Analytics2012
CONFERENCE
SERIES

# Influence of variable interactions versus segmentation in credit scoring: a case study

Martin Řezáč, Jan Kovář

# Introduction

- The industry standard for modelling the probability of client default is the logistic regression.

- However, details such as inclusion of interactions of predictors or segmentation of given data sample may determine the success.

- The presentation deals with measures of quality of a credit scoring model and with the influence of inclusion of variable interactions versus segmentation on this quality.

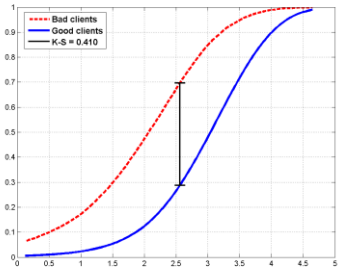
# Measuring the quality of a credit scoring model

- There are many measures of the quality of a credit scoring models. We focus only on:
  - Common (the most favourite) measures:
    - Kolmogorov-Smirnov statistic (KS)
    - Gini index (and Lorenz curve)
    - Lift (QLift)
  - Advanced measures (Lift based)
    - Lift Ratio
    - Integrated Relative Lift

# KS, Lorenz curve, Gini index

- KS is defined as maximal absolute difference between CDFs of scores of good and bad clients:

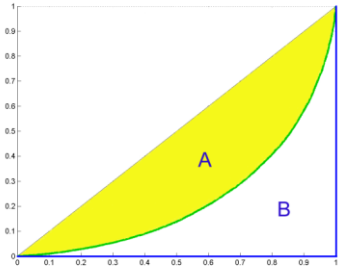
$$KS = \sup_{s \in \Re} \left| F_{BAD}(s) - F_{GOOD}(s) \right|$$




  - It takes values from 0 to 1. The value 0 corresponds to random model, the value 1 corresponds to ideal model.

- Lorenz curve is defined parametrically:

$$x = F_{BAD}(s)$$
$$y = F_{GOOD}(s)$$




- Gini index is defined as:

$$Gini = 1 - 2\int_0^1 F_{GOOD}\left(F_{BAD}^{-1}(s)\right)ds = \frac{A}{A+B} = 2A$$

  - It takes values from 0 to 1. The value 0 corresponds to random model, the value 1 corresponds to ideal model. It is connected to AUC by Gini=2*AUC-1.

# Lift and Qlift

- *Cumulative Lift* says how many times, at a given level of rejection, is the scoring model better than random selection (random model).
- Lift can be expressed and computed by formula:

$$Lift(s) = \frac{F_{BAD}(s)}{F_{ALL}(s)}$$

- In practice, Lift is computed corresponding to 10%, 20%, . . . , 100% of clients with the worst score. Hence we define:

$$QLift(q) = \frac{F_{BAD}(F_{ALL}^{-1}(q))}{F_{ALL}(F_{ALL}^{-1}(q))} = \frac{1}{q} F_{BAD}(F_{ALL}^{-1}(q)), \quad q \in (0,1]$$

$$F_{ALL}^{-1}(q) = \min\{s, F_{ALL}(s) \geq q\}$$

- Typical value of $q$ is 0.1. Then we have

$$QLift_{10\%} = QLift(0.1) = 10 \cdot F_{BAD}(F_{ALL}^{-1}(0.1))$$

  - It takes values from 1 (random model) to some upper limit.

