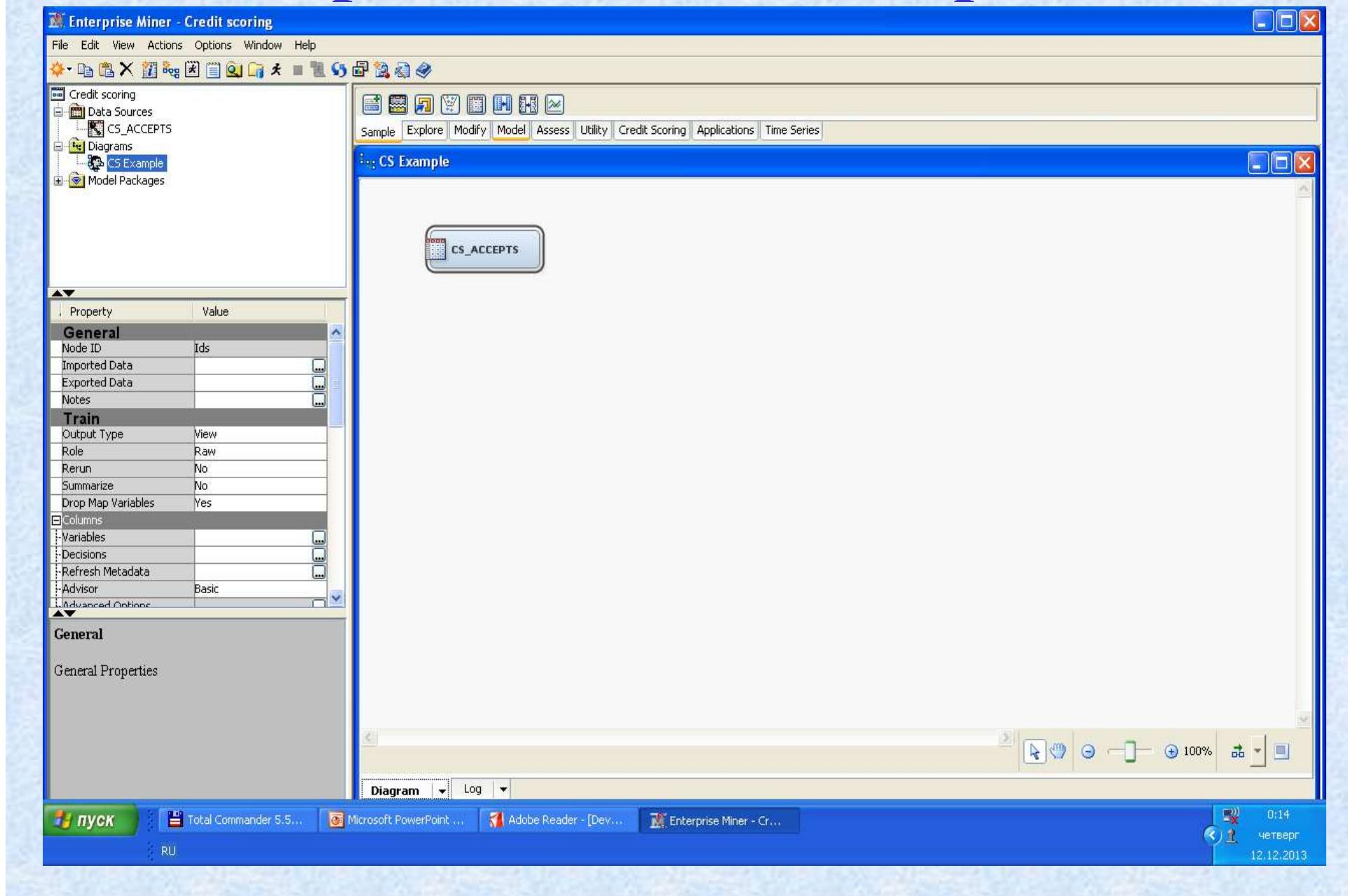
Property	Value
ID	csaccepts
Name	CS_ACCEPTS
Variables	[...]
Decisions	[...]
Role	Raw
Notes	[...]
Library	SAMPSIO
Table	CS_ACCEPTS
Sample Data Set	
Size Type	
Sample Size	
Type	DATA
No. Obs	3000
No. Cols	28
No. Bytes	1082368
Segment	
Created By	1A1-admin0

ID

Data Source identifier. The metadata tables associated with the data source are stored in the EMDS SAS library and use this identifier as the prefix for naming these tables.

пуск Total Commander 5.5... Microsoft PowerPoint... Adobe Reader - [Dev... Enterprise Miner - Cr... 0:10 четверг 12.12.2013 RU

Построение новой диаграммы



Лекция по кредитному评分ингу

Lec_Scor_R2014_s1 [Режим совместимости] - Microsoft PowerPoint

Главная Вставка Дизайн Анимация Показ слайдов Рецензирование Вид SAS

Макет Восстановить

Создать слайд

Вставить Буфер обмена

Предупреждение системы

Слайды Структура

70

71 Построение новой диаграммы

72

73 Разделение данных на тренировочные и тестовые

74 Анализ в группах для извлечения информации

Заметки к слайду

Слайд 72 из 116 | "Оформление по умолчанию" | русский | 68% | + | - | X

Explore - SAMPSON_CS_ACCEPTS

File View Actions Window

Sample Statistics

Obs #	Variable ...	Label	Type	Percent ...	Minimum	Maximum	Mean	Number o...	Mode Per...	Mode
1CAR	Type of Vehicle	CLASS	CLASS	0	.	.	.	3	76.65	CAR
2CARDS	Credit Cards	CLASS	CLASS	0	.	.	.	7	67.2	NO CREDI...
3NAT	Nationality	CLASS	CLASS	0	.	.	.	8	86.6	GERMAN
4PRODUCT	Type of Business	CLASS	CLASS	0.6	.	.	.	7	43.05	RADIO, TV, ...
5PROF	Profession	CLASS	CLASS	0.05	.	.	.	10	71.75	OTHERS
6RESID	Residence Type	CLASS	CLASS	16.1	.	.	.	3	79.55	LEASE
7STATUS	Status	CLASS	CLASS	0	.	.	.	6	51.5	V
8TITLE	Title	CLASS	CLASS	0	.	.	.	2	69.9	H
9AGE	Age	VAR	VAR	0	18	71	34.1175	.	.	.
10BUREAU	Credit Bureau Risk Class	VAR	VAR	0	1	3	1.829	.	.	.
11CASH	Requested cash	VAR	VAR	0	0	100000	2546.35	.	.	.
12CHILDREN	Num of Children	VAR	VAR	0	0	23	0.8025	.	.	.
13DIV	Large region	VAR	VAR	0	0	1	0.634	.	.	.
14EC_CARD	EC_card holders	VAR	VAR	0	0	1	0.301	.	.	.
15FINLOAN	Num finished Loans	VAR	VAR	0	0	1	0.3525	.	.	.
16GB	Good/Bad	VAR	VAR	0	0	1	0.5115	.	.	.
17INC	Salary	VAR	VAR	0	0	100000	25931.25	.	.	.
18INC1	Salary+ec_card	VAR	VAR	0	0	5	2.236	.	.	.
19INCOME	Income	VAR	VAR	0	0	10000	1767.35	.	.	.
20LOANS	Num of running loans	VAR	VAR	0	0	9	0.818	.	.	.
21LOCATION	Location of Credit Bureau	VAR	VAR	0	0	1	0.999	.	.	.
22NMBLOAN	Num Mybank Loans	VAR	VAR	0	0	2	0.1305	.	.	.
23PERS_H	Num in Household	VAR	VAR	0	1	25	2.3235	.	.	.
24REGN	Region	VAR	VAR	0	0	9	2.875	.	.	.
25TEL	Telephone	VAR	VAR	0	0	2	1.7725	.	.	.
26TMADD	Time at Address	VAR	VAR	0	0	999	127.8615	.	.	.
27TMJOB1	Time at Job	VAR	VAR	0	0	999	75.7005	.	.	.
28_freq_		VAR	VAR	0	1	30	15.1665	.	.	.

Выбор примеров

- **Historical** – все примеры
- **Validation** – используются для проверки правильности полученных результатов
- **Training** – используются для построения предиктивной модели

Разбиение данных и анализ

Enterprise Miner - Credit scoring

File Edit View Actions Options Window Help

Credit scoring
Data Sources CS_ACCEPTS
Diagrams CS Example
Model Packages

Sample Explore Modify Model Assess Utility Credit Scoring Applications Time Series

..: CS Example

CS_ACCEPTS → Data Partition

Property Value

General

Node ID	Part
Imported Data	[...]
Exported Data	[...]
Notes	[...]

Train

Variables	[...]
Output Type	Data
Partitioning Method	Default
Random Seed	12345

Data Set Allocations

Training	70.0
Validation	30.0
Test	0.0

Report

Interval Targets	Yes
Class Targets	Vec

Training

Specifies the allocation to the training data set.
The default value is 40 percent.

Diagram Log

0:17 четверг 12.12.2013

пуск Total Commander 5.5... Microsoft PowerPoint... Adobe Reader - [Dev... Enterprise Miner - Cr... RU

Анализ и группировка параметров

Enterprise Miner - Credit scoring

File Edit View Actions Options Window Help

Credit scoring
Data Sources
CS_ACCEPTS
Diagrams
CS Example
Model Packages

Sample Explore Modify Model Assess Utility Credit Scoring Applications Time Series

... CS Example

Property Value

Interval Variable Binning Op

- Apply Level Rule: No
- Binning Method: Quantile
- Number of Bins: 20

Special Code Options

- Use Special Codes: No
- Special Codes Data Set:

Grouping Options

- Interval Grouping Method: Monotonic Event Rate
- Ordinal Grouping Method: Monotonic Event Rate
- Tree Based Grouping Option:
- Constrained Optimal Option:
- Advanced Constrained Opt:
- Maximum Number of Groups: 10
- Significant Digits: 2
- Apply Restrictions: Yes

Maximum Number of Groups

Specifies the maximum number of groups to be generated.

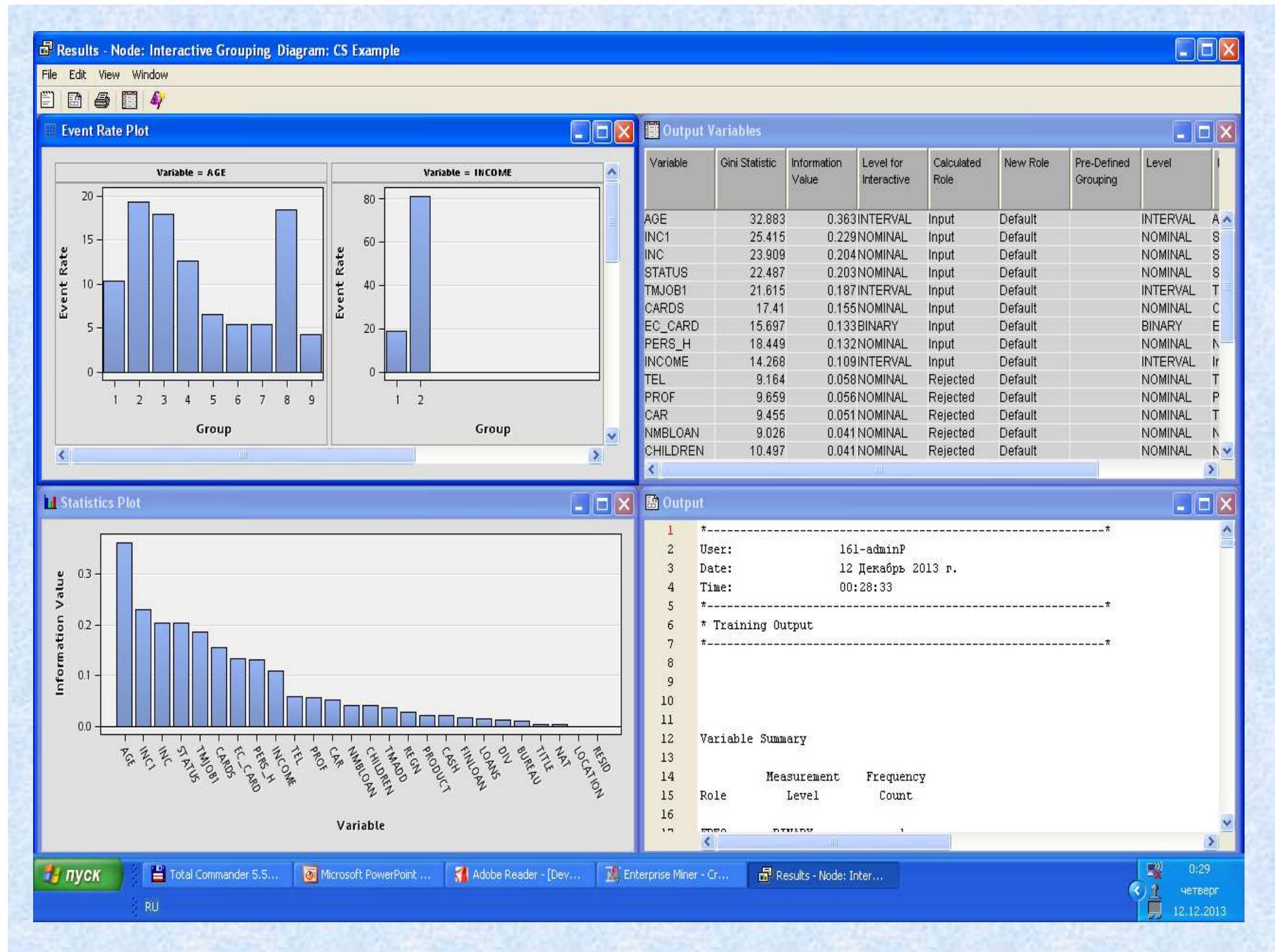
Diagram Log 100% RU 0:24 четверг 12.12.2013

