Credit scoringové modely - vývoj, implementace, praxe

Martin Řezáč

Bratislava, 6.12.2013

Obsah

Úvod

- Credit scoring je množina prediktivních modelů a jim příslušných statistických technik, které pomáhají finančním institucím při automatickém posuzování žádostí o úvěr.

- Se znalostí pravděpodobnosti selhání žadatele, očekávané míry zamítání, očekávaného zisku/ztráty, popř. dalších business předpokladů, lze efektivně rozhodnout kdo úvěr dostane, v jaké výši a jaké další strategie maximalizují profit plynoucí z daného žadatele/klienta.

Úvod

- Zatímco historie úvěru sahá 4000 let nazpět (první zaznamenaná zmínka o úvěru pochází ze starověkého Babylonu - 2000 let před n.l.), historie credit scoringu je pouze 50-70 let stará.

- První přístup k řešení problému identifikace skupin v populaci představil ve statistice Fisher (1936). V roce 1941, Durand jako první rozpoznal, že tyto techniky mohou být použity k rozlišování mezi dobrými a špatnými úvěry.

Úvod

- Významným milníkem při posuzování úvěrů byla druhá světová válka.
- Do té doby bylo standardem individuální posuzování žadatele o úvěr. Dále bylo standardem, že ve finanční sféře byli zaměstnáni (téměř) výhradně muži.
- Odchod značné části mužské populace do služeb armády měl za následek potřebu předat zkušenosti dosavadních posuzovatelů žádostí o úvěr novým pracovníkům.
- Díky tomu vznikla jakási rozhodovací pravidla a došlo k „automatizaci“ posuzování žádostí o úvěr.

Úvod

- Příchod kreditních karet ke konci šedesátých let minulého století a růst výpočetního výkonu způsobil obrovský rozvoj a využití credit scoringových technik. Událost, která zajistila plnou akceptaci credit scoringu, bylo přijětí zákonů „Equal Credit Opportunity Acts" (o rovné příležitosti přístupu k úvěrům) a jeho pozdějších znění přijatých v USA v roce 1975 a 1976. Tyto stanovily za nezákonné diskriminace v poskytování úvěru, vyjma situace, pokud tato diskriminace „byla empiricky odvozená a statisticky validní".

Úvod

- V osmdesátých letech minulého století začala být využívána logistická regrese, dodnes v mnoha oblastech považovaná za průmyslový standard, a lineární programování. O něco později se objevily na scéně metody umělé inteligence, např. neuronové sítě.

Cesta k automatickému scoringu

- “HISTORICAL EVOLUTION”:

Money lender

- lend only to people which he knows

Operators

- they make decision based on client's information and their experience

Automatic scoring

- make decision on statistical base

PAST EXPERIENCE -> ESTIMATION FOR FUTURE

Vývoj credit scoringového modelu

Main Stages – Development

- <u>Stage 1: Preliminaries and Planning</u>
  - Create Business Plan
    - Identify organizational objectives
    - Internal versus External development, and scorecard type
  - Create Project Plan
    - Identify project risks
    - Identify project team.

Main Stages – Development

- Stage 2: Data Review and Project Parameters
  - Data availability and quality
  - Data gathering for definition of project parameters
  - Definition of project parameters
    - Performance window and sample window
    - Performance categories definition (target)
    - Exclusions
  - Segmentation
  - Methodology
  - Review of implementation plan.

Main Stages – Development

- Stage 3: Development Database Creation
  - Development sample specification
  - Sampling
  - Development data collection and construction
  - Adjusting for prior probabilities.

Main Stages – Development

- Stage 4: Scorecard Development
  - Missing values and outliers
  - Initial characteristic analysis
  - Preliminary scorecard
  - Reject inference
  - Final scorecard production
    - Scaling
    - Points allocation
    - Misclassification
  - Scorecard strength
  - Validation.

Main Stages – Development

- Stage 5: Scorecard Management Reports
  - Gains tables and charts
  - Characteristic reports.

Main Stages – Implementation

- Stage 1: Pre-Implementation Validation

- Stage 2: Strategy Development
  - Scoring strategy
  - Setting cutoffs
  - Strategy considerations
  - Policy rules
  - Overrides.

Main Stages – Post Implementation

- Post-Implementation
  - Scorecard and Portfolio Monitoring Reports
  - Review.

Default – definice cílové prom. (good/bad)

- Obvykle je tato definice založena na klientově počtu dnů po splatnosti (Days Past Due, DPD) a částce po splatnosti. S částkou po splatnosti je spojena potřeba stanovení jisté míry tolerance, tedy stanovení co je považováno za významný dluh a co nikoli. Např. nemusí dávat smysl považovat za dluh částky menší než 100 Kč.
- Dále je třeba stanovit časový horizont (performance window), na kterém jsou dva zmíněné parametry sledovány.
- Za dobrého klienta lze např. označit klienta, který:
  - je po splatnosti méně než 60 dnů(s tolerancí 100 Kč) v prvních 6-ti měsících od první splátky,
  - je po splatnosti méně než 90 dnů (s tolerancí 30 Kč) v průběhu celé své platební historie (ever).

Default – definice cílové prom.

- Volba těchto parametrů závisí do značné míry na typu finančního produktu (jistě se bude lišit volba parametrů pro spotřebitelské úvěry pro malé částky se splatností kolem jednoho roku a pro hypotéky, které jsou obvykle spojeny s velmi vysokou finanční částkou a se splatností až několik desítek let) a na dalším využití této definice (řízení rizik, marketing, ...).

Default – definice cílové prom.

- Další praktickým problémem definice dobrého klienta je souběh několika smluv jednoho klienta. Například je možné, že zákazník je po lhůtě splatnosti na více smlouvách, ale s rozdílnými dny po splatnosti a s různými částkami. V tomto případě jsou většinou částky klienta dlužné v jednom konkrétním časovém okamžiku sečteny, a ze dnů po splatnosti na jednotlivých smlouvách je brána maximální hodnota. Tento přístup lze uplatnit pouze v některých případech, a to zejména v situaci, kdy jsou k dispozici kompletní účetní data. Situace je podstatně složitější v případě agregovaných údajů, např. na měsíční bázi.

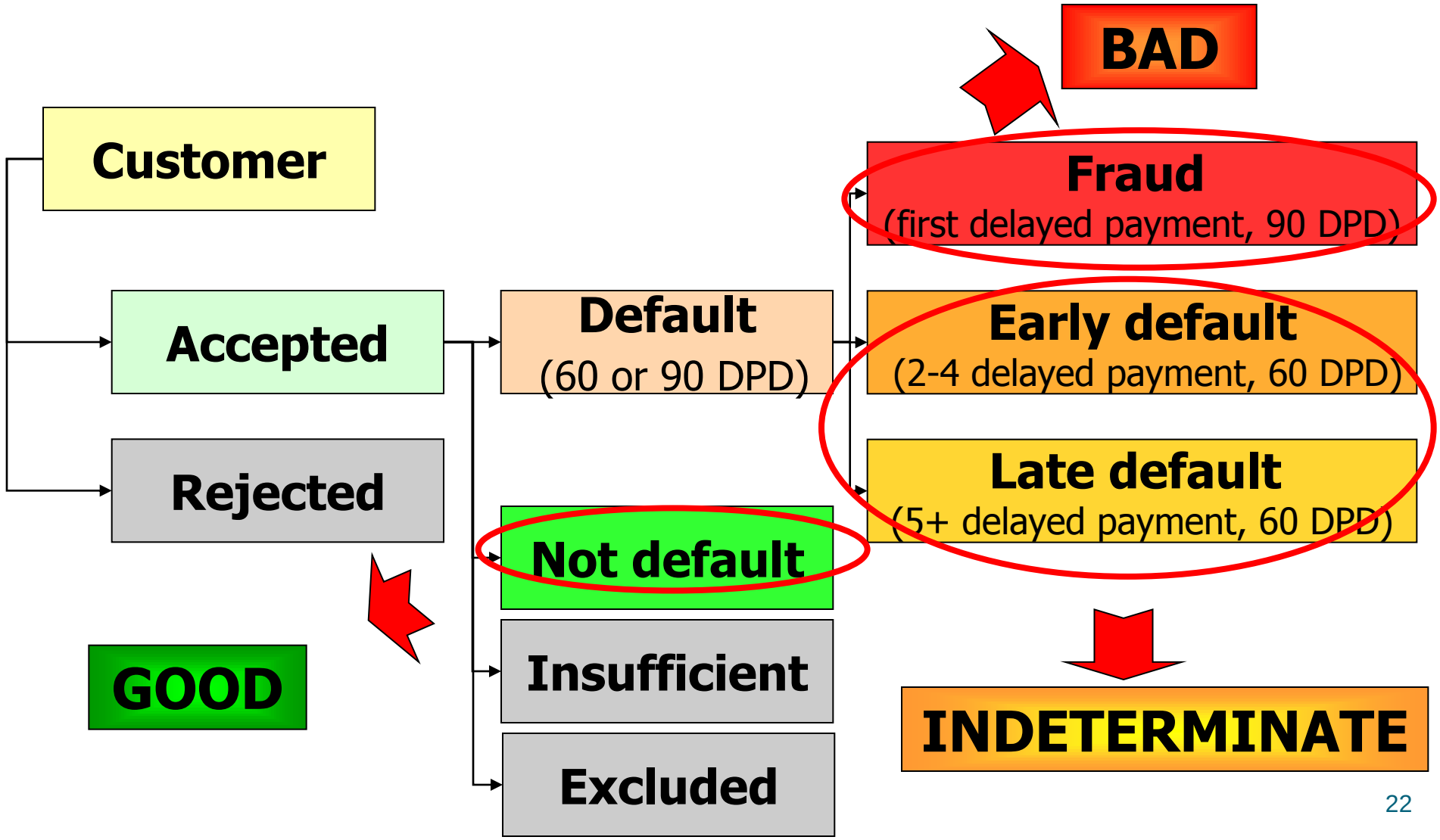
Default – definice cílové prom.