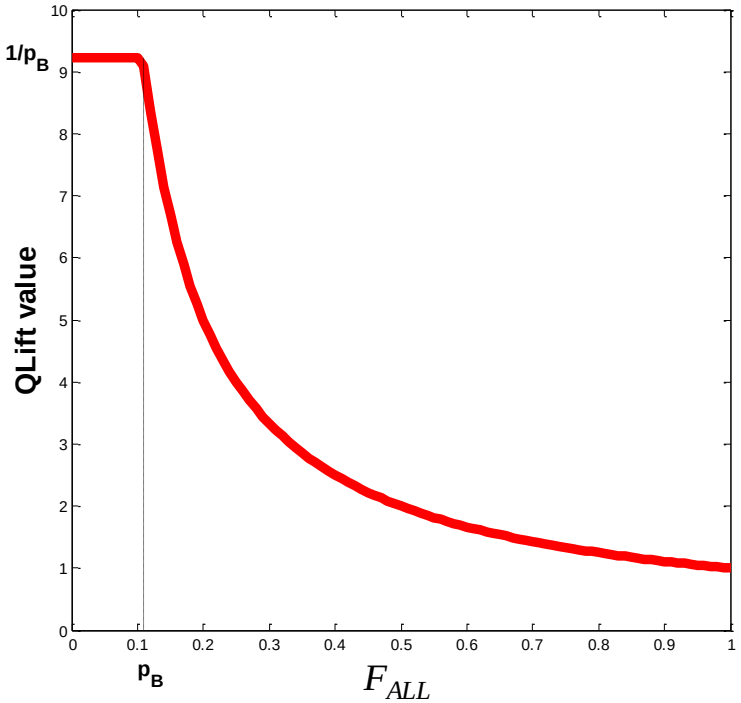
# Lift and Qlift for ideal model

- It is natural to ask how look Lift and QLift in case of ideal model. Hence we derived (ŘEZÁČ, Martin and Jan KOLÁČEK. On Aspects of Quality Indexes for Scoring Models. In *19th International Conference on Computational Statistics, Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. Paris: SpringerLink) , following formulas.
  - » Lift for ideal model:

$$Lift_{ideal}(s) = \begin{cases} \dfrac{1}{p_B} & s \le c \\ \dfrac{1}{F_{ALL}(s)} & s > c \end{cases}$$

  - » QLift for ideal model:

$$QLift_{ideal}(q) = \begin{cases} \dfrac{1}{p_B} & q \in (0, p_B] \\ \dfrac{1}{q} & q \in (p_B, 1] \end{cases}$$

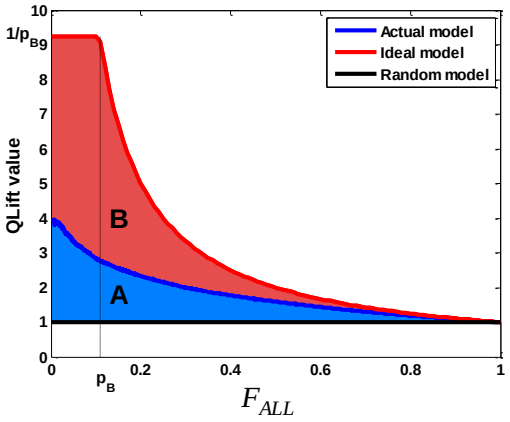



We can see that the upper limit of Lift and QLift is equal to $1/p_B$ , where $p_B$ is the proportion of bad clients.

# Lift Ratio, RLift, IRL

- Once we know form of QLift for ideal model, we can define Lift Ratio, Relative Lift and Integrated Relative Lift.
  - Lift ratio (LR):

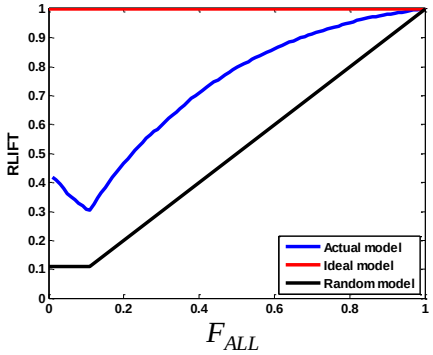
$$LR = \frac{\int_0^1 QLift(q)\,dq - 1}{\int_0^1 QLift_{ideal}(q)\,dq - 1} = \frac{A}{A+B}$$




  - It is analogy to Gini index. It takes values from 0 to 1. Value 0 corresponds to random model, value 1 match to ideal model. Important feature is that Lift Ratio allows us to fairly compare two models developed on different data samples, which is not possible with Lift.

