Matematické modelování úvěrového rizika v praxi

Mgr. Jiří Tesař (Home Credit, a.s.),
Mgr. Martin Řezáč, Ph.D. (PřF MU Brno)

Brno, 20.4.2010

Obsah

PPF a Home Credit Group

PPF Group

Růst na domácím trhu
(Česká republika)
1991-98

Expanze na
regionální trhy
(Střední a východní Evropa)
1999 - 2003

Globalizace
(SNS a Asie)
Od r. 2004 do současnosti

- Mezinárodní investiční skupina ve střední a východní Evropě
- Aktiva > **10 miliard eur** (ke dni 30. června 2009)
- Oblasti zájmu:
  - finanční služby (bankovnictví, spotřebitelské financování, pojištění, …)
  - investice do nemovitostí
  - vyhledávání investičních příležitostí na vznikajících trzích
- *Více o PPF Group: www.ppf.eu*

Home Credit Group

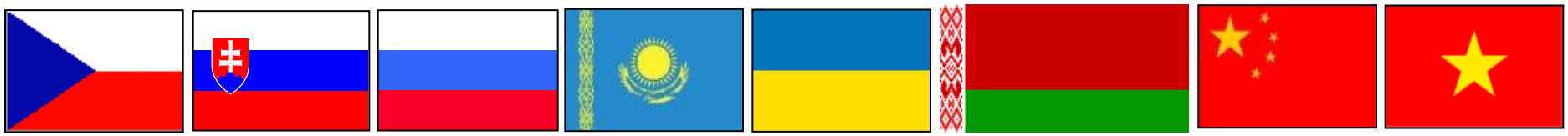
- **Přední poskytovatel spotřebitelského financování ve střední a východní Evropě**

- **Strategie Home Credit Group**
  - disciplinovaný růst
  - dlouhodobý nárůst zisku
  - stabilní správa rizik

- **Společnost Home Credit International**
  - poradenství a služby v oblasti IT
  - strategické řízení jednotlivých společností skupiny

Skupina Home Credit

- Významný poskytovatel spotřebitelského financování
- 14 200 zaměstnanců, více než 5,7 milionu zákazníků (údaj ke dni 30. června 2009)
- Působnost ve státech střední a východní Evropy a Asie :
  - **Česká republika** (Home Credit a.s., od roku 1997)
  - **Slovensko** (Home Credit Slovakia, a.s., od roku 1999)
  - **Ruská federace** (OOO Home Credit & Finance Bank, od roku 2002)
  - **Kazachstán** (AO Home Credit Bank, od roku 2005)
  - **Ukrajina** (OAO Home Credit Bank, od roku 2006)
  - **Bělorusko** (OAO Home Credit Bank, od roku 2007)
  - **Čína** (HC Asia N.V., od roku 2007)
  - **Vietnam** (PPF Vietnam Finance Company Ltd., od roku 2009)
- Více o skupině Home Credit: www.homecredit.net



Home Credit po produktech

SPOTŘEBITELSKÉ ÚVĚRY

- Home Credit / 71 % populace ČR
- konkurence získala například:
  - Česká spořitelna 34%
  - Cetelem 42%
  - GE Money Multiservis 52%

REVOLVINGOVÉ ÚVĚRY (KREDITNÍ NEBO ÚVĚROVÉ KARTY)

Home Credit / 45 % populace ČR

- konkurence získala například:
  - Česká spořitelna 76%
  - Cetelem 28%
  - GE Money Multiservis 34%

HOTOVOSTNÍ PŮJČKY

- Home Credit / 35 % populace ČR
- konkurence získala například:
  - Česká spořitelna 74%,
  - Cetelem 26%
  - GE Money Multiservis 21%

Absolventi MU v HC

- **Studijní obor**: Matematika nebo matematika – ekonomie

- **Počty absolventů v HC a HCI**:

Matematika	10
Matematika – ekonomie	8

- **Oddělení:**

  - Řízení rizik HC
  - Řízení rizik HCI
  - Ostatní oddělení

  - Celkem : cca 20 zaměstnanců

Přednáška pro studenty

Prezentace HC a Odboru řízení rizik - posílení analytických týmů o absolventy a studenty posledních ročníků vysokých škol na pozice:

SPECIALISTA ŘÍZENÍ RIZIK a
ANALYTIK ODD. VYMÁHÁNÍ POHLEDÁVEK

Kdy: 19.3.2009

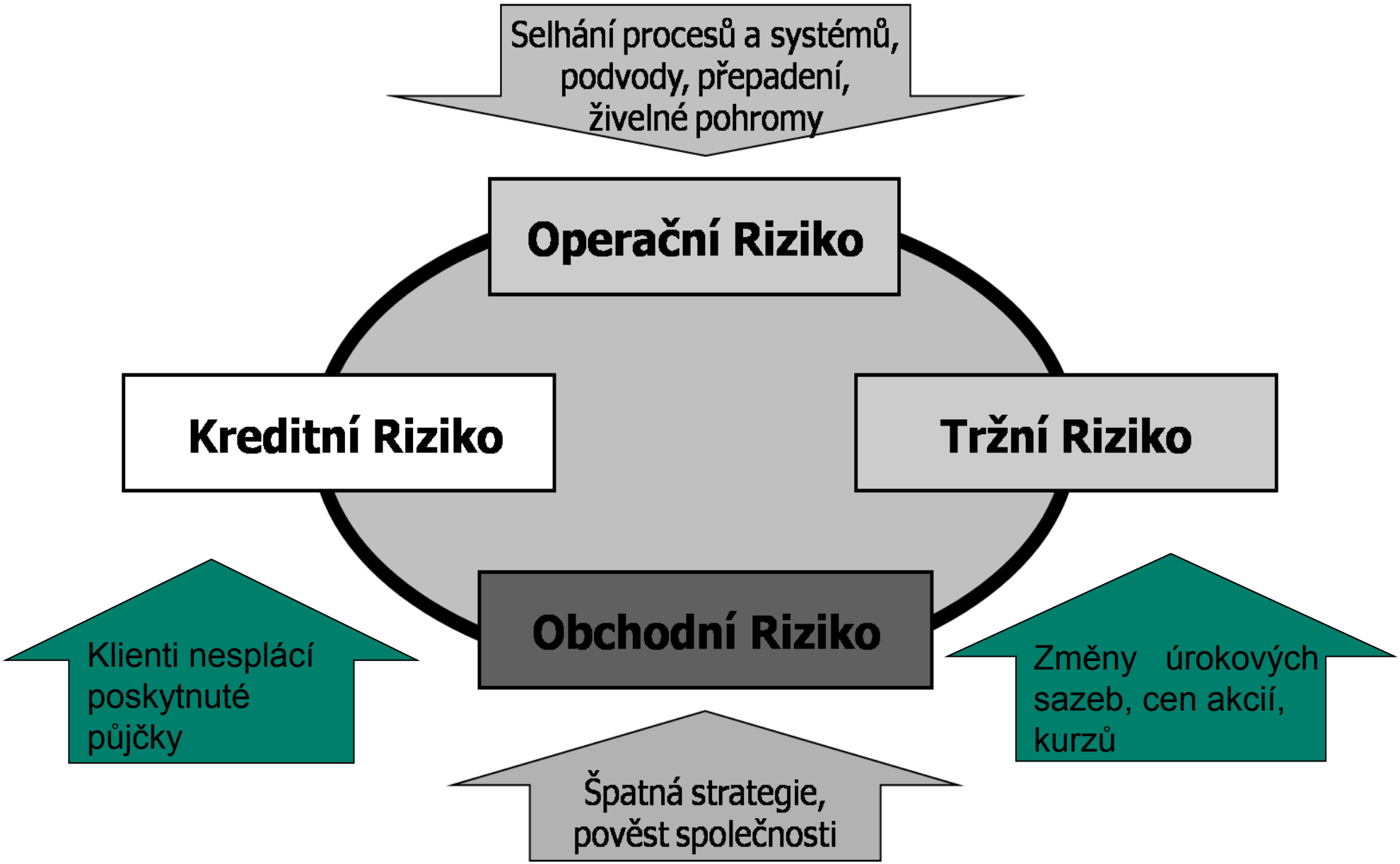
Účast: přibližně 40 studentů Přírodovědecké fakulty

Program:

- představení HC
- Risk management a druhy rizik
- Odbor řízení rizik

Scoring – obecné principy

Risk Management a druhy rizik



Why score?

$$sum\_points = 3.45 + 0.12\,(is\ women) - 0.32\,(is\ only\ 24\ years\ old) + 0.05\,(married) \ldots$$