Группировка параметров

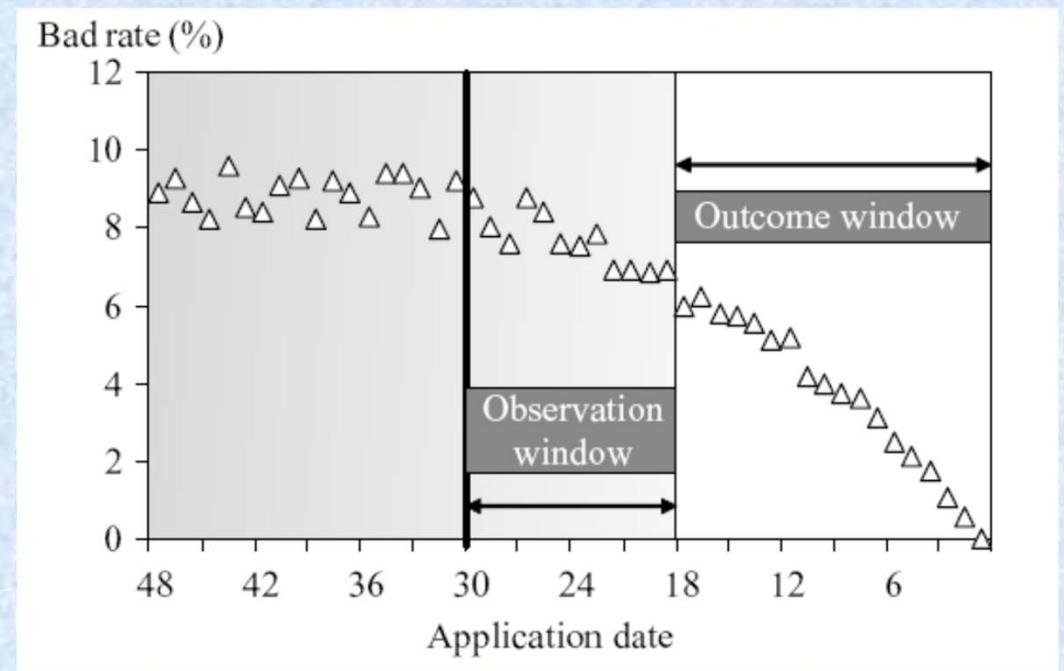
Узел позволяет в интерактивном режиме выбрать один из критериев отбора информативных признаков и сегментации признакового пространства:

- Критерий Джини
 - критерий Пирсона хи-квадрат
 - критерий минимума энтропии.

Допускается ручное задание сегментов.



Узел позволяет в интерактивном режиме выбрать один из критериев отбора информативных признаков и сегментации признакового пространства



Enterprise Miner - my_scor

File Edit View Actions Options Window Help

Interactive Grouping

Variable Selection
Selected Variable AGE

Variables Groupings

Value	Group	Cutoff	Event Count	Non Event Count	Total	Event Rate	WOE
MISSING_	8		0.0	0.0	0.0	0.0	0.0
AGE < 22	1	22.0	108.0	1050.0	1158.0	0.093	-1.12774
22 <= AGE < 24	2	24.0	141.0	1650.0	1791.0	0.079	-0.94238
24 <= AGE < 25	2	25.0	61.0	1200.0	1261.0	0.048	-0.42295
25 <= AGE < 27	3	27.0	126.0	1980.0	2106.0	0.06	-0.64758
27 <= AGE < 28	3	28.0	62.0	1140.0	1202.0	0.052	-0.4905
28 <= AGE < 29	4	29.0	32.0	1170.0	1202.0	0.027	0.19687
29 <= AGE < 31	4	31.0	100.0	2430.0	2530.0	0.04	-0.21167
31 <= AGE < 32	5	32.0	43.0	1230.0	1273.0	0.034	-0.04858
32 <= AGE < 33	5	33.0	26.0	1230.0	1256.0	0.021	0.45452
33 <= AGE < 35	6	35.0	56.0	2130.0	2186.0	0.026	0.23638
35 <= AGE < 37	7	37.0	33.0	1680.0	1713.0	0.019	0.52789
37 <= AGE < 38	7	38.0	23.0	1170.0	1193.0	0.019	0.52711
38 <= AGE < 39	8	39.0	24.0	960.0	984.0	0.024	0.28673
39 <= AGE < 42	8	42.0	61.0	2730.0	2791.0	0.022	0.39903
42 <= AGE < 44	8	44.0	26.0	1200.0	1226.0	0.021	0.42983

Variable Statistics
Original Gini 32,883
New Gini 32,883
Original Information Value 0,363
New Information Value 0,363

Detail Level
 Fine
 Coarse

Select Selected Variable: AGE

Reset All Changes Close

Monotonic Event Rate requests grouping to be

Run completed

Connected to 161-adminP as 161-adminP

1:40 среда 04.12.2013

ПУСК Total Commander 5.5... Enterprise Miner - my... Adobe Reader - [Dev... Results - Node: Inter... RU

- Вес факторов (WOE) измеряет способность каждого атрибута отделять “хороших” от “плохих”, показывает различия между частью “хороших” и частью “плохих”.
- Вес факторов (WOE) получается в результате расчета логарифма шансов:

$$\ln(\text{часть “хороших”} / \text{часть “плохих”})$$

- $< 0,02$ предсказательная сила отсутствует
- $0,02 - 0,1$ предсказательная сила мала
- $0,1 - 0,3$ предсказательная сила средняя
- $> 0,3$ предсказательная сила велика

Enterprise Miner - Credit_Scoring

File Edit View Actions Options Window Help

Credit_Scoring

CS_ACCEPTS
CS_REJECTS
CS Example
Model Packages

Variables - Scorecard

(none) not Equal to

Columns: Label Mining Basic Statistics

Name Use Report Role Level

Name	Use	Report	Role	Level
AGE	Default	No	Rejected	Interval
BUREAU	Default	No	Rejected	Nominal
CAR	Default	No	Rejected	Nominal
CARDS	Default	No	Rejected	Nominal
CASH	Default	No	Rejected	Interval
CHILDREN	Default	No	Rejected	Nominal
DIV	Default	No	Rejected	Binary
EC_CARD	Default	No	Rejected	Binary
FINLOAN	Default	No	Rejected	Binary
GB	Yes	No	Target	Binary
GRP_AGE	Default	No	Input	Ordinal
GRP_CARDS	Default	No	Input	Ordinal
GRP_EC_CARD	Default	No	Input	Ordinal
GRP_INC	Default	No	Input	Ordinal
GRP_INCI	Default	No	Input	Ordinal
GRP_INCOME	Default	No	Input	Ordinal
GRP_PERS_H	Default	No	Input	Ordinal
GRP_STATUS	Default	No	Input	Ordinal
GRP_TMJOB1	Default	No	Input	Ordinal
INC	Default	No	Rejected	Nominal
INC1	Default	No	Rejected	Nominal
INCOME	Default	No	Rejected	Interval
LOANS	Default	No	Rejected	Nominal
LOCATION	Default	No	Rejected	Binary
NAT	Default	No	Rejected	Nominal
NMBLOAN	Default	No	Rejected	Nominal
PERC_U	Default	No	Rejected	Nominal

Explore... Update Path OK Cancel

Diagram Log

Run completed Vlad_2 as Vlad_2 Connected to gateway

start Total Commander 6.5... Enterprise Miner - Cr... Adobe Acrobat Profe... Microsoft PowerPoint ... nero @SEARCH EN 0:28

Построение скоринговой карты

Enterprise Miner - Credit scoring

File Edit View Actions Options Window Help

Credit scoring
Data Sources
CS_ACCEPTS
Diagrams
CS Example
Model Packages

Sample Explore Modify Model Assess Utility Credit Scoring Applications Time Series

... CS Example

CS_ACCEPTS → Data Partition → Interactive Grouping → Scorecard

Scaling Options

Intercept Based Scorecard	No
Reverse Scorecard	No
Odds	50.0
Scorecard Points	200.0
Points to Double Odds	20.0
Scorecard Type	Summary
Precision	0
Bucketing Method	Min/Max Distribution
Number of Buckets	25
Use Indeterminate Values	No
Revenue Accepted Good	1000
Cost Accepted Bad	50000
Current Approval Rate	70.0
Current Event Rate	2.5
Generate Characteristic An/No	

General Properties

Edit Variables...
Update
Run
Create Model Package...
Results...
Export Path as SAS Program
Cut
Copy
Delete
Rename
Select All
Select Nodes
Connect Nodes
Disconnect Nodes

Diagram Log 100% RU 0:32 четверг 12.12.2013

```
graph LR; CS[CS_ACCEPTS] --> DP[Data Partition]; DP --> IG[Interactive Grouping]; IG --> SC[Scorecard]
```

Results - Node: Scorecard Diagram: CS Example

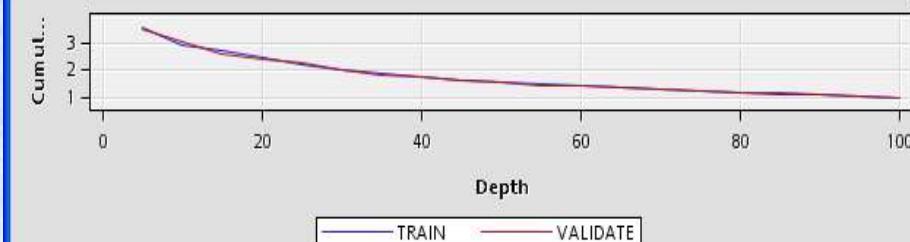


File Edit View Window



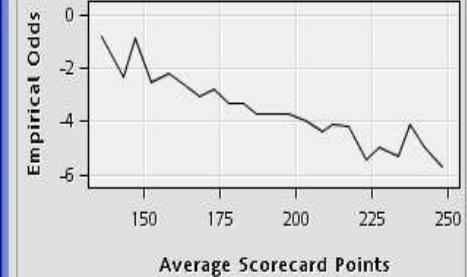
Score Rankings Overlay: Good/Bad

Cumulative Lift

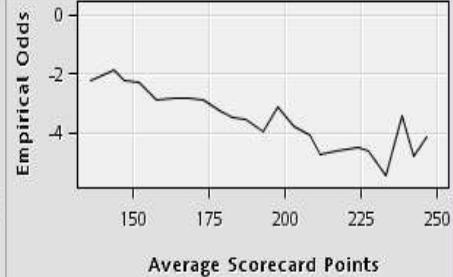


Empirical Odds Plot

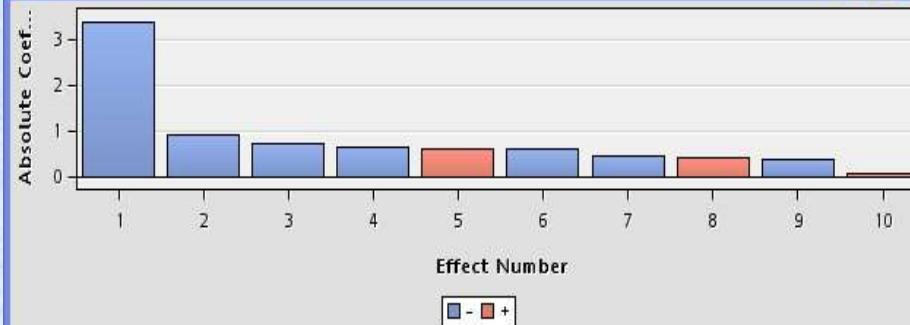
Data Role = TRAIN



Data Role = VALID



Effects Plot



Fit Statistics

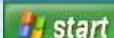
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
GB	Good/Bad	AIC_	Akaike's Inf...	8679.432	.	.
GB	Good/Bad	ASE_	Average Sq...	0.030479	0.030576	.
GB	Good/Bad	AVERR_	Average Err...	0.133021	0.133453	.
GB	Good/Bad	DFE_	Degrees of ...	32539	.	.
GB	Good/Bad	DFM_	Model Degr...	10	.	.
GB	Good/Bad	DFT_	Total Degr...	32549	.	.
GB	Good/Bad	DIV_	Divisor for A...	65098	27902	.
GB	Good/Bad	ERR_	Error Functi...	8659.432	3723.615	.
GB	Good/Bad	FPE_	Final Predic...	0.030498	.	.
GB	Good/Bad	MAX_	Maximum A	0.995281	0.996512	.