- Since Lift Ratio compares areas under Lift function for actual and ideal models, next concept is focused on comparison of Lift functions themselves. We define Relative Lift function by :

$$RLif(q) = \frac{QLift(q)}{QLift_{ideal}(q)},\ q \in (0,1]$$




- Integrated Relative Lift (IRL):

$$IRL = \int_0^1 RLift(q)\,dq$$

  - It takes values from $0.5(1+p_B^2)$, for random model, to 1, for ideal model.
  - Both the LR and IRL, compared to Gini and AUC, penalize more that models, which have weak performance on the left side of a score scale.

# Case study

- We have got a data file, developed a SAS macros (for automatic creation of segments and development of credit scoring models) and tried to study influence of inclusion of variable interactions and segmentation on the quality of a credit scoring model. We focused on following questions:
  - How big is the influence of inclusion of interactions?
  - How big is the influence of segmentation?
  - Which variable(s) to use for segmentation? Is there a relationship between Information Value of selected variable and the quality of final scoring model?
- Using some results based on another data we tried to find an answer to the question:
  - What is the time stability of a complex segmentation? What level of complexity is too complex?

# Case study 1

- Data file was obtained from KAGGLE competition „Give Me Some Credit“
- Data description:
  - 150 000 cases
  - 10 explanatory variables (+ some more created from these vars.)
  - Target variable: default (90 DPD) in the next two years


| Variable Name | Description | Type |
|---|---|---|
| **SeriousDlqin2yrs** | **Person experienced 90 days past due delinquency or worse** | **Y/N** |
| RevolvingUtilizationOfUnsecuredLines | Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits | percentage |
| age | Age of borrower in years | integer |
| NumberOfTime30-59DaysPastDueNotWorse | Number of times borrower has been 30-59 days past due but no worse in the last 2 years. | integer |
| DebtRatio | Monthly debt payments, alimony,living costs divided by monthy gross income | percentage |
| MonthlyIncome | Monthly income | real |
| NumberOfOpenCreditLinesAndLoans | Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards) | integer |
| NumberOfTimes90DaysLate | Number of times borrower has been 90 days or more past due. | integer |
| NumberRealEstateLoansOrLines | Number of mortgage and real estate loans including home equity lines of credit | integer |
| NumberOfTime60-89DaysPastDueNotWorse | Number of times borrower has been 60-89 days past due but no worse in the last 2 years. | integer |
| NumberOfDependents | Number of dependents in family excluding themselves (spouse, children etc.) | integer |

- **Weight of evidence, information value:**

$r$ ... number of levels (categories) of the variable
$g_i$ ... number of ”goods” the in $i$-th category
$b_i$ ... number of ”bads” the in $i$-th category
$G := \Sigma\, g_i$ ... total number of ”goods”
$B := \Sigma\, b_i$ ... total number of ”bads”

**Weight of evidence** for the $i$-th category: $\boldsymbol{woe_i = ln\,(g_i / G) - ln\,(b_i / B)}$

**Information value** for the $i$-th category: $\boldsymbol{Inf\_val_i = [(g_i / G) - (b_i / B)] \cdot woe_i}$

**Total information value** for the corresponding variable: $\boldsymbol{Inf\_val = \Sigma\, inf\_val_i}$

# Case study 1

- Quality indexes of considered models (segmentation into 2 data parts):