- Obecně uvažujeme následující typy klientů:
  - dobrý (good),
  - špatný (bad),
  - nedefinovaný (indeterminate),
  - s nedostatečnou úvěrovou historií (insufficient),
  - vyřazený (excluded),
  - zamítnutý (rejected).

Default – definice cílové prom.

- První dva typy byly diskutovány. Třetí typ, tj. indeterminate, je na hranici mezi dobrým a špatným klientem a při jeho použití přímo ovlivňuje definici dobrých/špatných klientů. Uvažujeme-li pouze DPD, klienti s vysokými DPD (např. 90 +) jsou typicky označeni za špatné, nedelikventní klienti (jejich DPD je rovno nule) jsou označeni za dobré. Za indeterminate jsou pak označeni delikventní klienti, kteří nepřekročí danou hranici DPD.
- Čtvrtý typ klientů jsou typicky klienti s velmi krátkou platební historií, u kterých je nemožná korektní definice cílové proměnné.
- Vyřazení klienti jsou klienti, jejichž data jsou natolik špatná, že by vedla ke zkreslení modelu(např. fraudy). Další skupinu tvoří klienti, kteří nejsou standardně hodnoceni daným modelem (VIP klienti).
- Poslední typ klientů jsou ti klienti, jejichž žádost o úvěr byla zamítnuta.

Default – definice cílové prom.



Čištění dat: Praktické zkušenosti

- Pokud vaše nová data obsahují více než 30 čísel, tak je v nich skoro jistě nějaká chyba.
- **Čištění a příprava dat zabírá obvykle 80 – 90 % analytikova času.**
- Pokud budete VELMI pečliví v této fázi, ušetříte si daleko víc času a nervů později – jinak stavíte dům na písku.
- GIGO…Garbage in, Garbage out (smetí dovnitř, smetí ven)
  - sebelepší model (proces) nevyrobí ze smetí nic jiného než opět smetí.

Co způsobí nekvalitní data

- Správa nekvalitních/nadbytečných dat
- Nedoručené zásilky (marketing, fakturace)
- Nesprávné výsledky zpracování (reporting, analýzy, data mining)
- Špatné fungování systému (nekompatibilita)
- Ztráta image, nespokojení klienti

Co způsobí nekvalitní data

- Při mailingové kampani jedné britské **maloobchodní společnosti** se ukázalo, že jedna pětina oslovených už zemřela. Přesto (nebo pro to?) byli obesláni s pozdravným oslovením „**Drahý pane Zesnulý**“. [1)]
- Jistá **pojišťovna** zjistila, že většina jejich zákazníků má **zaměstnání „Astronaut**“ – další pátrání ukázalo, že „Astronaut“ je první volba v seznamu v jejich CRM systému. [1)]
- 44 000-98 000 Američanů ročně umírá na základě **odvratitelné medicínské chyby** jako přepsání při psaní receptu, špatně popsaný výsledek krevní zkoušky, nečitelná informace v pacientských záznamech atd. Je to **osmá nejčastější příčina úmrtí** v USA [2)]
- 7.5.1999 bombardovaly **ozbrojené síly USA** čínské velvyslanectví v Jugoslávii. Vyšetřování zjistilo: CIA používá zastaralý mapový materiál; ještě k tomu pracovník předložil v důsledku chyby v datech **špatnou adresu** – „Doslovně nakreslil X na nesprávné místo“ [3)]

1) Peel, M: Letters to the dead and other data dereliction. © 2007 Financial Times Deutschland. http://www.ftd.de, vydání z 2.10.2007
2) Oash, J. (1999): IT Can Reduce Medical Errors. Obsaženo v: Wang, Pierce, Madnick: Information Quality, 2005
3) BBC: Americas chinese embassy warning ignored. © 1999 BBC. http://news/bbc.co.uk/1/hi/world/americas/37775.stm, vydání z 2.10.2007

Čištění dat: Průzkum proměnných

- Nabývá přípustných hodnot (x out of range)?
- „Divné“ kódy („xxx“, „9999“…)
- Duplicitní kódy pro stejnou věc („Ž“, „ž“, „žena“, „zena“…)
- Kódování češtiny/ruštiny/…
- Překlepy apod.
  - Editovací distance (Levenshteinova (Владимир Иосифович Левенштейн), …) pomohou odhalit překlep
  - Editovací distance = počet elementárních editovacích kroků potřebných pro změnu jednoho řetězce na druhý. Viz http://www.merriampark.com/ld.htm k Levenshteinově distanci
    - Je zde aplet, který ji umí počítat
  - Shlukování řetězců podle ED

Čištění dat: Průzkum proměnných

- Slučování podobných kategorií (prodavač – prodejce – prodavačka);
- Málo četné kategorie (národnost brazilská…) – je třeba sloučit/přiřadit k nějaké(kým) více četné(ným) kategorii(ím) na základě nějakého vhodného kriteria.
- Je distribuce přiměřená našemu očekávání (interval hodnot, rozptyl, šikmost, špičatost, modální hodnoty…)? Není např. příliš „ořezaná“ či naopak „roztažená“?
  - Někdy se obtížně poznává: Např. věk v části dat může být kódován jako poslední dvojčíslí roku narození, a v jiné části dat jako *2007 – rok narození*.

Čištění dat: Průzkum proměnných

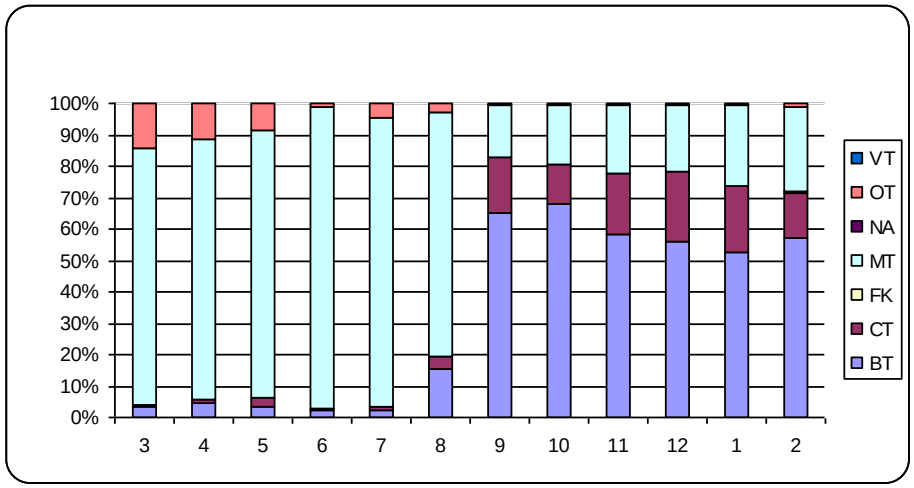
- Shluky (clumping), typicky kolem zaokrouhlených hodnot
  - Příjem –lidé rádi zaokrouhlují směrem nahoru.
  - Nebo třeba kolem hranic věkových kvót, vzniklé tím, jak tazatelé „upravují“ věky respondentů, aby se vešli do kvót.
- Chybějící hodnoty (příčiny vzniku, zastoupení,…)!!!
- Pozor na kódy časů (amer. x evrop. konvence), regionů apod.!

Čištění dat: Vazby mezi daty

- Více proměnných
  - Kontingenční tabulky, box ploty s kategoriemi, bodové grafy a jejich matice, korelační koeficienty
  - Logické vazby (např. 10tiletý nemůže být ženatý, 30tiletý nemůže pracovat 20let,…)
    - Hledání pomocí programu/kódu – podmínky vyjádříme pomocí prostředků matematické logiky a necháme počítač, aby vyhledal případy, kde nejsou splněny.
  - Extrémní hodnoty vícerozměrného rozdělení
    - Bodový graf
    - Mahalanobisova vzdálenost od těžiště: $[(\boldsymbol{x}-\boldsymbol{t})^T \boldsymbol{S}^{-1} (\boldsymbol{x}-\boldsymbol{t})]^{-1/2}$, kde $\boldsymbol{t}$ je vektor těžiště, $\boldsymbol{x}$ zkoumaný bod a $\boldsymbol{S}$ kovarianční matice
      - např. P. Filzmoser (2004) A multivariate outlier detection method, http://www.statistik.tuwien.ac.at/public/filz/papers/minsko4.pdf
  - Další vlastnosti; např. existují očekávané korelace?