Scorecard

Scorecard	Group	Scorecard Points	Weight of Evidence	Event Rate GB = 1

Output

```
1 *-----*
2 User: Vlad_2
3 Date: 31, грудня 2013
4 Time: 00:33:41
5 *-----*
6 * Training Output
7 *-----*
```



Total Commander 6.5...

Enterprise Miner - Cr...

Results - Node: Score...

Adobe Acrobat Profes...

Microsoft PowerPoint ...

nero @SEARCH

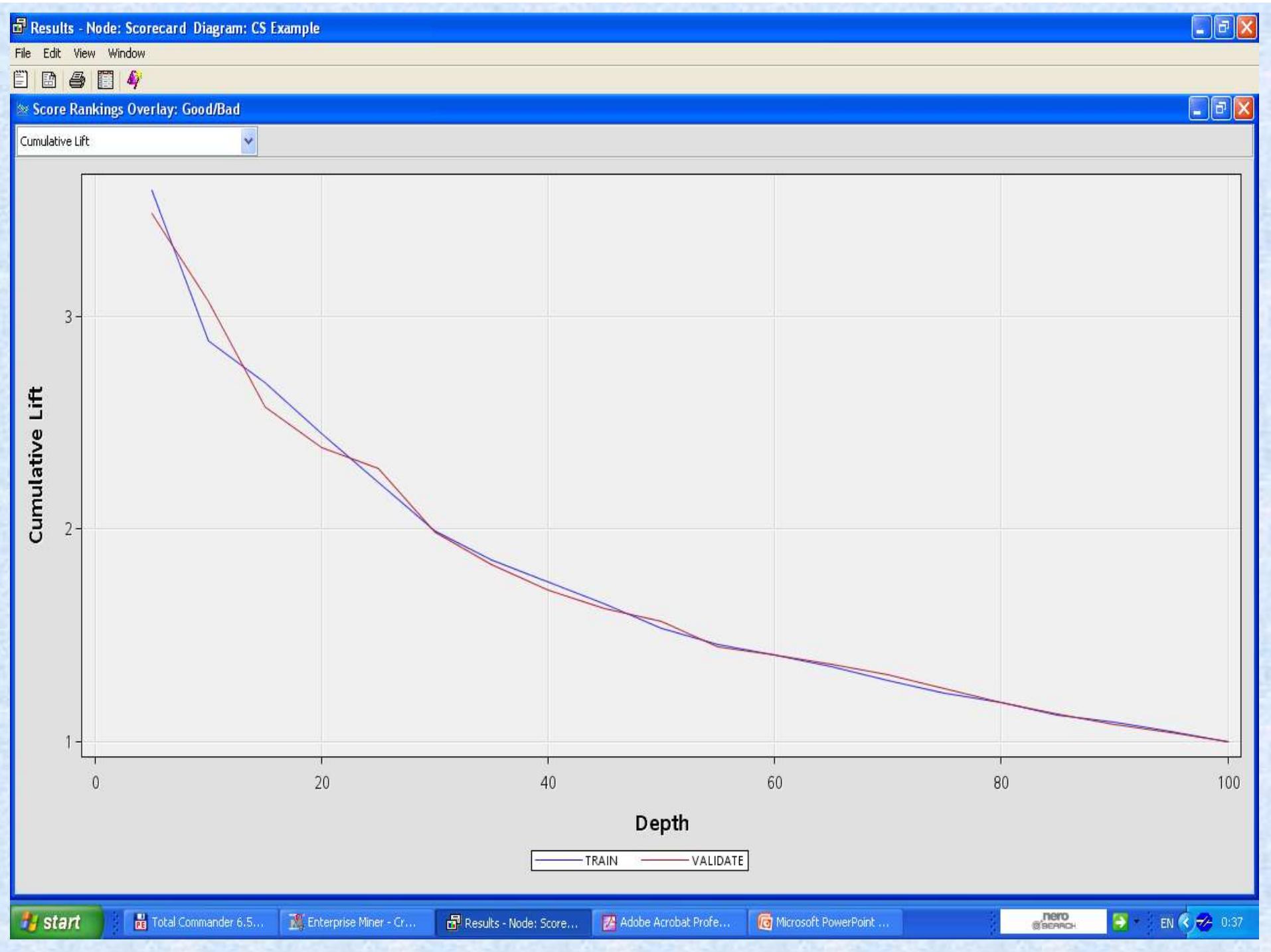


EN

0:34

Анализ скоринговой карты





Издержки ошибочной классификации (Lift)

- $L = (k/K)/(n/N)$ – коэффициент лифта

N – вся выборка, n – результат положителен, для каждого примера модель выдает вероятность его принадлежности к определенному классу.

Задача аналитика – выбор подмножества примеров, которые имеют наибольшую вероятность положительного исхода.

K – выборка для которой положительные результаты наиболее значительны.

- Максимальный лифт будет при объеме выборки = 1.
- При увеличении выборки в нее попадает все больше отрицательных исходов и лифт уменьшается.
- 100%, лифт стремиться к 1.

Results - Node: Scorecard Diagram: CS Example



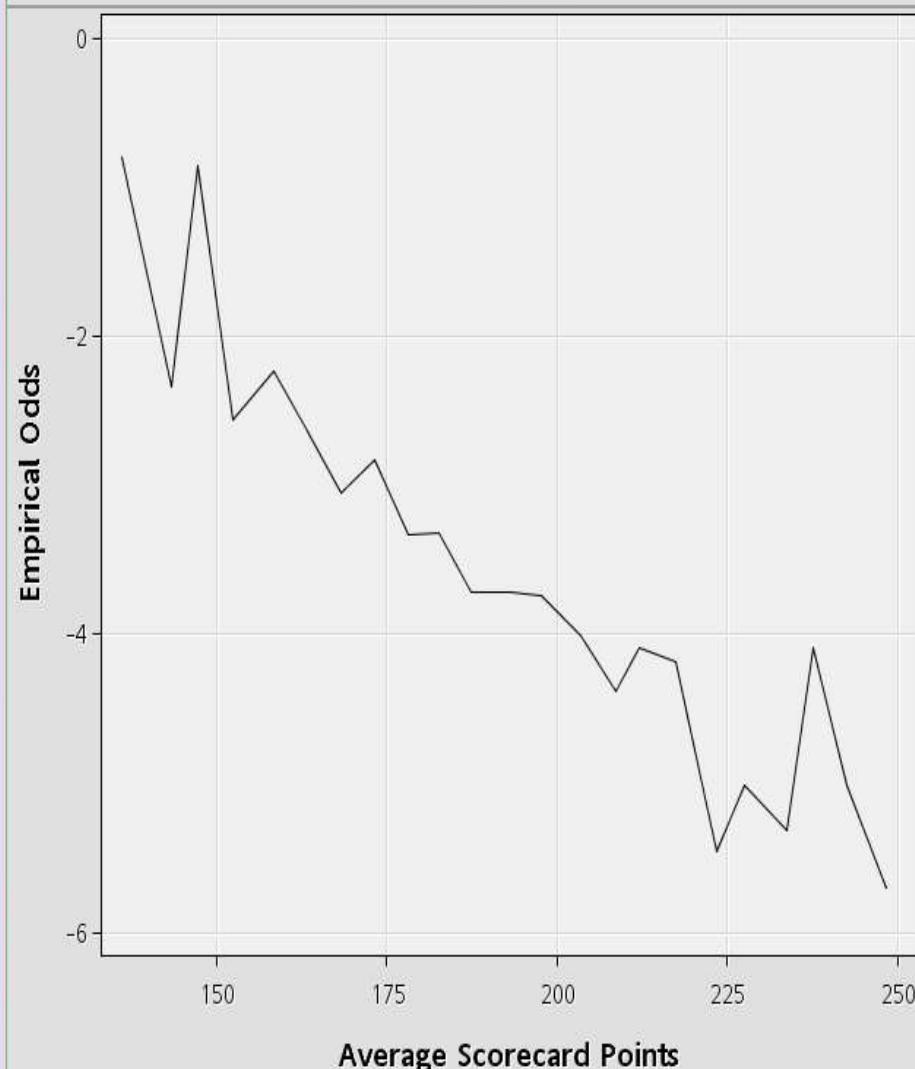
File Edit View Window



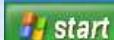
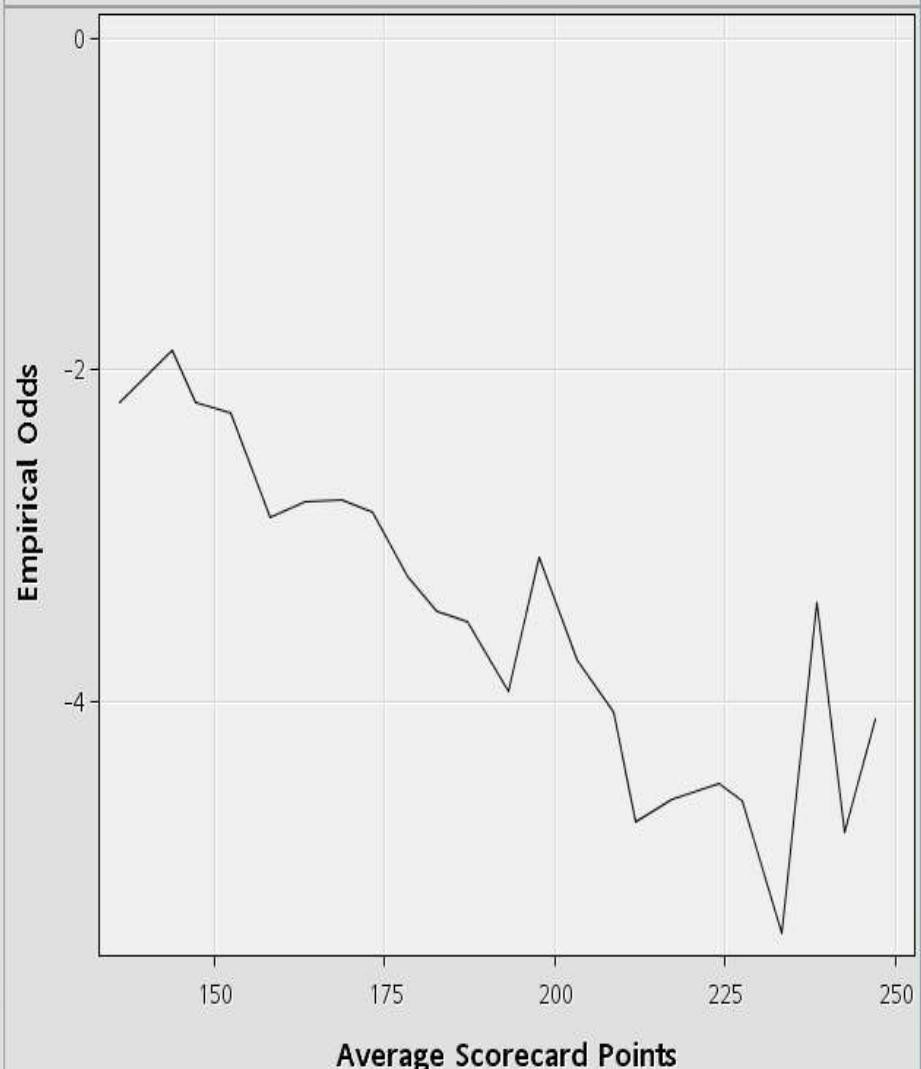
Empirical Odds Plot



Data Role = TRAIN



Data Role = VALID



Total Commander 6.5...

Enterprise Miner - Cr...