| | | Index: | | | | | |
|---|---|---|---|---|---|---|---|
| | IV of var. | Gini | KS | $QLift_{10\%}$ | $QLift_{20\%}$ | LR | IRL |
| base model, no segments, no interactions | | 0,719 | 0,563 | 5,419 | 3,583 | 0,598 | 0,855 |
| base model, no segments, with interactions | | 0,720 | 0,565 | 5,413 | 3,592 | 0,597 | 0,856 |
| **segmented (2 seg.) by:** | | | | | | | |
| DPD | 1,4112 | 0,726 | 0,569 | 5,454 | 3,615 | 0,602 | 0,858 |
| RevUtil | 1,1134 | 0,725 | 0,562 | 5,454 | 3,582 | 0,603 | 0,858 |
| N90DaysLate | 0,8376 | 0,724 | 0,568 | 5,441 | 3,597 | 0,602 | 0,858 |
| age | 0,2569 | 0,723 | 0,568 | 5,424 | 3,614 | 0,600 | 0,857 |
| NOfOpenCreditLines | 0,0821 | 0,725 | 0,571 | 5,432 | 3,617 | 0,600 | 0,858 |
| MonthlyIncome | 0,0801 | 0,721 | 0,564 | 5,426 | 3,577 | 0,600 | 0,857 |
| DebtRatio | 0,075 | 0,723 | 0,569 | 5,434 | 3,589 | 0,601 | 0,857 |
| NRealEstateL | 0,0554 | 0,723 | 0,568 | 5,418 | 3,606 | 0,599 | 0,857 |
| NumDepend | 0,0353 | 0,720 | 0,565 | 5,401 | 3,594 | 0,597 | 0,856 |
| income_near5000 | 0,0167 | 0,721 | 0,565 | 5,413 | 3,588 | 0,598 | 0,856 |
| DR1 | 0,0129 | 0,721 | 0,564 | 5,404 | 3,591 | 0,597 | 0,856 |
| Spearman coeff. of rank corr. b/t. IV and index | | 0,8500 | 0,3182 | 0,8886 | 0,3000 | 0,9182 | 0,8636 |

| | Index: | | | | | |
|---|---|---|---|---|---|---|
| | Gini | KS | $QLift_{10\%}$ | $QLift_{20\%}$ | LR | IRL |
| | improvements (compared to base model without interactions and with no segmentation) | | | | | |
| | 0,14% | 0,32% | -0,11% | 0,25% | -0,12% | 0,02% |
| | | | | | | |
| | 0,97% | 1,17% | 0,65% | 0,89% | 0,75% | 0,35% |
| | 0,83% | -0,09% | 0,65% | -0,03% | 0,82% | 0,35% |
| | 0,70% | 0,87% | 0,41% | 0,39% | 0,62% | 0,30% |
| | 0,56% | 0,91% | 0,09% | 0,87% | 0,35% | 0,18% |
| | 0,83% | 1,39% | 0,24% | 0,95% | 0,38% | 0,30% |
| | 0,28% | 0,20% | 0,13% | -0,17% | 0,30% | 0,15% |
| | 0,56% | 1,03% | 0,28% | 0,17% | 0,52% | 0,25% |
| | 0,56% | 0,89% | -0,02% | 0,64% | 0,17% | 0,23% |
| | 0,14% | 0,43% | -0,33% | 0,31% | -0,20% | 0,04% |
| | 0,28% | 0,34% | -0,11% | 0,14% | 0,05% | 0,12% |
| | 0,28% | 0,14% | -0,28% | 0,22% | -0,22% | 0,06% |

- worst
- average
- best performance

- The influence of interactions is imponderable, the influence of segmentation is very low in this case (below 1%).
- For some models increased Gini and KS, but decreased $QLift_{10\%}$ and LR.
- Spearman rank correlation is high for LR, it is very low for KS and $QLift_{20\%}$.

# Case study 1

- Quality indexes of considered models (segmentation into 3 data parts):

| | | Index: | | | | | |
|---|---|---|---|---|---|---|---|
| segmented (3 seg.) by: | IV of var. | Gini | KS | $QLift_{10\%}$ | $QLift_{20\%}$ | LR | IRL |
| DPD | 1,4112 | 0,728 | 0,572 | 5,483 | 3,625 | 0,606 | 0,860 |
| RevUtil | 1,1134 | 0,728 | 0,566 | 5,461 | 3,597 | 0,604 | 0,859 |
| N90DaysLate | 0,8376 | 0,724 | 0,568 | 5,440 | 3,597 | 0,601 | 0,858 |
| age | 0,2569 | 0,722 | 0,565 | 5,445 | 3,603 | 0,601 | 0,857 |
| NOfOpenCreditLines | 0,0821 | 0,727 | 0,570 | 5,445 | 3,619 | 0,602 | 0,859 |
| MonthlyIncome | 0,0801 | 0,723 | 0,565 | 5,448 | 3,594 | 0,602 | 0,858 |
| DebtRatio | 0,075 | 0,724 | 0,570 | 5,426 | 3,604 | 0,599 | 0,857 |
| NRealEstateL | 0,0554 | 0,724 | 0,568 | 5,417 | 3,610 | 0,599 | 0,858 |
| NumDepend | 0,0353 | 0,721 | 0,566 | 5,372 | 3,591 | 0,594 | 0,855 |
| income_near5000 | 0,0167 | 0,722 | 0,567 | 5,418 | 3,587 | 0,599 | 0,857 |
| DR1 | 0,0129 | 0,721 | 0,565 | 5,322 | 3,562 | 0,589 | 0,854 |
| Spearman coeff. of rank corr. b/t. IV and index | | 0,7841 | 0,3364 | 0,8886 | 0,6341 | 0,8909 | 0,9045 |