Čištění dat: Vazby mezi daty

- korektní vkládání dat do DB
  - text. pole s názvem zboží vs. rolovací seznam s typem zboží



  - pořadí hodnot v rolovacím seznamu – problém první (defaultní) hodnoty

Explorační analýza – PROČ?

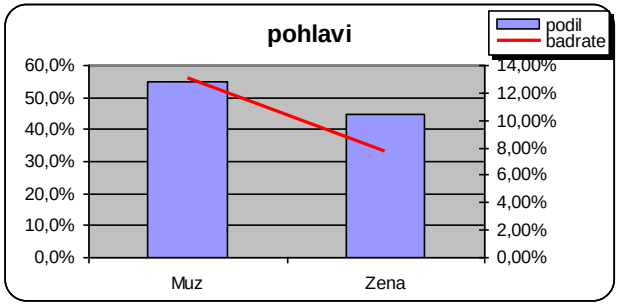
- Je třeba pochopit data:
  - najít chyby v datech
  - najít vzory v datech
  - najít porušení statistických předpokladů, testování hypotéz
  - …a především proto, že pokud to neuděláme, budeme mít velké problémy později.

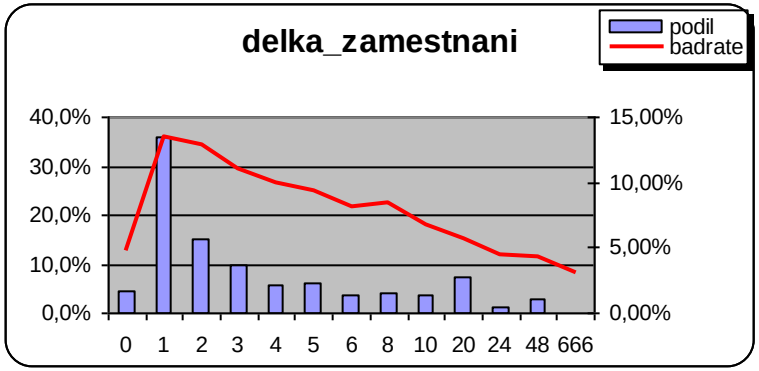
Explorace dat - jednorozměrná

- Frekvenční tabulky, histogramy:

	pocet	podil	badrate
Muz	248 768	55,0%	13,08%
Zena	203 194	45,0%	7,69%
Total	451 962	100,0%	10,66%

delka_zamestnani	pocet	podil	badrate
0	20 825	4,6%	4,69%
1	163 144	36,1%	13,43%
2	67 462	14,9%	12,80%
3	43 778	9,7%	10,97%
4	26 256	5,8%	10,01%
5	27 526	6,1%	9,32%
6	15 893	3,5%	8,16%
8	18 036	4,0%	8,39%
10	17 195	3,8%	6,72%
20	33 641	7,4%	5,60%
24	5 176	1,1%	4,48%
48	12 934	2,9%	4,28%
666	96	0,0%	3,13%
Total	451 962	100,0%	10,66%

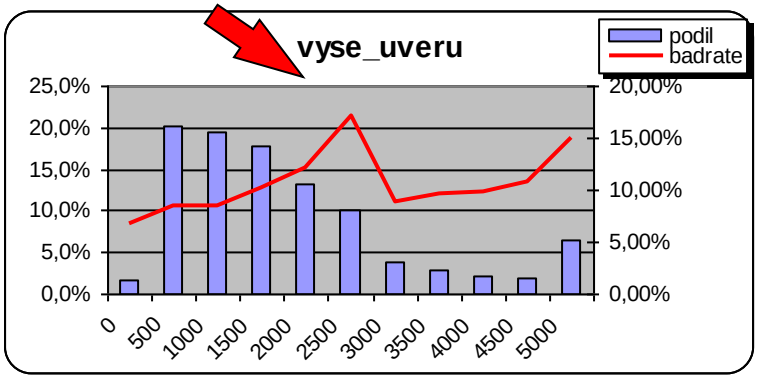




Explorace dat - jednorozměrná

- výše úvěru vs. cílová proměnná (bad rate).
  - je třeba vysvětlit veškeré „nestandardní“ závislosti
  - úplné pochopení dat vede k interpretovatelným modelům s vysokou prediktivní silou

OK? Nebo je to způsobeno jiným faktorem???



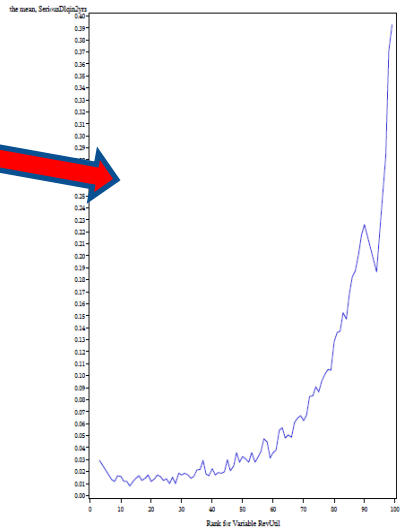
Explorace dat - jednorozměrná

- spojité proměnné:
  - průměr
  - modus
  - kvantily
  - rozptyl
  - min./maximální hodnota
  - vztah k cílové proměnné

Variable	Mean	Std Dev	Variance	Minimum	Maximum	N Miss	1st Pctl	5th Pctl
Bo_Age	36.7928463	10.0283282	100.5673670	18.0000000	99.0000000	0	21.0000000	23.0000000
Ln Orig	153467.57	68370.61	4674539752	19600.00	599000.00	0	42750.00	62000.00
Orig_LTV_Ratio_Pct	93.0798522	8.8537162	78.3882906	20.0000000	111.0000000	0	63.0000000	80.0000000
Credit_score	687.6683165	62.9002322	3956.44	440.0000000	999.0000000	0	527.0000000	575.0000000
Tot_mthly_incm	5024.71	2952.16	8715254.06	500.0000000	65000.00	0	1473.00	2000.00
Median_state_inc	44945.07	5431.51	29501323.98	32589.00	57352.00	0	33948.00	38550.00
DTI_Ratio	0.3747207	0.1758619	0.0309274	0	3.4280770	0	0	0
orig_apprd_val_amt	170661.44	81775.07	6687162030	0	870000.00	0	47000.00	68000.00
pur_prc_amt	164681.56	79719.84	6355252570	20000.00	870000.00	0	45000.00	65000.00
Tot_mthly_debt_exp	1745.46	1089.20	1186348.36	0	17225.00	0	0	0

Variable	10th Pctl	Lower Quartile	Median	Upper Quartile	90th Pctl	95th Pctl	99th Pctl
Bo_Age	25.0000000	30.0000000	37.0000000	41.0000000	50.0000000	56.0000000	69.0000000
Ln_Orig	75000.00	103500.00	141500.00	190950.00	255000.00	285600.00	322700.00
Orig_LTV_Ratio_Pct	80.0000000	90.0000000	95.0000000	100.0000000	100.0000000	100.0000000	102.0000000
Credit_score	606.0000000	647.0000000	688.0000000	737.0000000	769.0000000	783.0000000	801.0000000
Tot_mthly_incm	2402.00	3245.00	4632.00	6000.00	7975.00	9611.00	14830.00
Median_state_inc	39000.00	40171.00	43988.00	49894.00	53275.00	56763.00	56772.00
DTI_Ratio	0.1738380	0.2778700	0.3761930	0.4710300	0.5746010	0.6383500	0.8506940
orig_apprd_val_amt	82000.00	113000.00	154000.00	214000.00	285000.00	327000.00	415000.00
pur_prc_amt	78000.00	108900.00	148650.00	205000.00	275512.00	319900.00	405000.00
Tot_mthly_debt_exp	655.0000000	1073.00	1578.00	2253.00	3008.00	3604.00	5290.00

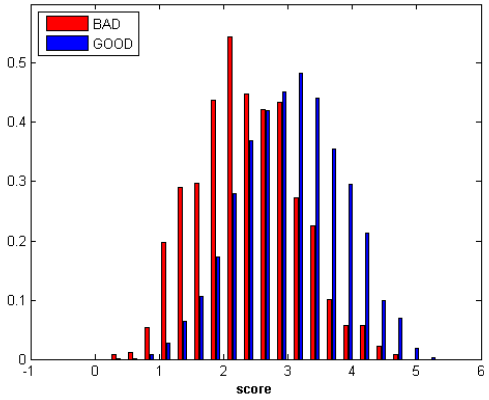
- často je vhodná kategorizace (následně frekvenční tabulky, vztah k cílové proměnné)