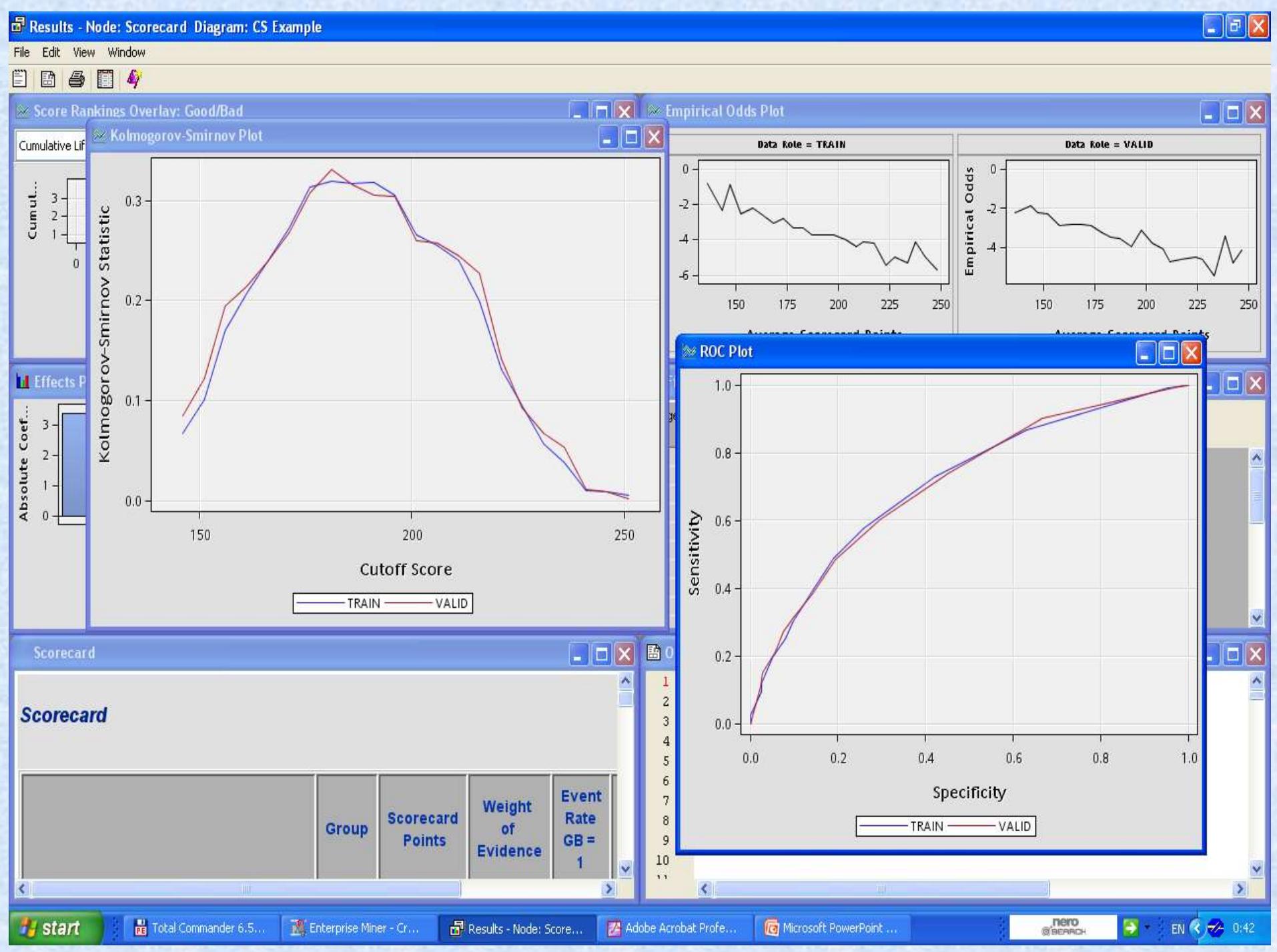
Results - Node: Score...

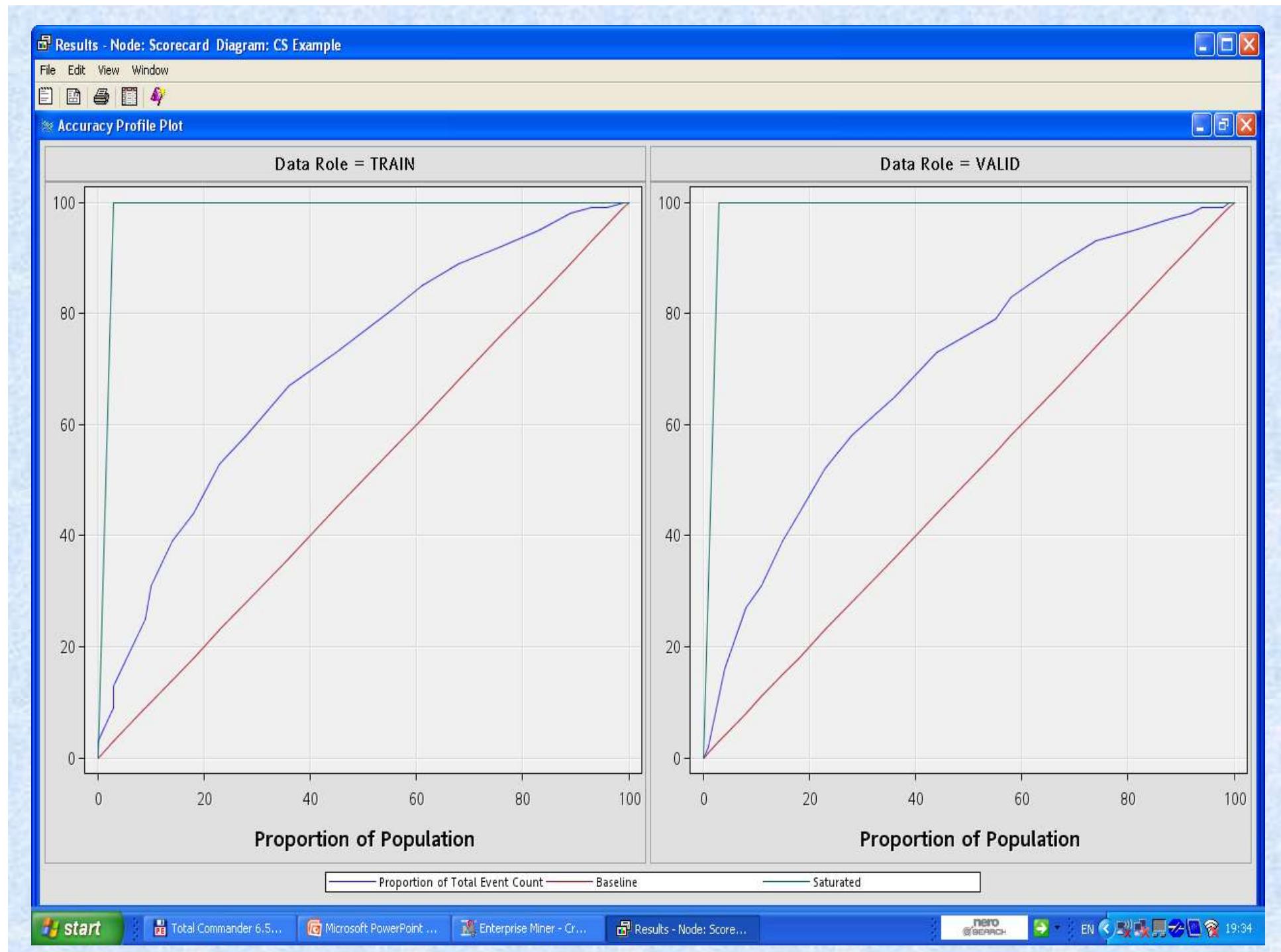
Adobe Acrobat Profes...

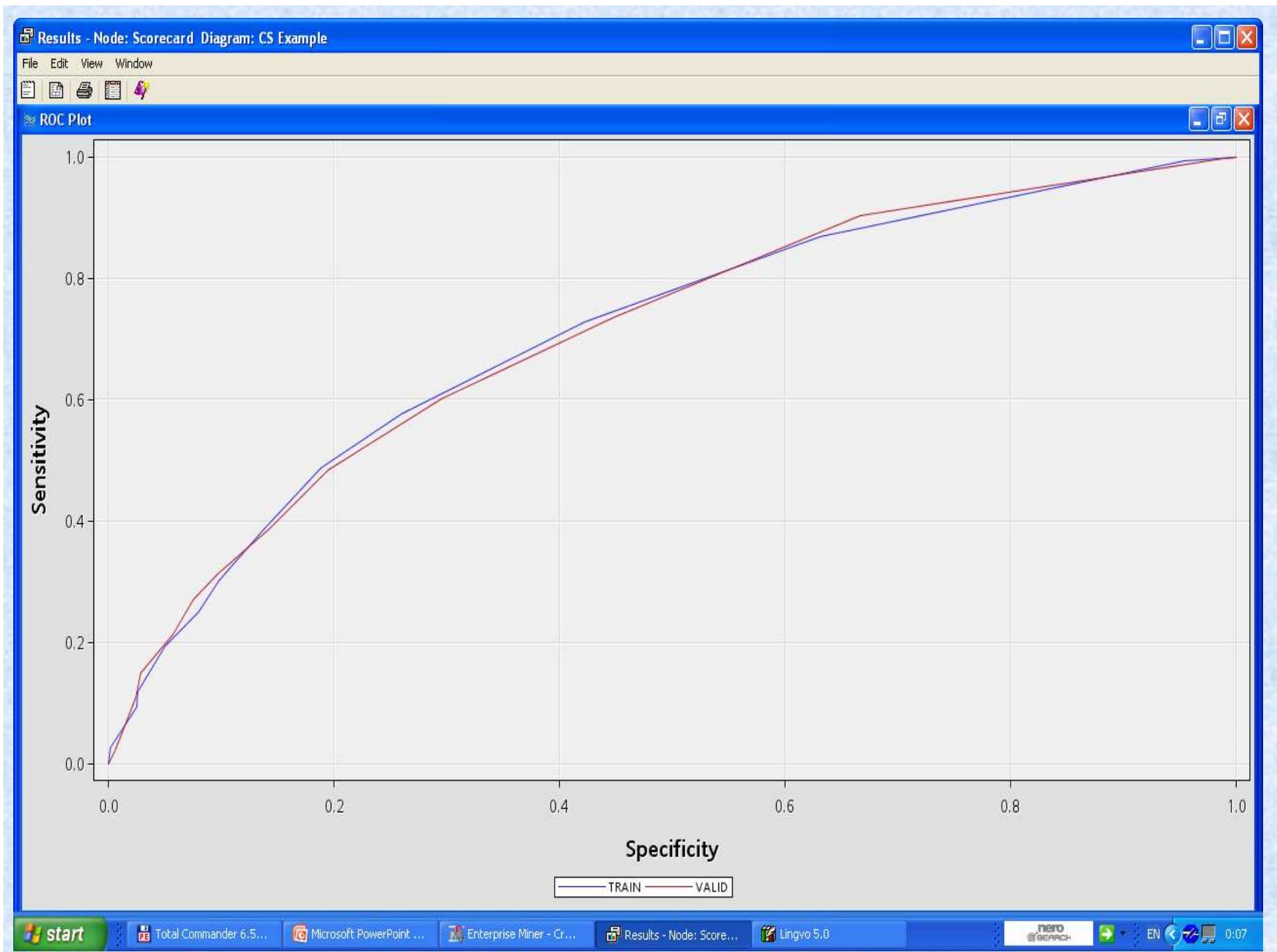
Microsoft PowerPoint ...

nero
@SEARCH

EN 0:39







Sensitivity

- Цель ROC анализа – выбрать точку отсечения (разделяющую два класса), которая обеспечит максимум чувствительности и специфичности.
- Специфичность – отношение числа истинноотрицательных наблюдений к числу фактически отрицательных наблюдений. $Sp = TN / (TN + FP)$
- Чувствительность - отношение числа истинноположительных наблюдений к числу фактически положительных наблюдений. $Se = TP / (TP + FN)$
- При уменьшении порога отсечения увеличивается вероятность ошибочного распознавания положительных наблюдений (ложноположительных), а при увеличении возрастает вероятность неправильного распознавания отрицательных наблюдений.
- ROC – receiver operating characteristic