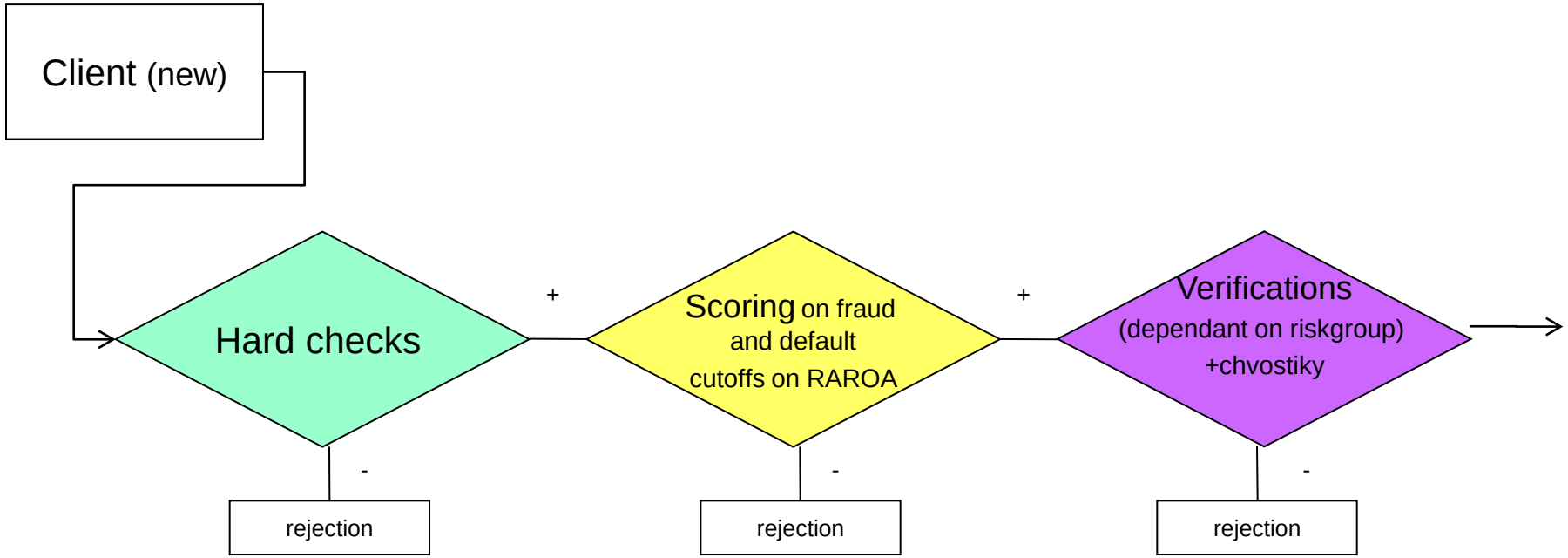
ADVANTAGES:

- Automatization of approval proces
- Cost – effective
- Less fraud possibilities

DISADVANTAGES

- Statistical based, not take in account client like individual

Score in approval process



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

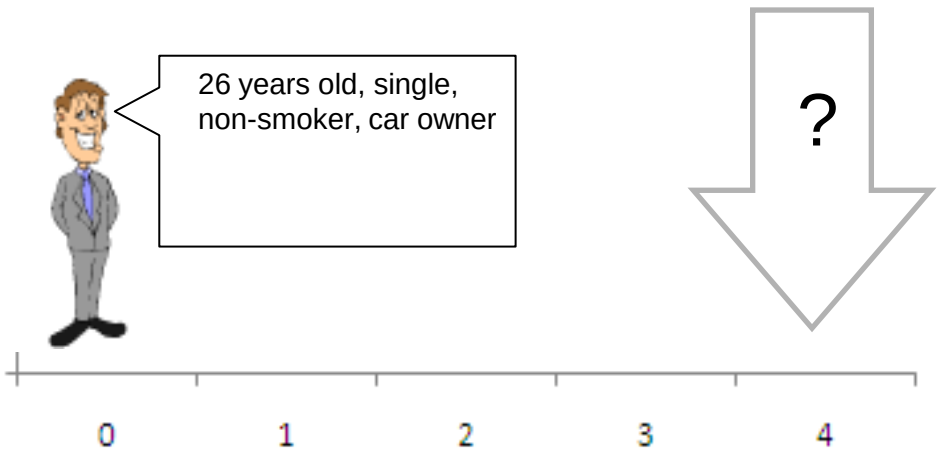
What is the probability that client will pay?