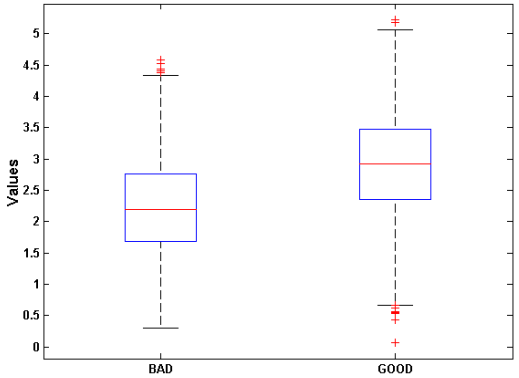


- U dichotomické cílové prom. (0/1) jde o relativní zastoupení vybrané kategorie (např. bad rate) pro vhodné intervaly zkoumané proměnné. Intervaly můžou být:
  - pevně dané, např. 0-10,10-20,...
  - decily/percentily
  - klouzavé okno

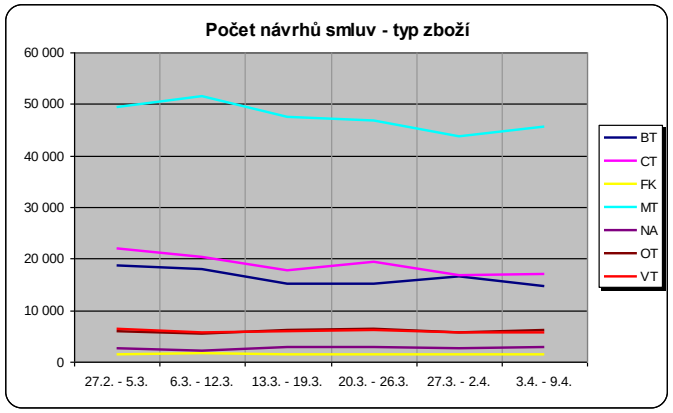
Explorace dat - jednorozměrná

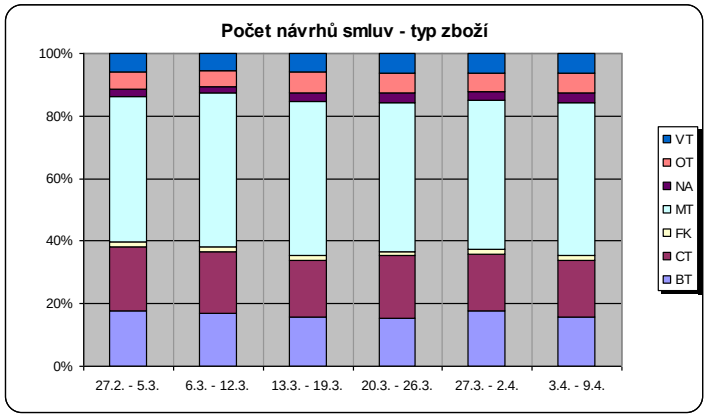
- Histogramy, box ploty





- Stabilita v čase





Explorace dat - vícerozměrná

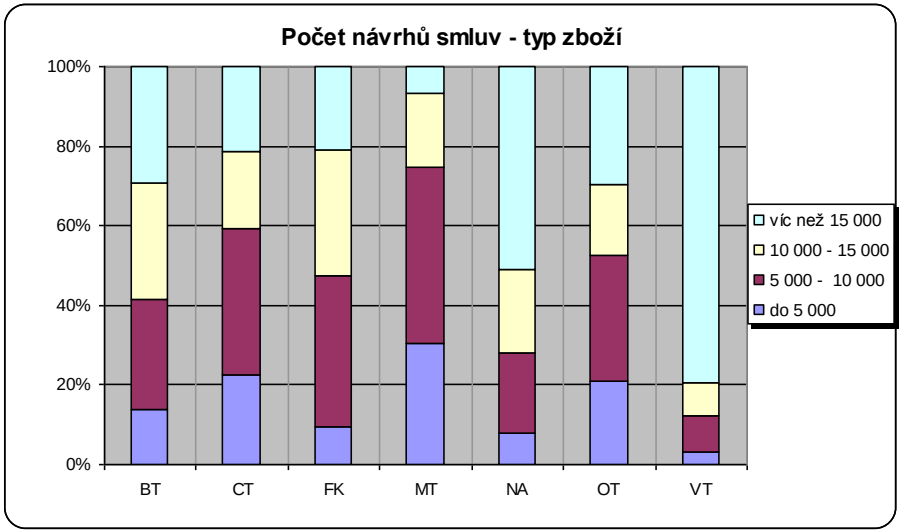
- Kontingenční tabulky
  - absolutní četnosti slouží ke kontrole jestli některá kombinace hodnot není příliš málo četná
  - relativní četnosti (řádkově + sloupcově podmíněné) slouží k odhalení vztahů mezi proměnnými

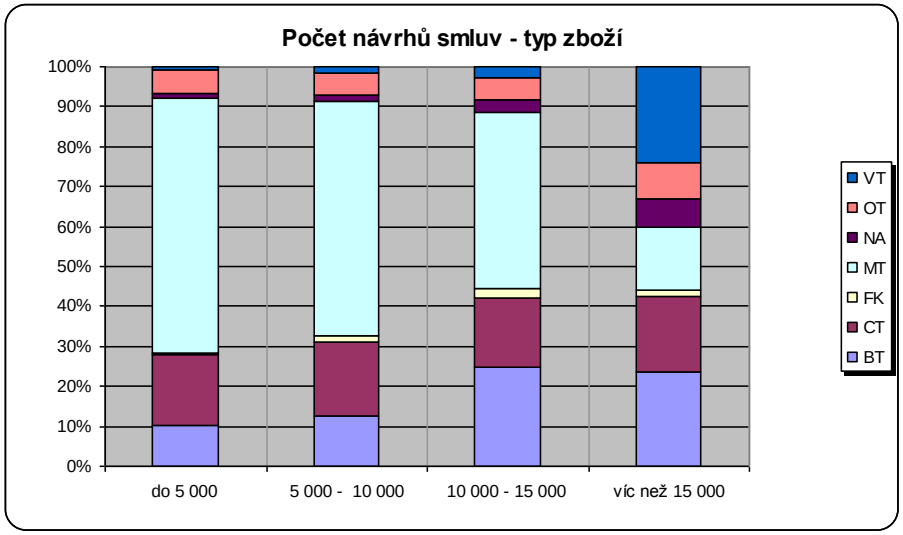
	do 5 000	5 000 - 10 000	10 000 - 15 000	víc než 15 000
BT	4 291	8 581	9 176	9 044
CT	7 587	12 493	6 500	7 236
FK	258	1 017	851	557
MT	27 191	39 551	16 524	5 992
NA	426	1 088	1 114	2 737
OT	2 478	3 689	2 103	3 475
VT	384	1 001	963	9 086

row%	do 5 000	5 000 - 10 000	10 000 - 15 000	víc než 15 000
BT	13,8%	27,6%	29,5%	29,1%
CT	22,4%	36,9%	19,2%	21,4%
FK	9,6%	37,9%	31,7%	20,8%
MT	30,5%	44,3%	18,5%	6,7%
NA	7,9%	20,3%	20,8%	51,0%
OT	21,1%	31,4%	17,9%	29,6%
VT	3,4%	8,8%	8,4%	79,5%

col%	do 5 000	5 000 - 10 000	10 000 - 15 000	víc než 15 000
BT	10,1%	12,7%	24,6%	23,7%
CT	17,8%	18,5%	17,5%	19,0%
FK	0,6%	1,5%	2,3%	1,5%
MT	63,8%	58,7%	44,4%	15,7%
NA	1,0%	1,6%	3,0%	7,2%
OT	5,8%	5,5%	5,6%	9,1%
VT	0,9%	1,5%	2,6%	23,8%