Results - Node: Scorecard Diagram: CS Example

File Edit View Window

Score Rankings Overlay: Good/Bad

Cumulative Lift

Gains Table

Bucket	Score Bucket	Data Role	Count	Cumulative Count	Event Count	Non-Event Count	Cumulative Event Count
25 Score >= 256 TRAIN		TRAIN	0	0	0	0	0
24 251 <= Sco... TRAIN		TRAIN	0	0	0	0	0
23 248 <= Sco... TRAIN		TRAIN	150	150	0	150	150
22 241 <= Sco... TRAIN		TRAIN	151	301	1	150	150
21 236 <= Sco... TRAIN		TRAIN	122	423	2	120	120
20 231 <= Sco... TRAIN		TRAIN	1025	1448	5	1020	1020
19 226 <= Sco... TRAIN		TRAIN	755	2203	5	750	750
18 221 <= Sco... TRAIN		TRAIN	1416	3619	6	1410	1410
17 216 <= Sco... TRAIN		TRAIN	2071	5690	31	2040	2040
16 211 <= Sco... TRAIN		TRAIN	2013	7703	33	1980	1980
15 206 <= Sco... TRAIN		TRAIN	2673	10376	33	2640	2640
14 201 <= Sco... TRAIN		TRAIN	2474	12850	44	2430	2430
13 196 <= Sco... TRAIN		TRAIN	1505	14355	35	1470	1470
12 191 <= Sco... TRAIN		TRAIN	3595	17950	85	3510	3510
11 186 <= Sco... TRAIN		TRAIN	2827	20777	67	2780	2780
10 181 <= Sco... TRAIN		TRAIN	2580	23357	90	2490	2490
9 176 <= Sco... TRAIN		TRAIN	1647	25004	57	1590	1590
8 171 <= Sco... TRAIN		TRAIN	1716	26720	96	1620	1620
7 166 <= Sco... TRAIN		TRAIN	1225	27945	55	1170	1170
6 161 <= Sco... TRAIN		TRAIN	1191	29136	81	1110	1110
5 156 <= Sco... TRAIN		TRAIN	631	29787	61	570	570

Effects P

Absolute Coef

Scorecard

Scorecard

Empirical Odds Plot

Data Role = TRAIN

Odds

Average Scorecard Points

Data Role = VALID

Empirical Odds

Average Scorecard Points

Statistics

	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Good/Bad	_AIC_	Akaike's Infor...		8679.432		
Good/Bad	_ASE_	Average Squa...		0.030479	0.030576	
Good/Bad	_AVERR_	Average Error ...		0.133021	0.133453	
Good/Bad	_DFE_	Degrees of Fr...		32539		
Good/Bad	_DFM_	Model Degree...		10		
Good/Bad	_DFT_	Total Degree...		32549		
Good/Bad	_DIV_	Divisor for ASE		65098	27902	
Good/Bad	_ERR_	Error Function		8659.432	3723.615	
Good/Bad	_FPE_	Final Predicti...		0.030498		
Good/Bad	_MAX_	Maximum Abs...		0.995281	0.995512	
Good/Bad	_MSE_	Mean Square		0.030489	0.030576	

Output

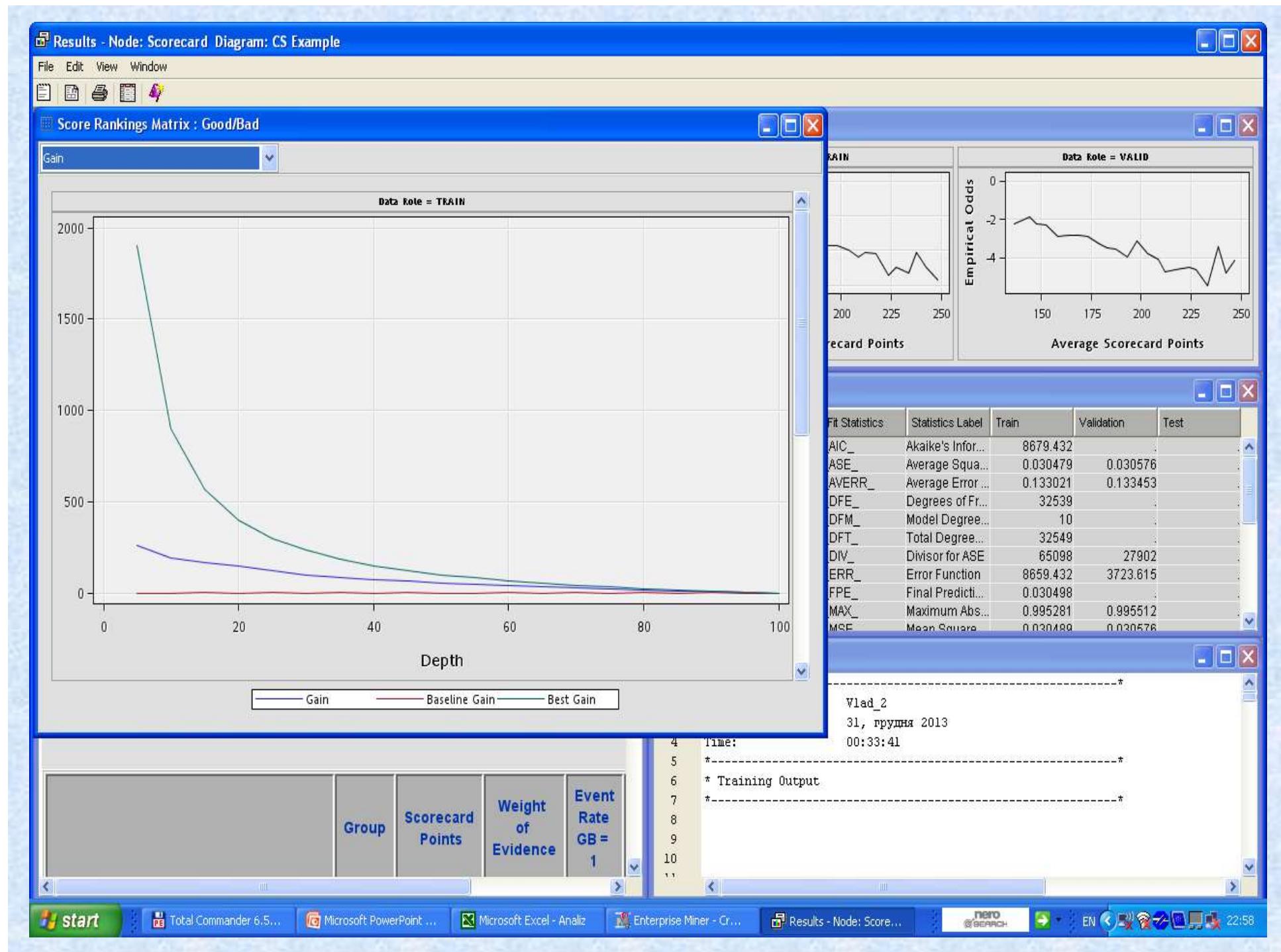
User: Vlad_2

Date: 31, грудня 2013

Time: 00:33:41

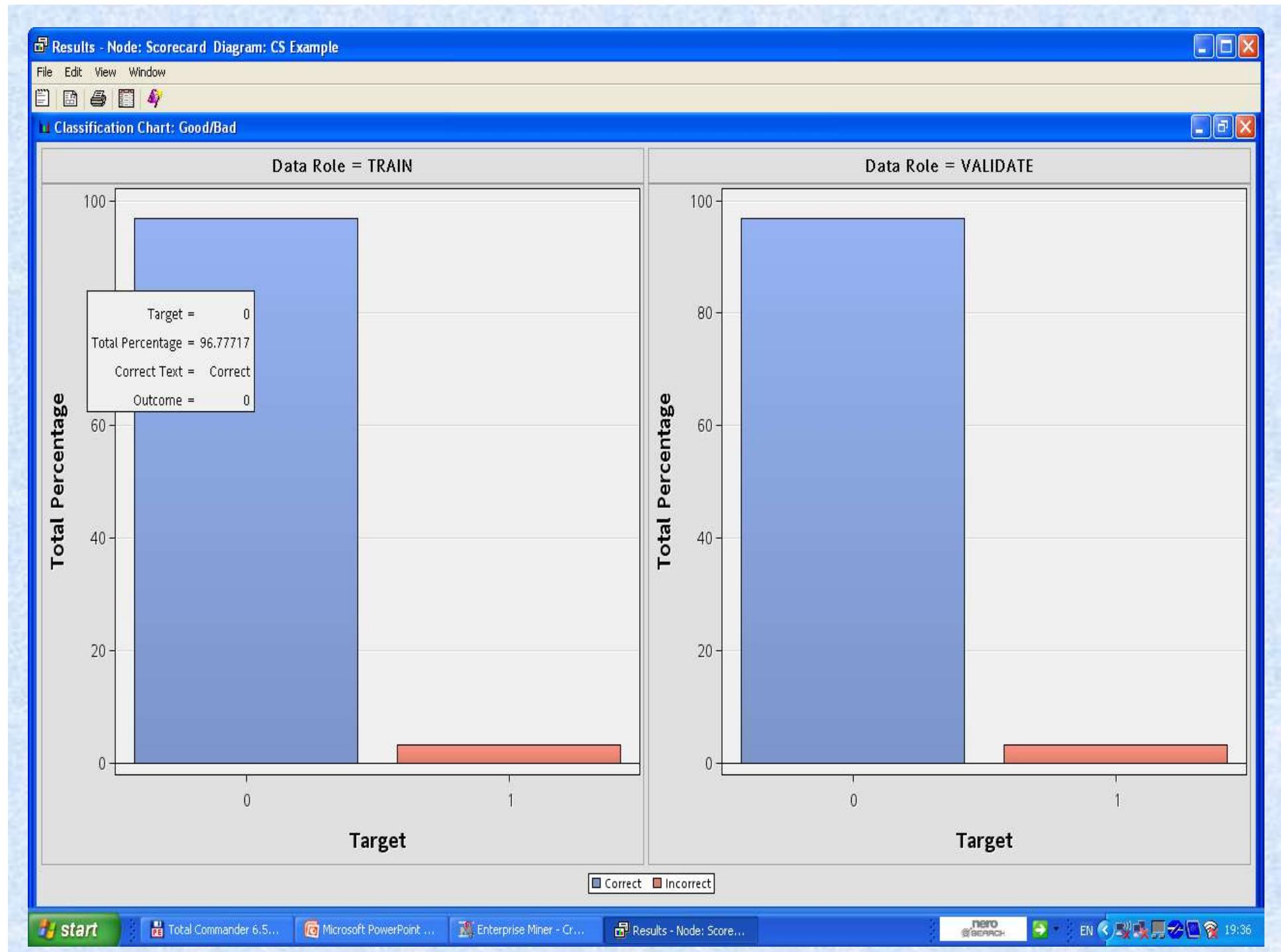
* Training Output

Start Total Commander 6.5... Enterprise Miner - Cr... Results - Node: Score... Microsoft PowerPoint ... nero @SEARCH EN



Gain – диаграмма = разновидность Lift – диаграммы (кумулятивная Lift -диаграмма

- По вертикальной – отношение числа истинноположительных наблюдений, попавших в выборку к числу всех наблюдений, классифицированных как положительные. $TP / (TP + FN)$
- По горизонтальной оси – размер выборки.
- Увеличивая размер выборки, мы увеличиваем количество ложноположительных наблюдений, риск ошибочной классификации растет.