| | Index: | | | | | |
|---|---|---|---|---|---|---|
| | Gini | KS | $QLift_{10\%}$ | $QLift_{20\%}$ | LR | IRL |
| | improvements (compared to base model without interactions and with no segmentation) | | | | | |
| DPD | 1,25% | 1,71% | 1,18% | 1,17% | 1,34% | 0,50% |
| RevUtil | 1,25% | 0,57% | 0,78% | 0,39% | 1,00% | 0,48% |
| N90DaysLate | 0,70% | 0,87% | 0,39% | 0,39% | 0,60% | 0,30% |
| age | 0,42% | 0,36% | 0,48% | 0,56% | 0,49% | 0,20% |
| NOfOpenCreditLines | 1,11% | 1,33% | 0,48% | 1,00% | 0,67% | 0,39% |
| MonthlyIncome | 0,56% | 0,41% | 0,54% | 0,31% | 0,69% | 0,28% |
| DebtRatio | 0,70% | 1,21% | 0,13% | 0,59% | 0,25% | 0,20% |
| NRealEstateL | 0,70% | 0,91% | -0,04% | 0,75% | 0,15% | 0,26% |
| NumDepend | 0,28% | 0,53% | -0,87% | 0,22% | -0,64% | -0,01% |
| income_near5000 | 0,42% | 0,75% | -0,02% | 0,11% | 0,18% | 0,15% |
| DR1 | 0,28% | 0,37% | -1,79% | -0,59% | -1,51% | -0,21% |

- worst
- average
- best performance

- The influence of segmentation is low, but higher than on the previous slide.
- Again, for some models increased Gini and KS, but decreased $QLift_{10\%}$, LR and IRL.
- Spearman rank correlation is high for IRL, LR and $QLift_{10\%}$, it is very low for KS.

# Case study 1

- Quality indexes of considered models (segmentation into 2 data parts, using variable interaction):

| interaction, segmented (2 seg.) by: | IV of var. | Index: Gini | KS | $QLift_{10\%}$ | $QLift_{20\%}$ | LR | IRL |
|---|---|---|---|---|---|---|---|
| DPD | 1,4112 | 0,728 | 0,571 | 5,463 | 3,626 | 0,604 | 0,859 |
| RevUtil | 1,1134 | 0,726 | 0,565 | 5,460 | 3,579 | 0,604 | 0,859 |
| N90DaysLate | 0,8376 | 0,726 | 0,570 | 5,450 | 3,611 | 0,603 | 0,859 |
| age | 0,2569 | 0,724 | 0,570 | 5,429 | 3,620 | 0,599 | 0,857 |
| NOfOpenCreditLines | 0,0821 | 0,726 | 0,569 | 5,440 | 3,619 | 0,601 | 0,858 |
| MonthlyIncome | 0,0801 | 0,723 | 0,566 | 5,442 | 3,591 | 0,601 | 0,858 |
| DebtRatio | 0,075 | 0,725 | 0,570 | 5,426 | 3,601 | 0,600 | 0,858 |
| NRealEstateL | 0,0554 | 0,723 | 0,569 | 5,418 | 3,608 | 0,598 | 0,857 |
| NumDepend | 0,0353 | 0,720 | 0,567 | 5,423 | 3,593 | 0,599 | 0,856 |
| income_near5000 | 0,0167 | 0,722 | 0,564 | 5,415 | 3,594 | 0,599 | 0,857 |
| DR1 | 0,0129 | 0,721 | 0,564 | 5,341 | 3,582 | 0,590 | 0,854 |
| Spearman coeff. of rank corr. b/t. IV and index | | 0,9023 | 0,6455 | 0,9545 | 0,4545 | 0,8977 | 0,9273 |