Explorace dat - vícerozměrná



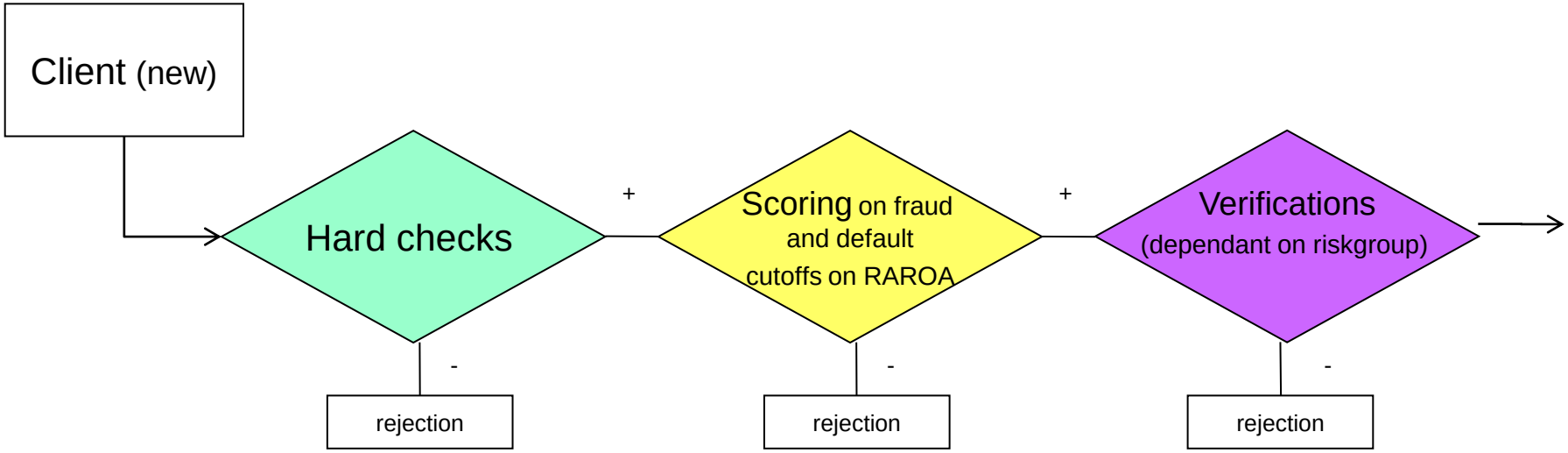


Implementace

Implementace

- Analytik sepíše požadavek a předá IT
  - SW model (SAS,…) -> papír -> schvalovací kolečko -> PL SQL,…
- Implementace s asistencí IT
  - Analytik sám vytvoří schvalovací kód (PL SQL, DLL,…), ale jakákoli úprava podléhá schválení a asistenci IT.
- Implementace plně pod kontrolou odd. risku

Schvalovací proces



Policy declines – low age, unsufficient length of employment, “terorrist” etc.

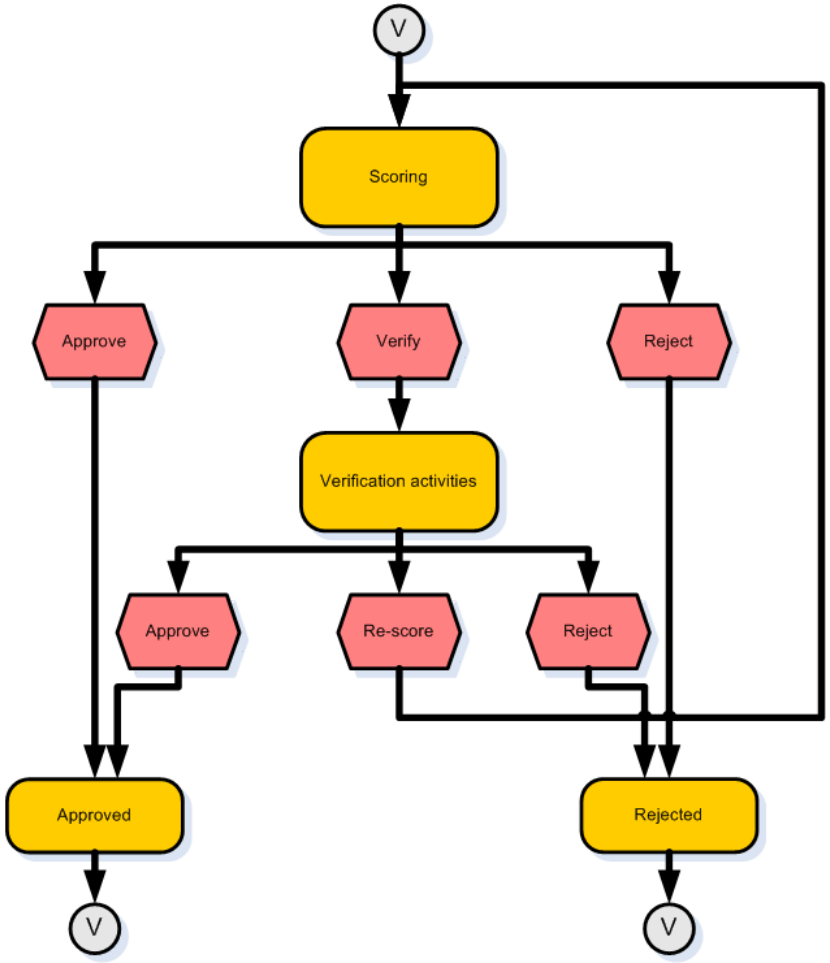
What is the probability that client will pay?

Will the contract be profitable?

Is the number of client‘s phone valid? Etc.

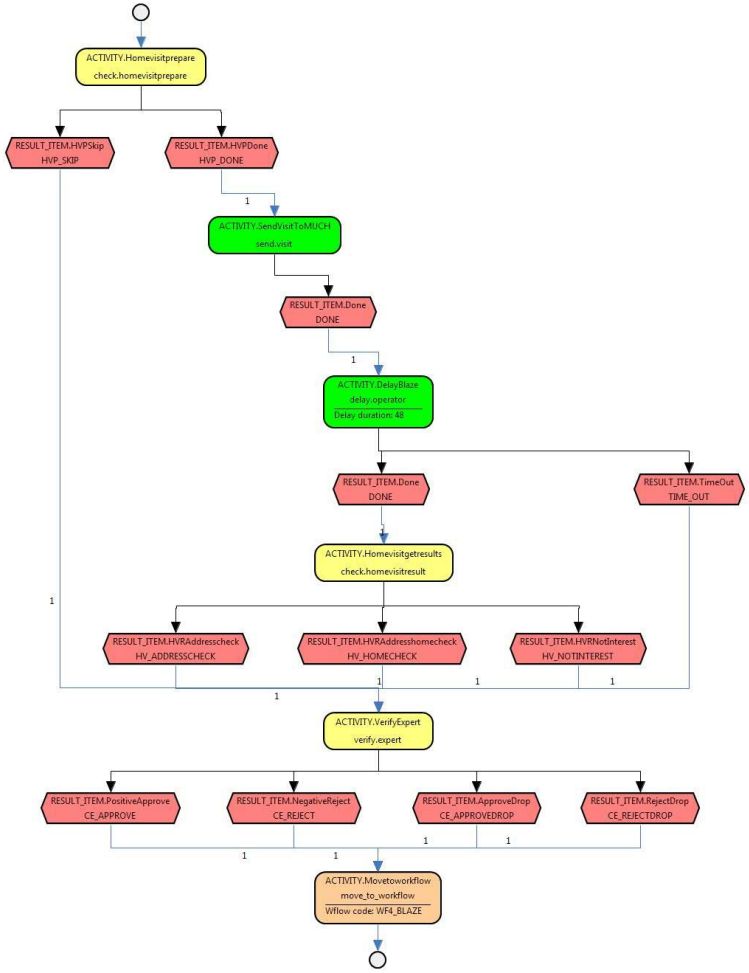
Schvalovací proces

- Identifikace žadatele, řešení problémů s duplicitami
- Nápočty prediktorů – na datovém skladu, online
- Schvalovací proces – sekvence aktivit vedoucí ke schválení nebo zamítnutí smlouvy
- Provádění verifikací – Call centrum
- Automatické kontroly – externí zdroje, skóring
- Ve špičce až desítky tisíc schvalování za hodinu (Rusko, Čína)



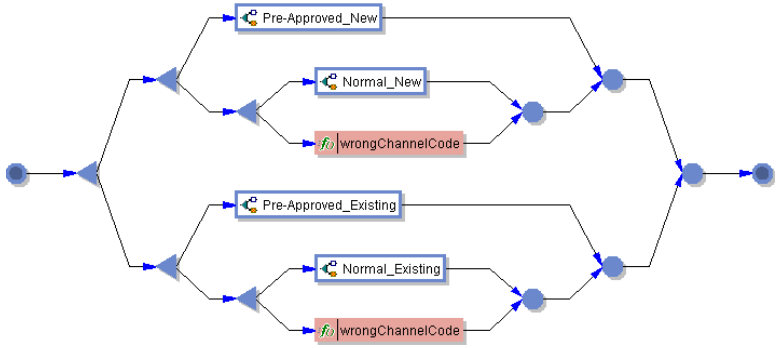
Workflow

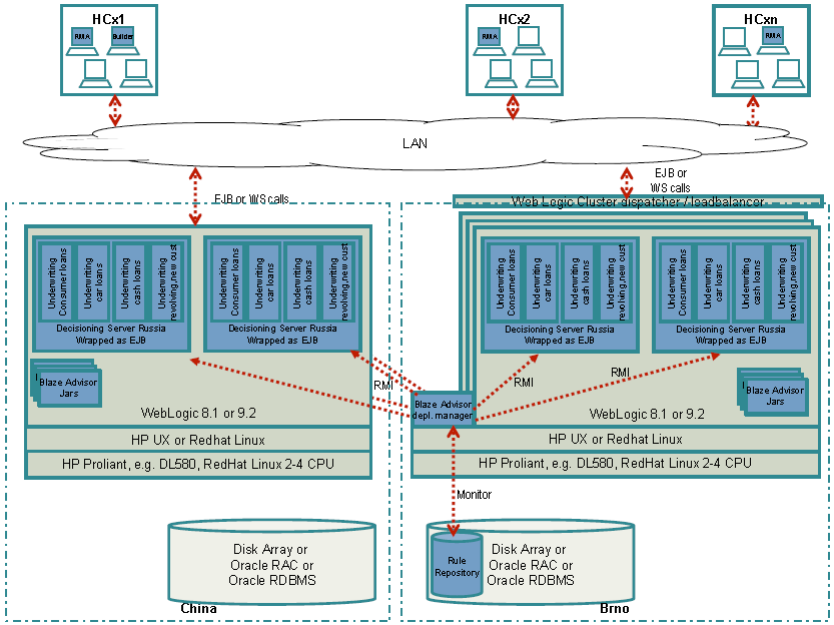
- Sekvence několika aktivit definovaných schvalovacím procesem
  - Manuální aktivity – v Call centru operátor vyplní výsledek
  - Automatické – výsledek aktivity zvolí systém
  - Speciální – čekání, přechod mezi workflow
- Grafický editor pro snadnou modifikaci atributů
- Automatická část a call centrum – samostatné aplikace, souběžný běh na více serverech