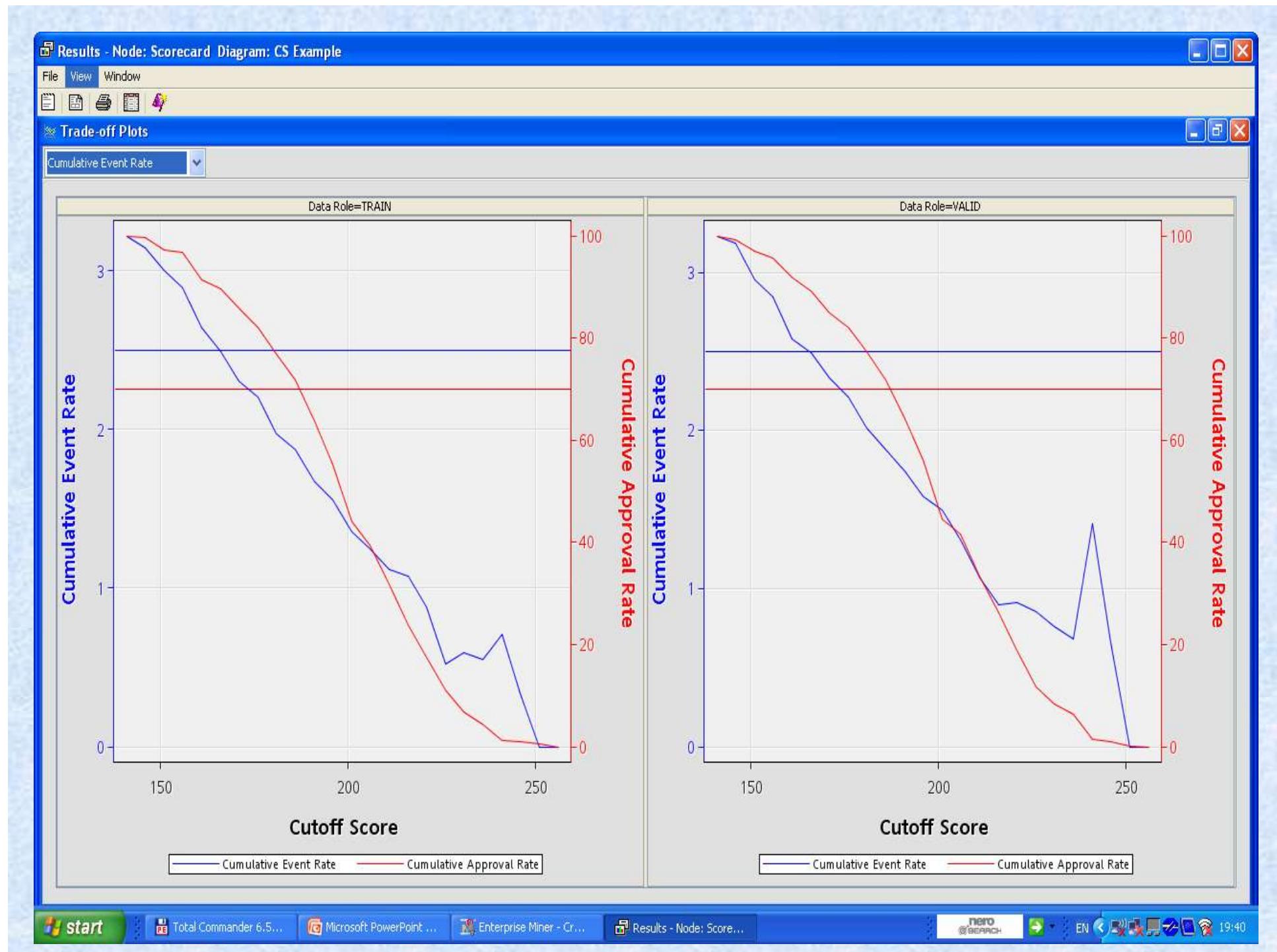


Таблицы сопряжения

- Таблицы сопряжения – это матрицы классификации: сколько хороших попали в хорошие и сколько в плохие, сколько плохих попали в плохие и сколько в хорошие.

Profit - кривые

- Lift – диаграммы – сравнение моделей и их оптимизация с точки зрения издержек классификации. Но не ясно в какие суммы обойдутся ошибочно выданные кредиты.
- Profit curve – кривая дохода. Чтобы узнать как бал связан с прибылью или убытками необходимо использовать числовые значения издержек и прибыли.



- Целью тестирования является оценка качества модели при её использовании на данных не входящих в выборку, которая была использована при построении модели. Можно выделить следующие основные тесты:
 - Эффективность (дифференцирующая способность)
 - Устойчивость (робастность)
 - Бэк-тестирование
 - Champion-Challenger стратегия
- Основным показателем эффективности является коэффициент Джини (Gini Statistic). Чем больше выборка, тем более надежным будет данный критерий.
- Для оценки надежности коэффициента Джини используют:
 - Бенчмаркинг
 - Аналитическую оценку уровня доверия
 - Эвристическую оценку уровня доверия путем повторных случайных выборок (Bootstrapping, Jackknifing, Метод скользящего среднего)
- Выборка для валидации
 - In-time: выборка построена случайным выбором из общей выборки
 - Out-of-time: случайный выбор из другого временного периода.

Сохранение скоринговой карты

Results - Node: Scorecard Diagram: CS Example

File Edit View Window

Score

Score

Properties

SAS Results

Scoring

Assessment

Model

Scorepoints Code

Score Distribution

Trade-off Plots

Empirical Odds Plot

Event Frequency Charts

Average Predicted Probability

Statistics Table

Gains Table

Scorecard

Effects Plot

Adverse Characteristics

Scorecard Points

	Scorecard Points
Age	-1
1	6
2	9
28<= AGE< 31	19
31<= AGE< 33	24
33<= AGE< 35	25
35<= AGE< 38	31
38<= AGE< 54, _MISSING_	31
54<= AGE	38
Credit Cards	38
CHEQUE CARD, MASTERCARD/EUROC, OTHER CREDIT CAR	38
AMERICAN EXPRESS, NO CREDIT CARDS, VISA MYBANK, VISA OTHERS, _MISSING_, _UNKNOWN_	14
EC_card holders	24
0.00, _MISSING_, _UNKNOWN_	24
1.00	10

0:44

четверг

12.12.2013

Модуль отчетности

- Целью тестирования является оценка качества модели при её использовании на данных не входящих в выборку, которая была использована при построении модели. Можно выделить следующие основные тесты:
 - Эффективность (дифференцирующая способность)
 - Устойчивость (робастность)
 - Бэк-тестирование
 - Champion-Challenger стратегия

- Существуют два совершенно разные вопросы, каждый из которых, тем не менее, можно упрощенно сформулировать в виде...: «какова величина кредитного риска данного займа?»
- А. Вопрос об ожидаемой частоте: каково ожидаемое значение частоты потерь по данному типу кредита?
- Б. X-Y-Z – вопрос: какова вероятность X того, что кредиты данного типа будут иметь частоту потерь Y за следующие Z лет?

Руководство по кредитному скорингу. Под ред. Элизабет Мэйз. – Минск, 2008. – 464 с.

Enterprise Miner - Credit_Scoring

File Edit View Actions Options Window Help

Credit Scoring Applications Time Series

CS Example

CS_ACCEPTS Data Partition Interactive Grouping Scorecard Credit Exchange

Property Value

General

Node ID	CreditEx
Imported Data	[...]
Exported Data	[...]
Notes	[...]

Train

Variables	[...]
-----------	-------

Status

Create Time	31.12.13 0:46
Run ID	915d470a-061b-487b-bfeb-6
Last Error	
Last Status	Complete
Last Run Time	31.12.13 0:47
Run Duration	0 Hr. 0 Min. 9,02 Sec.
Grid Host	
User-Added Node	No

General Properties

Diagram Log 100% 100%

Run completed Vlad_2 as Vlad_2 Connected to gateway

start Total Commander 6.5... Enterprise Miner - Cr... Microsoft PowerPoint ...

nero @SEARCH EN 0:54

The screenshot shows the SAS Enterprise Miner interface with the 'Credit_Scoring' project open. The left sidebar displays the project tree with nodes like 'Data Sources' (containing 'CS_ACCEPTS' and 'CS_REJECTS'), 'Diagrams' (containing 'CS Example'), and 'Model Packages'. The main workspace is titled 'CS Example' and contains a flow diagram with five nodes connected sequentially: 'CS_ACCEPTS' (input), 'Data Partition', 'Interactive Grouping', 'Scorecard', and 'Credit Exchange' (output). Each node has a green checkmark indicating successful execution. Below the workspace is a properties panel showing node details such as 'Node ID' (CreditEx), 'Create Time' (31.12.13 0:46), and 'Last Status' (Complete). The bottom status bar indicates the run completed and the user is connected to a gateway named 'Vlad_2 as Vlad_2'.

Модуль отчетности Credit Scoring Solution

Позволяет найти скоринг-балл, при котором ожидаемые потери по невозвратам кредитов в группе заемщиков с этим баллом уравновесятся процентным доходам по возвращенным кредитам (точка безубыточности).

Однако, значение точки безубыточности еще не позволяет определить балл отсечения клиентов.

Для оптимального балла отсечения математическое ожидание доходности всего кредитного портфеля должно быть максимальным.

Оценить оптимальный балл отсечения позволяет исследование зависимости средней доходности по заемщику совокупного кредитного портфеля от балла отсечения.

Модуль отчетности Credit Scoring Solution

- Для принятия решения относительно балла отсечения также полезна зависимость отношения вероятности возврата кредита к вероятности дефолта от скоринг-балла.
- Модуль отчетности легко интегрируется в информационную инфраструктуру банка (фронт-офис, СУБД, хранилище данных).

- Для проверки качества модели и её предикативной (прогнозной) силы используются стандартные статистические коэффициенты:
- Статистика Колмогорова-Смирнова;
- Площадь под ROC-кривой;
- Коэффициент Gini;

Статистика Колмогорова-Смирнова

- Ск разрабатываются для ранжирования заемщиков по шансам наступления определенного события. Важно, что в ск кредиты с которыми происходит данное событие и кредиты, с которыми оно не происходит, имели разные баллы.

Статистика Колмогорова – Смирнова используется для оценки способности скоринговой карты ранжировать заемщиков

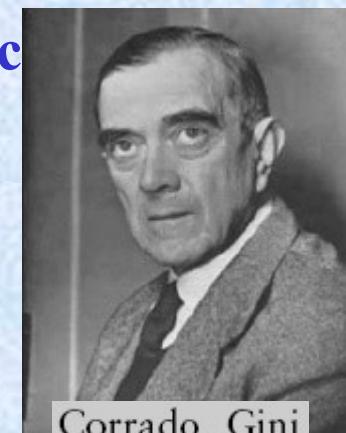
- Статистика Джини (Манна-Уитни)
- Индекс Джини является интегральной характеристикой, позволяющей судить о прогностической силе скоринговой карты.
- Если множество Т содержит элементы из n классов, $gini(T)$ определяется

$$gini(T) = 1 - \sum p_j^2$$

где p_j относительная частота класса j во множестве А
После разбиения Т на два подмножества

Т1 и Т2 индек

$$gini(T)_{split} = \frac{|T_1|}{|T|} gini(T_1) + \frac{|T_2|}{|T|} gini(T_2)$$



Corrado Gini

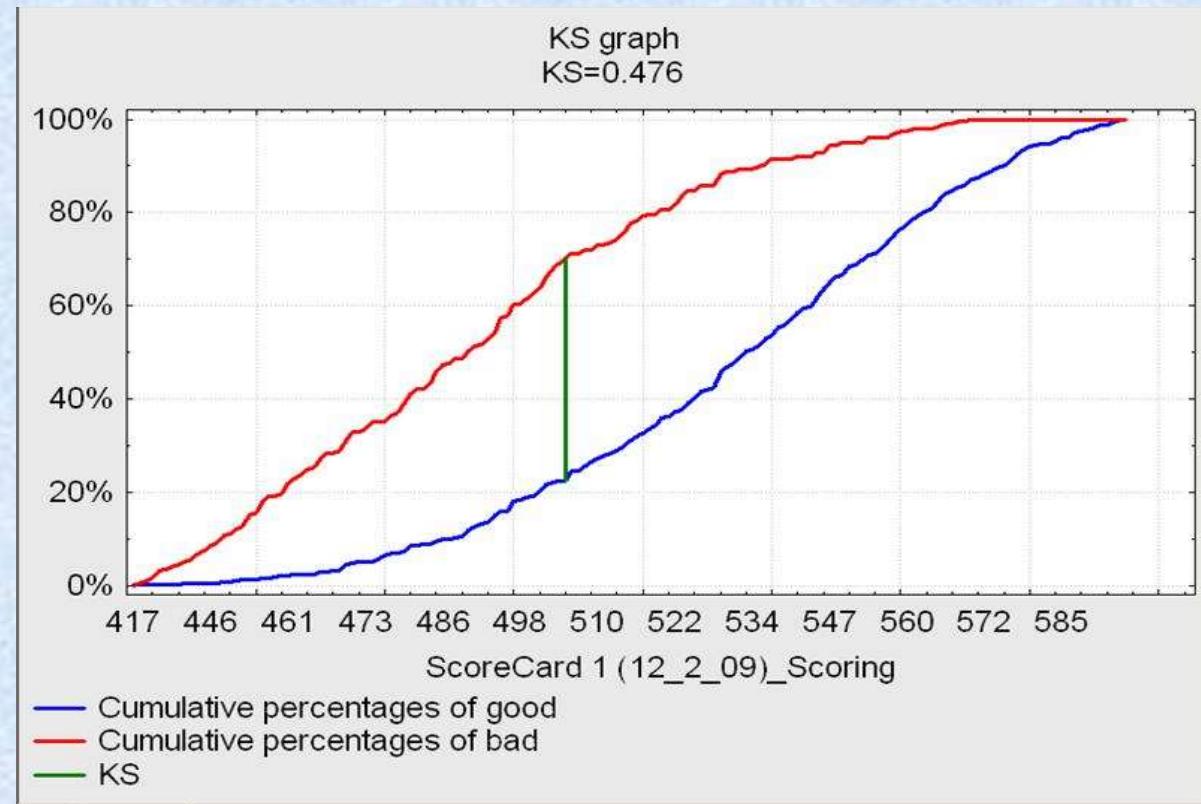
Gini Index

- Может быть использован для сравнения распределения признака (дохода) по разным группам населения (например, коэффициент Джини для сельского населения и коэффициент Джини для городского населения).
- Позволяет отслеживать динамику неравномерности распределения признака (дохода) в совокупности на разных этапах.