| Index: improvements (compared to base model without interactions and with no segmentation) Gini | KS | $QLift_{10\%}$ | $QLift_{20\%}$ | LR | IRL |
|---|---|---|---|---|---|
| 1,25% | 1,39% | 0,81% | 1,20% | 0,95% | 0,47% |
| 0,97% | 0,48% | 0,76% | -0,11% | 0,99% | 0,41% |
| 0,97% | 1,35% | 0,57% | 0,78% | 0,80% | 0,39% |
| 0,70% | 1,28% | 0,18% | 1,03% | 0,27% | 0,21% |
| 0,97% | 1,07% | 0,39% | 1,00% | 0,49% | 0,33% |
| 0,56% | 0,59% | 0,42% | 0,22% | 0,59% | 0,26% |
| 0,83% | 1,33% | 0,13% | 0,50% | 0,40% | 0,28% |
| 0,56% | 1,03% | -0,02% | 0,70% | 0,10% | 0,19% |
| 0,14% | 0,75% | 0,07% | 0,28% | 0,12% | 0,05% |
| 0,42% | 0,30% | -0,07% | 0,31% | 0,12% | 0,18% |
| 0,28% | 0,16% | -1,44% | -0,03% | -1,29% | -0,16% |

- worst
- average
- best performance

- The influence of adding interactions is comparable with the more complex segmentation on the previous slide.
- For one considered model increased Gini and KS, but decreased $QLift_{10\%}$, LR and IRL.
- Spearman rank correlation is very high for $QLift_{10\%}$, high for IRL and Gini and it is low for KS and $QLift_{20\%}$.

# Case study 1

- Quality indexes of considered models (segmentation into 3 data parts, using variable interaction):

| | | Index: | | | | | |
|---|---|---|---|---|---|---|---|
| | IV of var. | Gini | KS | $QLift_{10\%}$ | $QLift_{20\%}$ | LR | IRL |
| **interaction, segmented (3 seg.) by:** | | | | | | | |
| DPD | 1,4112 | 0,730 | 0,574 | 5,501 | 3,636 | 0,607 | 0,860 |
| RevUtil | 1,1134 | 0,729 | 0,569 | 5,466 | 3,595 | 0,605 | 0,860 |
| N90DaysLate | 0,8376 | 0,726 | 0,570 | 5,447 | 3,611 | 0,602 | 0,859 |
| age | 0,2569 | 0,731 | 0,577 | 5,645 | 3,692 | 0,598 | 0,861 |
| NOfOpenCreditLines | 0,0821 | 0,727 | 0,572 | 5,449 | 3,621 | 0,602 | 0,859 |
| MonthlyIncome | 0,0801 | 0,724 | 0,566 | 5,457 | 3,606 | 0,603 | 0,858 |
| DebtRatio | 0,075 | 0,725 | 0,571 | 5,421 | 3,610 | 0,599 | 0,857 |
| NRealEstateL | 0,0554 | 0,725 | 0,568 | 5,421 | 3,616 | 0,599 | 0,857 |
| NumDepend | 0,0353 | 0,721 | 0,565 | 5,399 | 3,599 | 0,597 | 0,856 |
| income_near5000 | 0,0167 | 0,723 | 0,566 | 5,428 | 3,596 | 0,600 | 0,857 |
| DR1 | 0,0129 | 0,721 | 0,564 | 5,270 | 3,555 | 0,584 | 0,852 |
| Spearman coeff. of rank corr. b/t. IV and index | | 0,8841 | 0,7727 | 0,8409 | 0,5364 | 0,7568 | 0,9159 |