Skórování žádosti

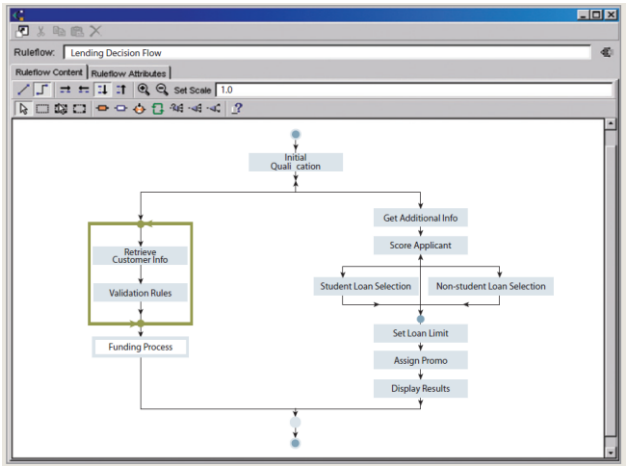
- Tvrdé kontroly a další pravidla
- Skóringové funkce – typ klienta, skóring
- Logistická regrese
- Rozhodnutí jak pokračovat dál – schválení, zamítnutí, jiné workflow
- Snadná modifikace všech nastavení
- Samostatná Java aplikace spouštěná schvalovacím procesem, běží paralelně na dvou serverech
- Volá se několikrát pro každý schvalovací proces

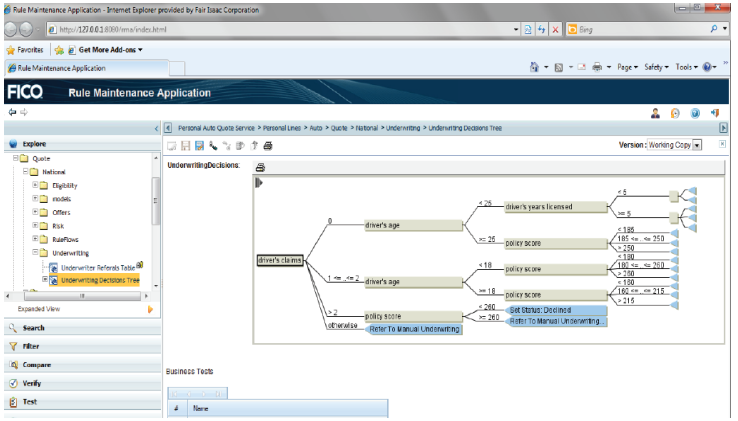




FICO™ Blaze Advisor®

- Business objectives:
  - Modifikace nastavení schvalovacího procesu bez zásahu IT
  - …
- Technical objectives:
  - Zpracování 500 000 žádostí/ hod.
  - Zpracování 1 000 000 požadavků skórovacího modulu/ hod.
  - Podpora Champion/challenger
  - Online i batch mode
  - …





Fungování modelu v praxi

Monitoring scoringových modelů

- Není překvapivé, že prediktivní modely se ve statistickém slova smyslu chovají nejlépe na vývojovém vzorku dat. Výstupy těchto modelů, např. skóre nebo rating klienta, jsou počítány pomocí jistých vzorců, jejichž koeficienty příslušející nezávislým proměnným (prediktorům) jsou odvozeny na datech vývojového vzorku. Posun distribuce výstupu daného modelu je pak zapříčiněn právě změnou vstupních hodnot modelu, tj. prediktorů, v průběhu času. V podstatě ihned (alespoň většinou) po nasazení prediktivního modelu do praxe dochází k jistému poklesu jeho prediktivní síly, který je způsoben určitou změnou vstupních hodnot modelu. Zásadní je v praxi nastavení takových procesů, které odhalí, že se tak děje, proč se tak děje a jak vážný problém to ve svých důsledcích znamená.

Monitoring scoringových modelů

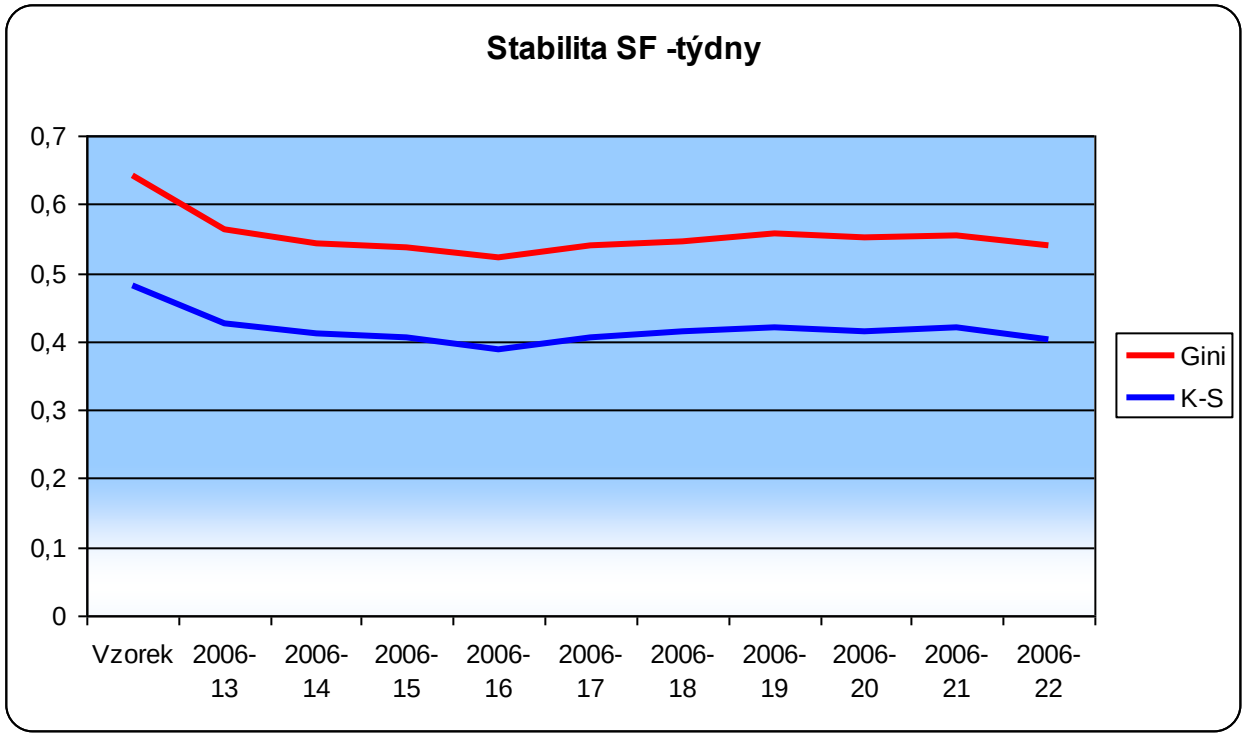
- Faktorů způsobujících posun v distribuci prediktorů, a následně posun v distribuci výstupu prediktivního modelu, je několik:

  - Přirozený posun v datech/změna demografické struktury dat
  - Databázové chyby
  - Změna datového zdroje
  - Změna definice/formátu vstupních dat
  - Změna datového univerza
  - Ostatní (organizovaný podvod,…)