U-критерий Манна — Уитни

- (*Mann — Whitney U-test*) — статистический критерий, используемый для оценки различий между двумя независимыми выборками по уровню какого-либо признака, измеренного количественно. Позволяет выявлять различия в значении параметра между малыми выборками.
- **Критерий согласия Колмогорова** предназначен для проверки гипотезы о принадлежности выборки некоторому закону распределения, то есть проверки того, что эмпирическое распределение соответствует предполагаемой модели.

Статистика Колмогорова-Смирнова (K-S) – это максимальное расстояние между функциями распределения по баллу хороших и плохих.



KC < 20 - карта непригодна

KC от 20 до 40 : нормально

KC от 41 до 50 : хорошая

KC от 51 до 60 : очень хорошая

KC от 61 до 75 : супер

KC от 75 : ?

Адекватность модели

- Адекватность модели — совпадение свойств (функций/параметров/характеристик и т. п.) модели и соответствующих свойств моделируемого объекта. Адекватностью называется совпадение модели моделируемой системы в отношении цели моделирования.
- Оценка адекватности модели — проверка соответствия модели реальной системе. Оценка адекватности модели реальному объекту оценивается по близости результатов расчетов экспериментальным данным.

Два основных подхода к оценке адекватности:

1) по средним значениям откликов модели и системы

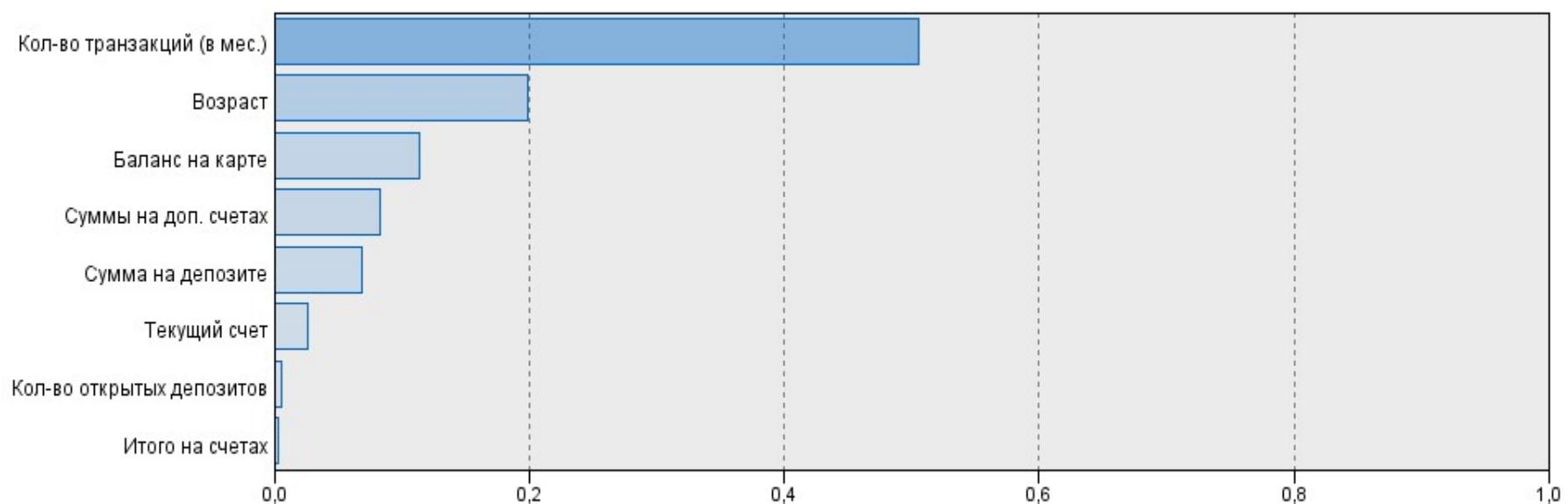
Проверяется гипотеза о близости средних значений каждой n -й компоненты откликов модели Y_n известным средним значениям n -й компоненты откликов реальной систем.

2) по дисперсиям отклонений откликов модели от среднего значения откликов систем

Сравнение дисперсии проводят с помощью критерия F (проверяют гипотезы о согласованности), с помощью критерия согласия (при больших выборках, $n > 100$), критерия Колмогорова-Смирнова (при малых выборках, известны средняя и дисперсия совокупности), Кохрена и др.

Важность предиктора

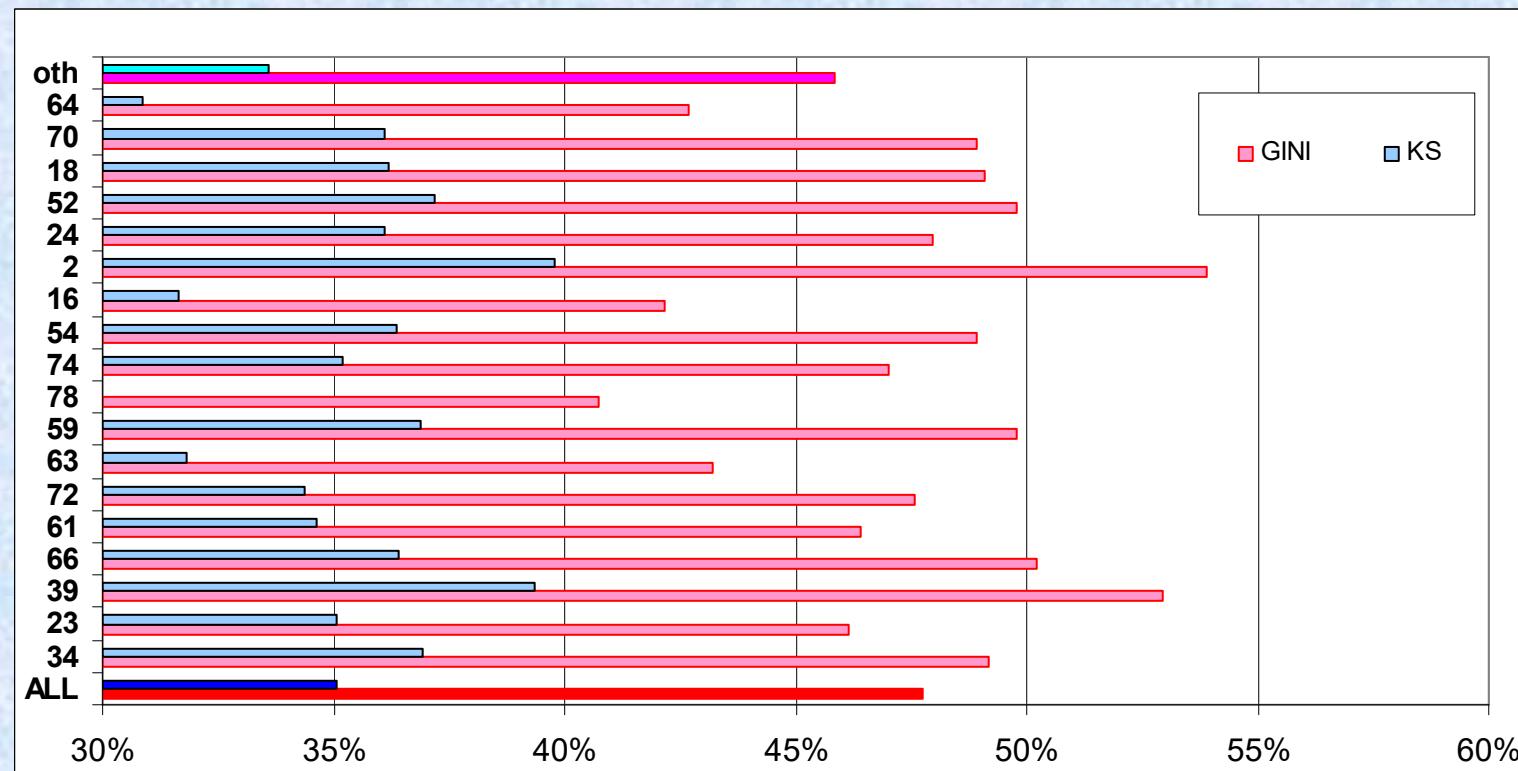
Целевое поле: Ипотека



Наименее важный

Наиболее важный

Статистики Колмогорова-Смирнова и Джини считаются лучшими



Проблемы построения карты

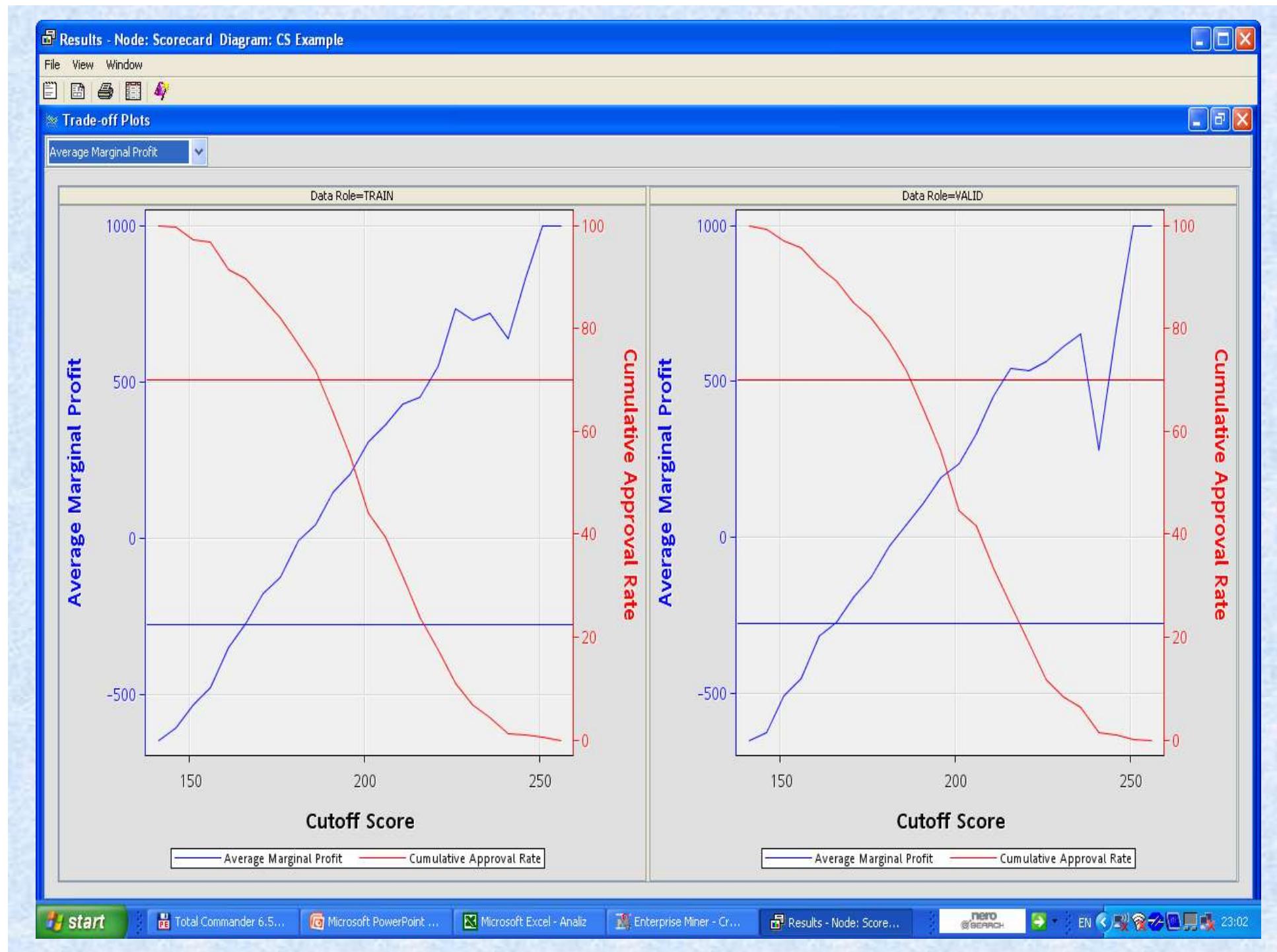
- Определение целевой переменой
- Оценка рисков на базе кредитной истории
- Подготовка предикторов
- Логистическая регрессия
- Оценка эффективности скоринговых карт