| Index: | | | | | |
|---|---|---|---|---|---|
| Gini | KS | $QLift_{10\%}$ | $QLift_{20\%}$ | LR | IRL |
| improvements (compared to base model without interactions and with no segmentation) | | | | | |
| 1,53% | 1,97% | 1,51% | 1,48% | 1,61% | 0,60% |
| 1,39% | 1,08% | 0,87% | 0,33% | 1,17% | 0,53% |
| 0,97% | 1,33% | 0,52% | 0,78% | 0,75% | 0,39% |
| 1,67% | 2,47% | 4,17% | 3,04% | 0,10% | 0,69% |
| 1,11% | 1,60% | 0,55% | 1,06% | 0,75% | 0,42% |
| 0,70% | 0,52% | 0,70% | 0,64% | 0,82% | 0,33% |
| 0,83% | 1,46% | 0,04% | 0,75% | 0,22% | 0,23% |
| 0,83% | 0,87% | 0,04% | 0,92% | 0,17% | 0,23% |
| 0,28% | 0,44% | -0,37% | 0,45% | -0,20% | 0,05% |
| 0,56% | 0,50% | 0,17% | 0,36% | 0,32% | 0,19% |
| 0,28% | 0,21% | -2,75% | -0,78% | -2,39% | -0,40% |

- worst
- average
- best performance

- The influence of adding interactions in combination with more complex segmentation is the biggest –compared with the previous slides.
- Again, for some models increased Gini and KS, but decreased $QLift_{10\%}$, LR and IRL.
- Spearman rank correlation is high for IRL, Gini and $QLift_{10\%}$, it is low for KS and LR, and it is very low for $QLift_{20\%}$.

# Case study 2

- Data provided by a financial company operating in Central and Eastern Europe providing small- and medium-sized consumer loans. Data were registered in 2004 - 2006. To preserve confidentiality, the data were selected in such a way as to provide heavy distortion in the parameters describing the true solvency situation of the financial company.
  - Around 1 100 000 cases of fraudsters, 2 500 000 cases of defaulters
  - 21 explanatory variables
  - Target variables:
    - Fraud: 90 DPD on first payment
    - Default: 60 DPD on $2^{nd}$ - $4^{th}$ payment

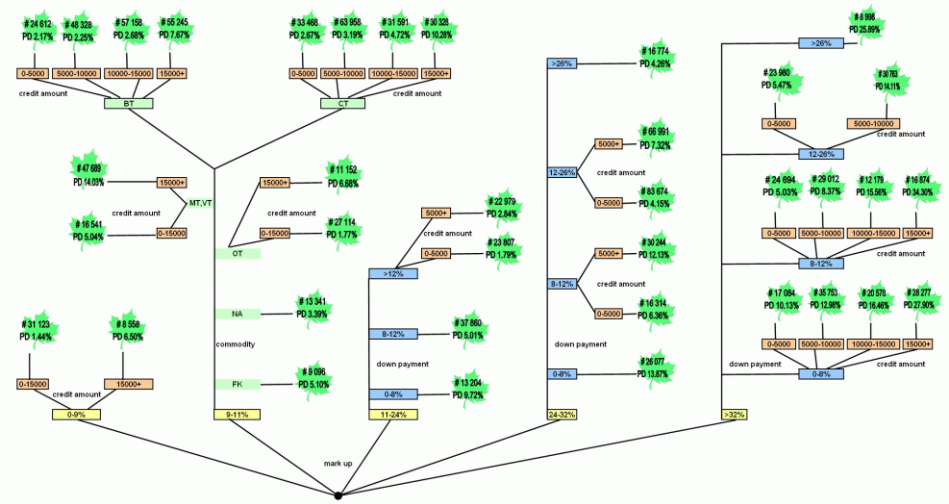
# Case study 2

- Tested variants:
  - 2 logistic regressions (LRs for whole sample frauds and defaulters)
  - Expert segments (**2x7** LRs for frauds and defaulters)
    - Using commodity (mobiles, furniture,…)
  - Expert segments (**2x29** LRs for frauds and defaulters)
    - Using commodity x distribution channel
  - Chaid-tree segments (**2x70** LRs for frauds and defaulters)
  - Product segments (**2x37** LRs for frauds and defaulters)

| | | Gini | improvement (compared to base model without segmentation) |
|---|---|---|---|
| Defaulters | unseg. | 0.423 | |
| | 7 seg. | 0.444 | 4.96% |
| | 29 seg. | 0.465 | 9.93% |
| | 70 seg | 0.513 | 21.28% |
| | 37 seg | 0.472 | 11.58% |
| Fraudsters | unseg. | 0.597 | |
| | 7 seg. | 0.625 | 4.69% |
| | 29 seg. | 0.664 | 11.22% |
| | 70 seg | 0.662 | 10.89% |
| | 37 seg | 0.649 | 8.71% |