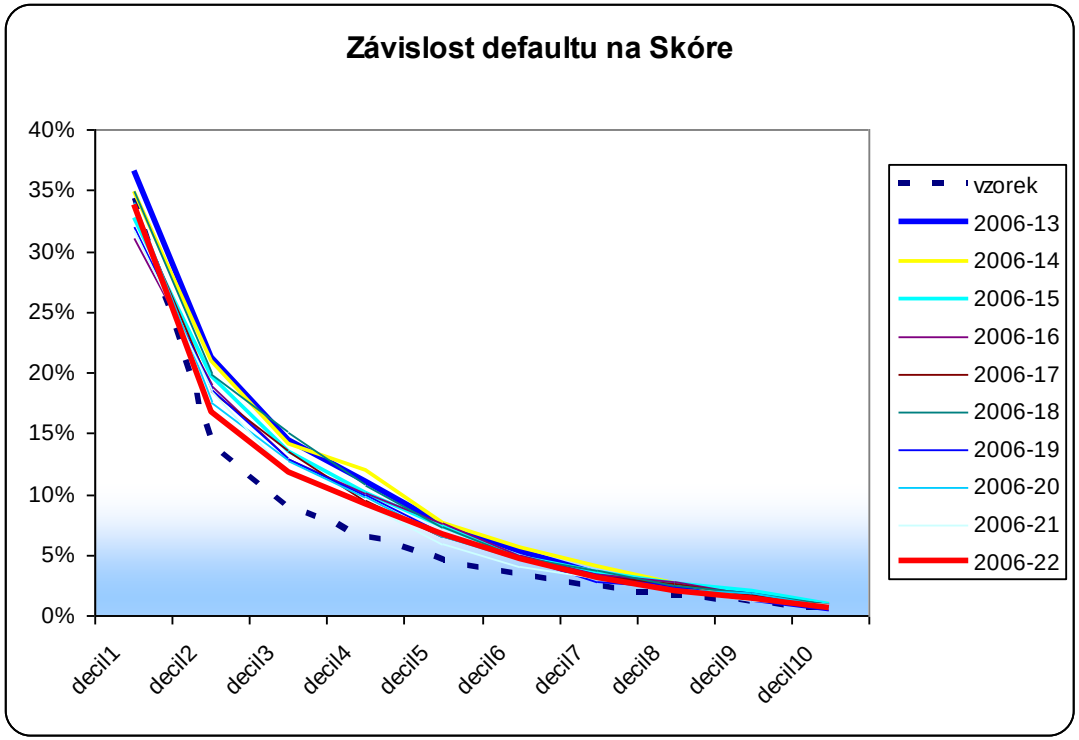
- Typickým příkladem prvního uvedeného důvodu je příjem klienta (všeobecným trendem je růst příjmu populace). Změnou definice/formátu vstupních dat je myšlena například situace, kdy je rozšířen číselník hodnot, kterých může vstupní proměnná nabývat. Změnou datového univerza je myšlen případ kdy je vyvinutý prediktivní model použit např. pro odlišný/nový segment portfolia nebo odlišný/nový produkt.

Monitoring scoringových modelů

- K-S, Gini:



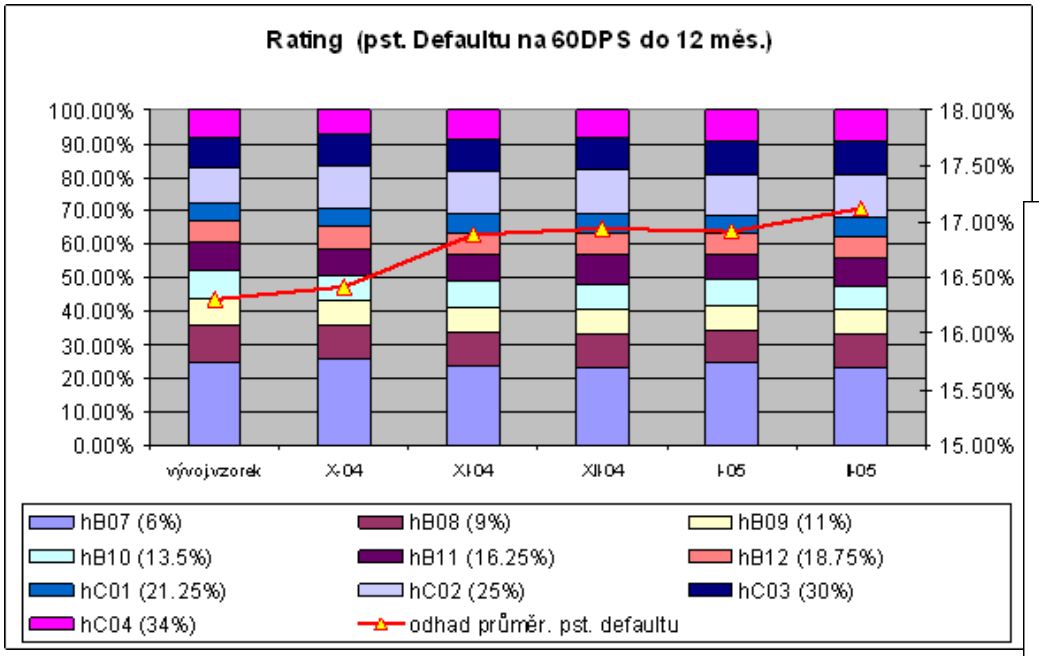
Monitoring scoringových modelů

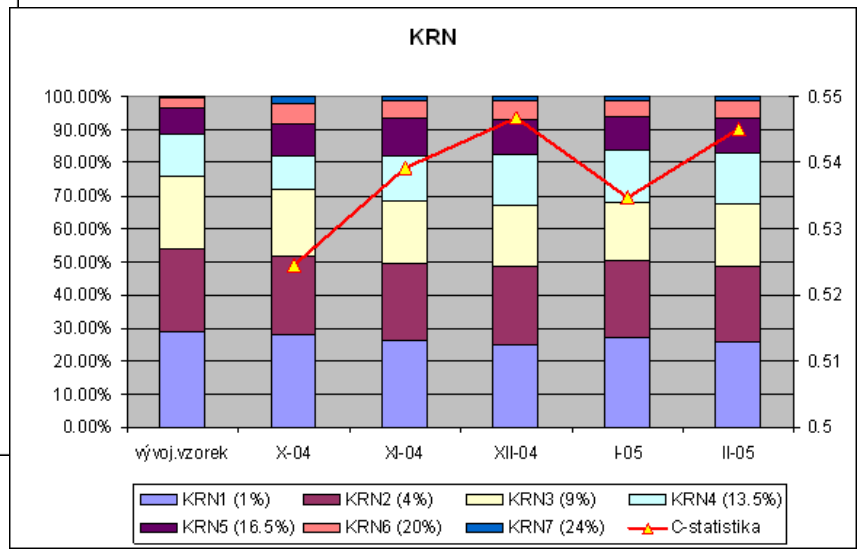


- Čím strmější křivka tím lépe.
- V průběhu času se zplošťuje – jde o to, jak moc.

Monitoring scoringových modelů

- c-statistika:





Monitoring scoringových modelů

- Chceme posoudit zda se distribuce skóre na vývojovém vzorku liší od distribuce skóre v daném časovém intervalu:

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i} \qquad PSI = \sum_{i=1}^{r} (O_i - E_i) \ln\left(\frac{O_i}{E_i}\right)$$

	výv. vzorek [1]	týden1 [2]	[3]=[2] -[1]	[4]=[2]/[1]	[5]=ln[4]	[6]=[3]*[5]
skóre_1	10,00%	5,63%	-0,044	0,563	-0,574	0,025
skóre_2	10,00%	11,21%	0,012	1,121	0,114	0,001
skóre_3	10,00%	11,00%	0,010	1,100	0,095	0,001
skóre_4	10,00%	10,97%	0,010	1,097	0,092	0,001
skóre_5	10,00%	10,31%	0,003	1,031	0,031	0,000
skóre_6	10,00%	10,12%	0,001	1,012	0,012	0,000
skóre_7	10,01%	9,62%	-0,004	0,961	-0,039	0,000
skóre_8	10,00%	9,89%	-0,001	0,989	-0,011	0,000
skóre_9	10,00%	10,31%	0,003	1,031	0,030	0,000
skóre_10	10,00%	10,94%	0,009	1,095	0,091	0,001
					PSI	0,030