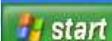
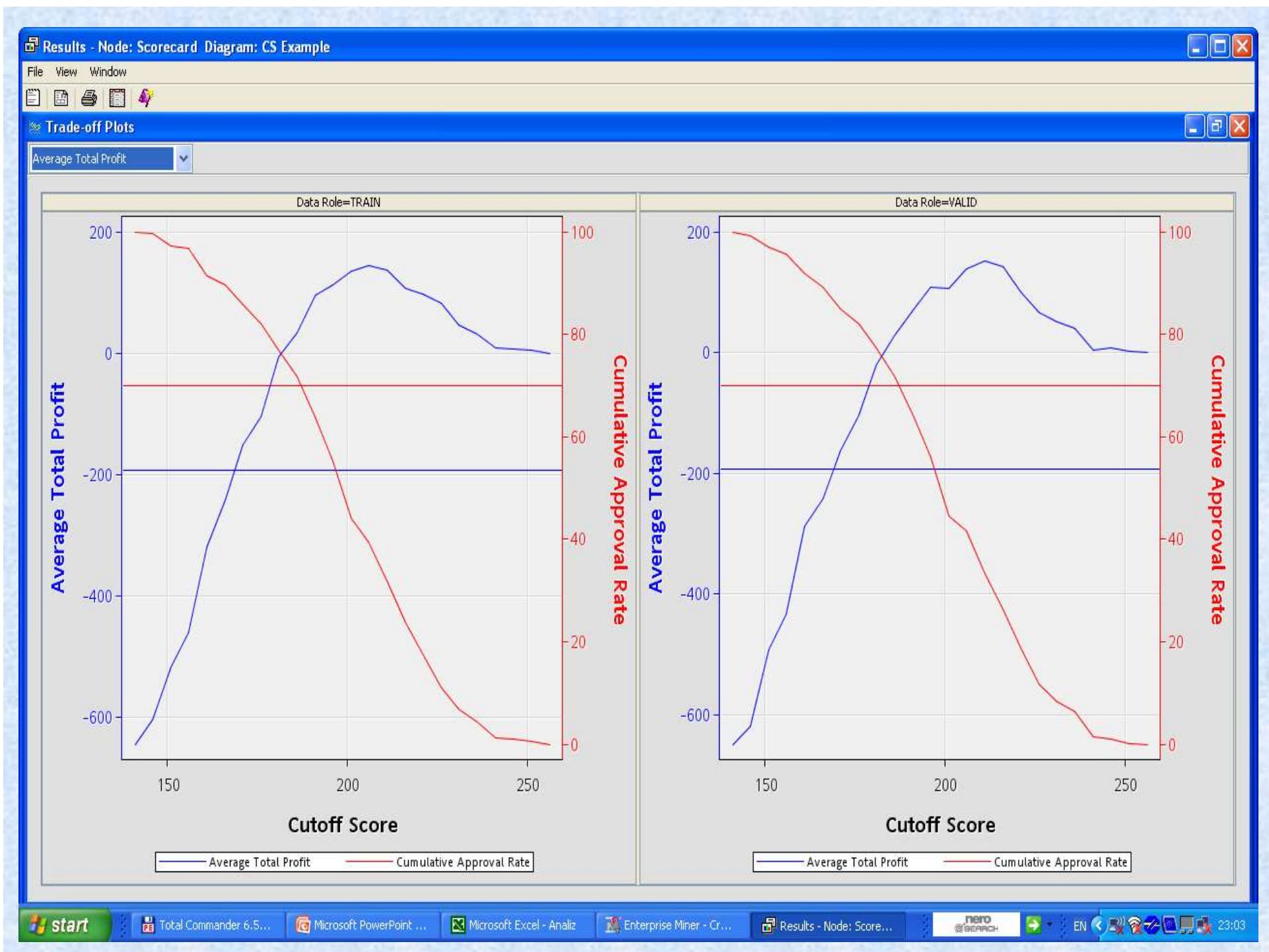
Использование скоринговой карты

- Установление лимитов по скоринговой карте
- Установление ставки по кредиту
- Прогноз прибыли
- Прогноз невозврата

Анализ, определение оптимального балла отсечения

- Понять, какой балл отсечения является оптимальным с точки зрения уровня продаж, поможет соотношение средней прибыли на одного заявителя и уровня продаж.
- Из графика видно, что максимальная прибыль на одного заявителя соответствует баллу отсечения 350.
- Помимо выбора балла отсечения целесообразным представляется автоматическое отнесение обрабатываемой заявки к «белой», «серой» или «черной» зоне.
- Черная зона (высокий уровень риска) - автоматический отказ.
- Белая зона (низкий уровень риска) - автоматическое согласие, наиболее выгодные предложения по кредитованию.
- Серая зона - более подробное рассмотрение заявки, возможно, корректировка запрашиваемой суммы кредита или условий кредитования.





Total Commander 6.5...

Microsoft PowerPoint ...

Microsoft Excel - Analiz

Enterprise Miner - Cr...

Results - Node: Score...

nero @SEARCH

EN

23:03

Модуль отчетности Credit Scoring Solution					
Балл отсечения	150	170	190	220	250
Уровень одобрения заявок	90%	65%	50%	30%	15%
Уровень дефолтов	20%	15%	10%	5%	2%
Прибыль на 1 заемщика	9000	11000	15000	17000	19000
Прибыль на 1 заявителя	1500	10000	8000	6000	3000

Графики поддержки принятия решений при выборе кредитной стратегии (Trade-off):

графики показывают следующие статистики в зависимости от выбора балла отсечения по скоринговой карте:

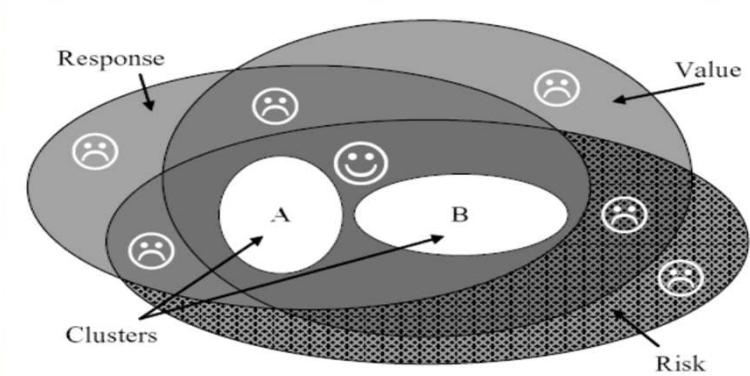
Ожидаемый уровень одобрения кредитных заявок (cumulative approval rate)

Ожидаемый уровень дефолтов (cumulative event rate)

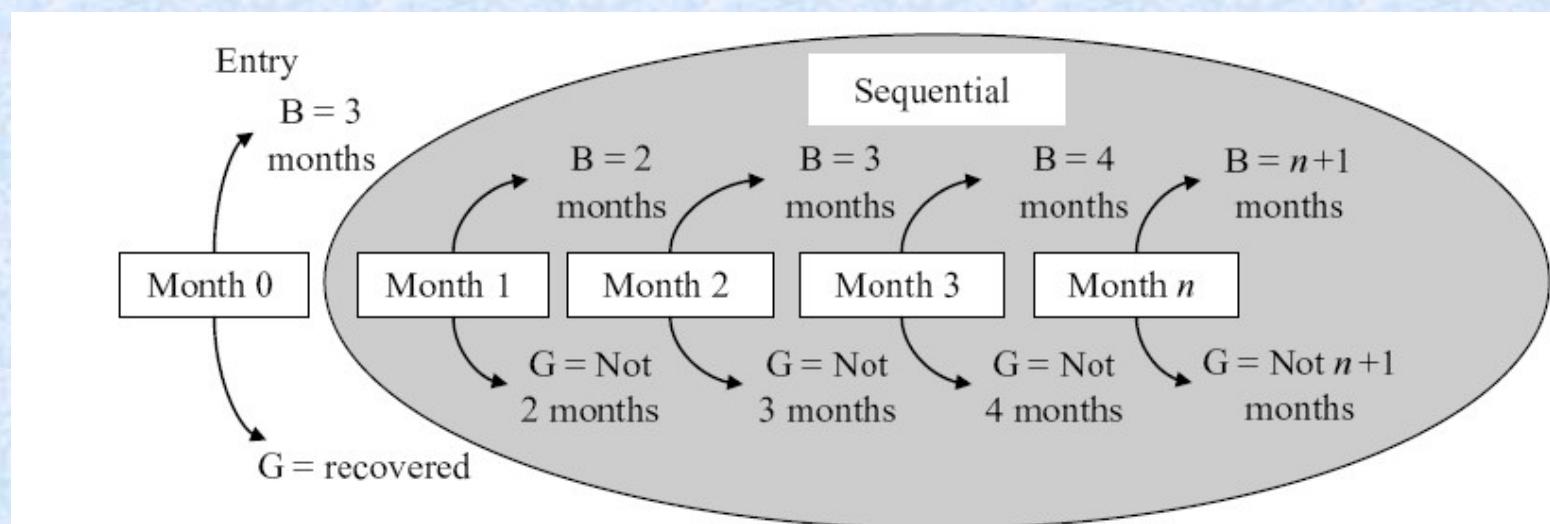
Средняя ожидаемая прибыль на одного заемщика (average marginal profit)

Средняя ожидаемая прибыль на одного заявителя (average total profit)

Score	Number of applicants	Loan amount requested (\$)	Insurance income (\$)	Interest income (%)	Bad rate (%)	Provision rate	Provision (\$)	Operating cost (\$)	Contribution (\$)
232	7,481	3,110	237	530	40.8	45.0	1,399	238	-870
247	6,209	2,713	200	445	30.5	33.5	910	195	-460
254	5,319	2,772	204	452	26.2	28.7	797	179	-320
259	4,695	2,725	203	438	23.2	25.3	690	167	-216
264	5,604	2,648	187	420	20.8	22.6	599	158	-150
268	6,256	2,721	198	435	18.7	20.3	551	151	-69
273	6,928	2,686	187	423	16.7	18.0	483	143	-16
.....									
277	8,157	2,771	194	439	14.9	15.9	442	137	54
282	9,200	2,794	192	441	13.2	14.0	392	131	110
286	9,505	2,868	202	456	11.7	12.4	356	126	176
291	10,508	2,888	197	463	10.4	10.9	315	122	223
295	11,180	2,891	191	452	9.2	9.5	276	117	250



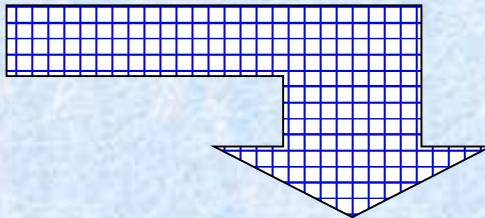
- Net return = value x interest x prob. of recovery - cost of action



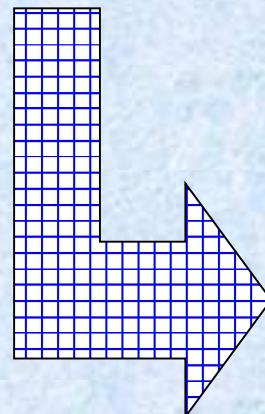
Добавляем параметр – время: сколько жить кредиту?

- Модель живучести (*Survival*) позволяет использовать Cox Model для построения скоринговой карты с информацией о времени возможного дефолта (когда заемщик перестает выплачивать кредит).
- Этот модуль позволяет оценить вероятность дефолта в заданный период времени (например, после 6 месяцев, 9 месяцев, и т.д.).
- $$h(t) = h_0(t) \times \exp(b_{\text{age}} \cdot \text{age} + b_{\text{sex}} \cdot \text{sex} + \dots + b_{\text{group}} \cdot \text{group})$$
- Аналогично добавляем параметры – макроэкономические, которые влияют на скоринговый бал (инфляция, курс валюты, ...)

Ресурсы



Скоринговая карта



**Кредитный
портфель**

Проблемы использования

- Есть трудности связанные с оценкой стоимости активов:
 - (i) изменение стоимости активов влияет на вероятность банкротства;
 - (ii) стоимость активов зависит от динамики экономических параметров, чье значение изменяется во времени и появления новых значимых факторов.
- Необходимо дополнительно строить модели прогноза потерь (убытков) связанных с динамикой макроэкономических переменных.

5 мифов о кредитном скоринге

- Кредит-скоринговая система представляет собой магический “черный” ящик.
- Можно заказать разработку кредит-скоринговых моделей у внешних компаний.
В чем же проблема?
- Кредит-скоринговая процедура применяется один раз при выдаче кредита.
- Процедура должна быть простой, не требует сложной аналитики.
- Приносит мало пользы, можно забыть об этом или отложить внедрение.

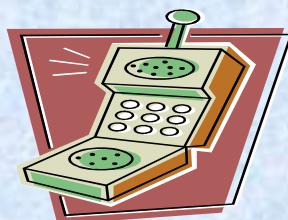


**Докладчики широко использовали
открытые материалы многих авторов и
организаций без указания ссылок
за что приносят им благодарность!**



Дякую за увагу!

**Кафедра прикладних
інформаційних систем**



0957166446



scoreinua@gmail.com