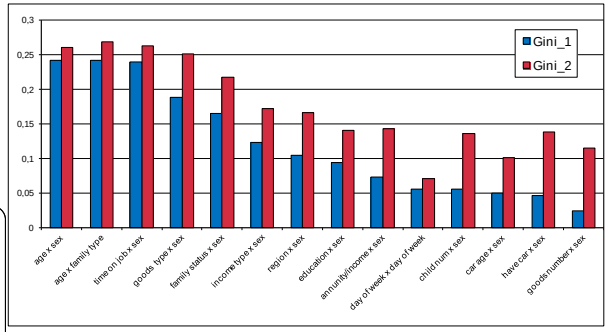
# Case study 2

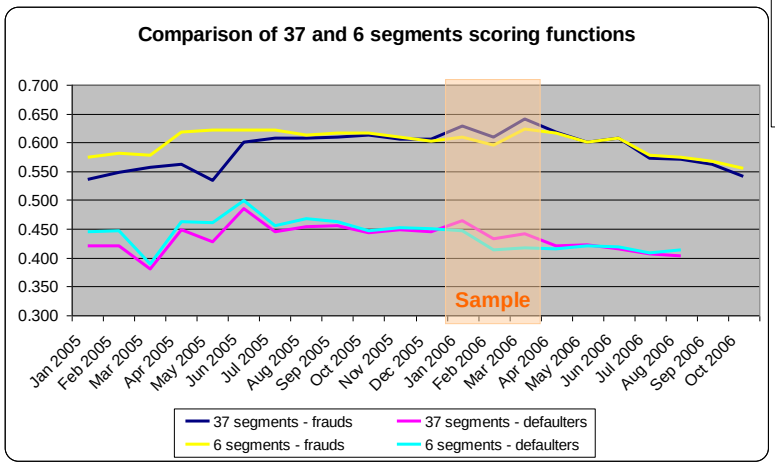
- Features taken into account (for 37 segments):
  - Experience from regression trees – the best Ginis
  - Product dimension – risk over products
  - Commodity dimension – risk over commodities
  - Statistic stability – sufficiently large sample in segment
  - Price control – bright insight to profit analysis
  - Variables for segmentation:
    - » Mark up
    - » Down payment
    - » Credit amount
    - » Commodity type
    - » …Information value were very high for all of these



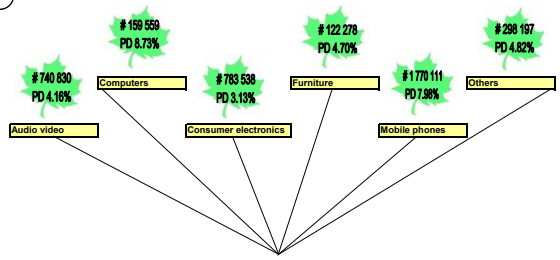
# Case study 2

- One year later = new segmentation, var. interactions
  - Developed new models with 6 segments and some more complex variable interactions and compared with models with 37 segments (with redeveloped coefficients)
  - Interactions between variables provided for a significant gain in Gini coefficient.
  - Time stability




- Advantages of new segmentation:
  - Clear structure of segments
  - Better time stability of developed models
  - More simple monitoring



# Summary

- Both segmentation and interactions leads to more powerful models compared to models without them.
- Influence of segmetation is higher than influance of interactions.
- Segmentation by variables with the highest Information values leads to the best performance.
- Validation of stability in time is crutial. Too complex segmentations and too complex interactions lead to overfitting and consequently to bad performance in real process.

Analytics2012
CONFERENCE
SERIES

# Thank you for your attention