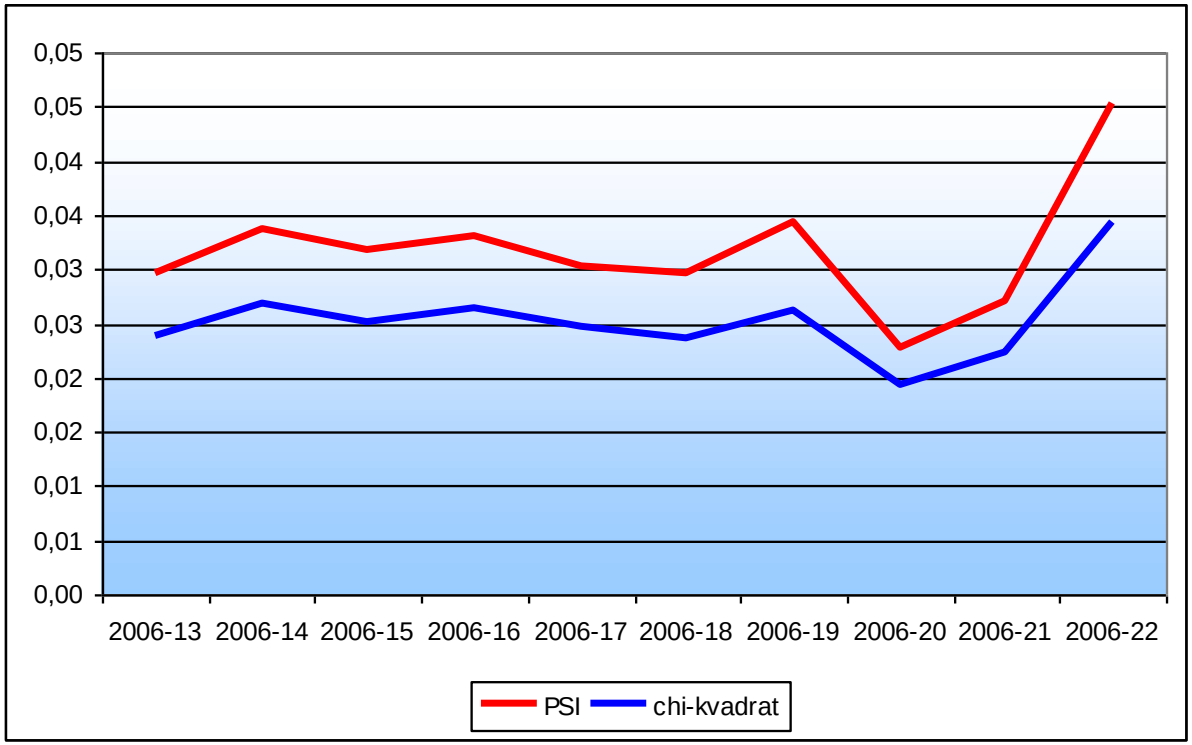
Monitoring scoringových modelů

$PSI \leq 0{,}1$ značí žádný nebo jen velmi malý rozdíl daných distribucí skóre.

$0{,}1 < PSI \leq 0{,}25$ znamená, že došlo k nějakému posunu distribuce, nicméně nikterak významnému.

$PSI > 0{,}25$ signalizuje významný posun v distribuci skóre, tj. zamítáme hypotézu o shodě daných distribucí.

Monitoring scoringových modelů

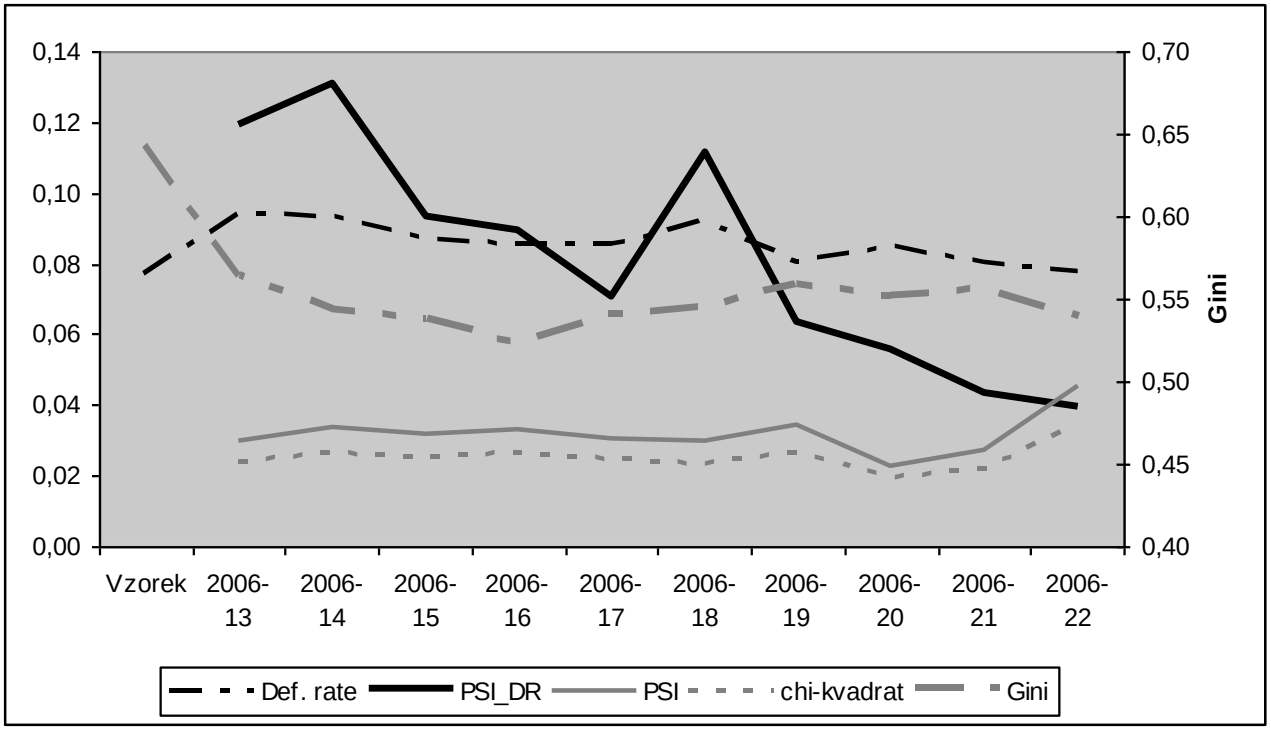


Monitoring scoringových modelů

$$PSI_{DR} = \sum_{i=1}^{r} (DR2_i - DR1_i) \ln\left(\frac{DR2_i}{DR1_i}\right)$$

	def_rate	Gini	PSI_DR	PSI	chi-kvardat
vzorek	7,69%	0,643			
200613	9,38%	0,564	0,120	0,030	0,024
200614	9,35%	0,542	0,131	0,034	0,027
200615	8,70%	0,537	0,093	0,032	0,025
200616	8,57%	0,523	0,089	0,033	0,026
200617	8,59%	0,540	0,071	0,030	0,025
200618	9,19%	0,544	0,111	0,030	0,024
200619	8,03%	0,558	0,063	0,034	0,026
200620	8,52%	0,552	0,055	0,023	0,019
200621	8,05%	0,555	0,043	0,027	0,022
200622	7,76%	0,539	0,039	0,045	0,034

Monitoring scoringových modelů



Champion-challenger (mistr – vyzyvatel)

- K rozšíření využití strategie champion-challenger došlo v devadesátých letech minulého století. Princip je velmi jednoduchý. Předpokládejme, že existuje nějaký způsob dělání něčeho (např. aktuálně používaný scoringový model pro schvalování/zamítání žádostí o úvěr). Tento způsob nazveme mistrem (champion). Nicméně existují další, jeden nebo více, alternativní způsoby jak dosáhnout téhož (nebo velmi podobného) cíle. Tyto nazveme vyzyvateli (challengers). Na náhodném vzorku otestujeme vyzyvatele a porovnáme s mistrem. To nám umožní nejen porovnat efektivnost vyzyvatelů a mistra, ale získáme možnost identifikovat existenci a rozsah vedlejších efektů. Výsledkem pak může být zjištění, že některý z vyzyvatelů je lepší než mistr a tento vyzyvatel se stane novým mistrem.

Segmentace

- 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

Segmentace

- 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)

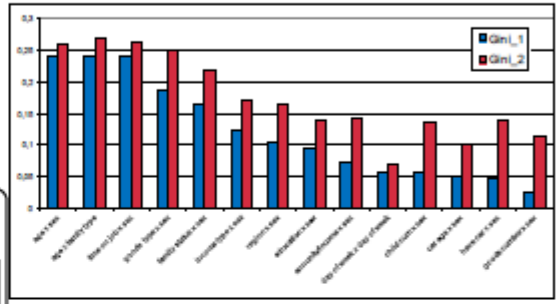
		Gini	improvement (compared to base model without segmentation)
Defaulters	unseg.	0.423	
	7 seg.	0.444	4.96%
	29 seg.	0.465	9.93%
	70 seg	0.513	21.28%
	37 seg	0.472	11.58%
Fraudsters	unseg.	0.597	
	7 seg.	0.625	4.69%
	29 seg.	0.664	11.22%
	70 seg	0.662	10.89%
	37 seg	0.649	8.71%

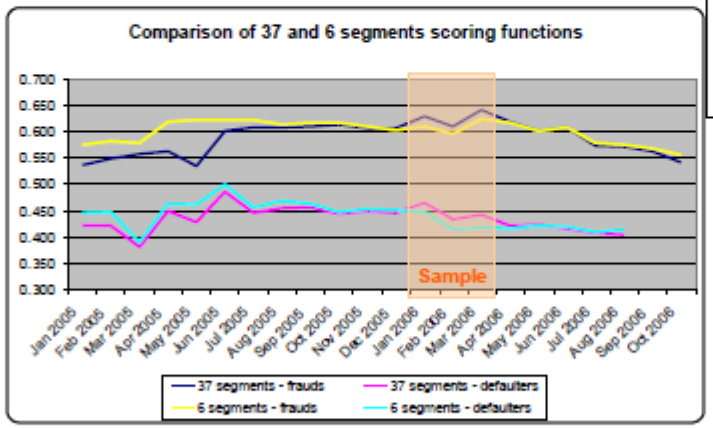
Segmentace

- 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

Segmentace

- 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

Děkuji za pozornost.