Will the contract be profitable?

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

Score development – which data do we use

Socio-demographic data	Product data	Behavioral data (for already known customers)
• Age • Sex • family status • Income • Profession • …	• Price • Term • Downpayment • …	• Maximum days past due • Number of credits which he already had • Number of instalments past due • …

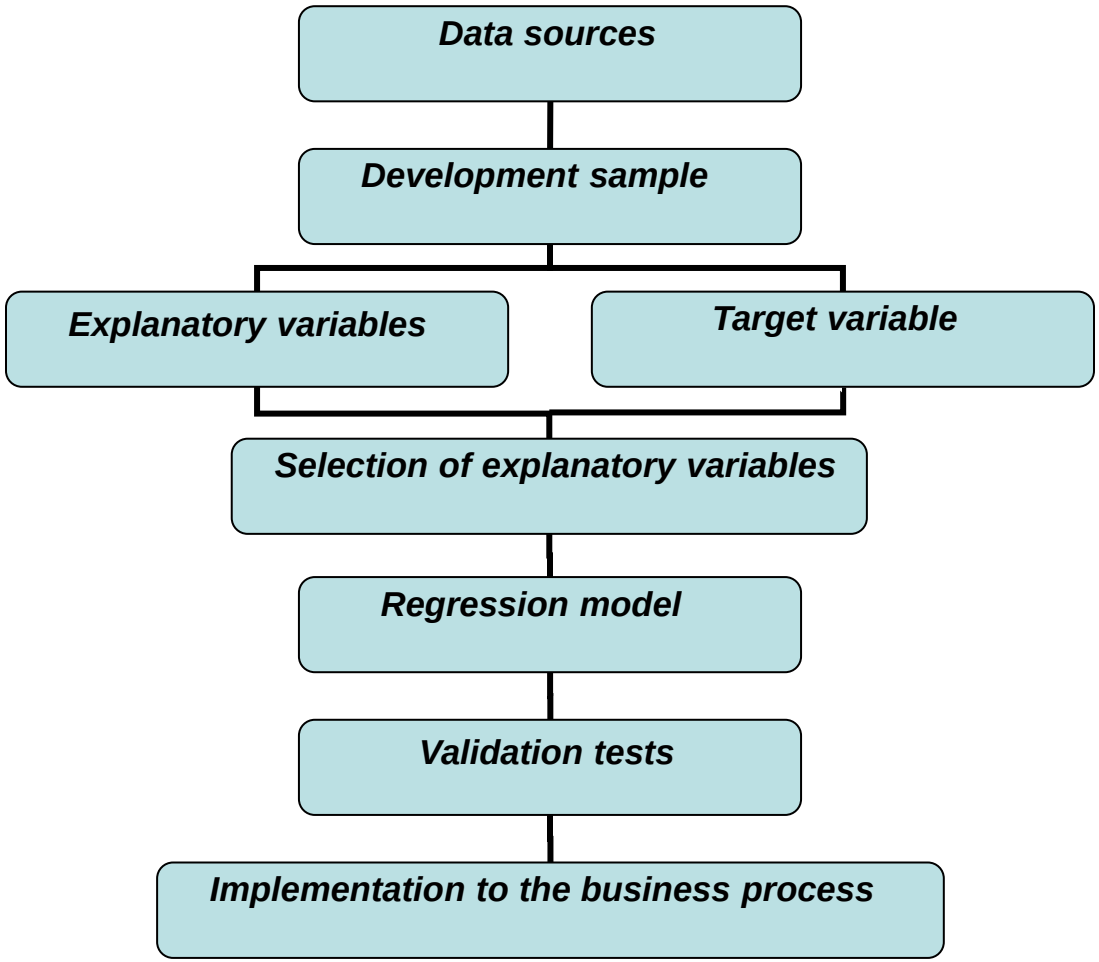


Scoring - Data sample preparation

Main reason for the scorecard development

- to update the existing scorecard

- to reflect the latest available history for the scorecard development



Target variable

The target (or explained) **variable** is a two valued (dichotomous) variable which indicates whether the loan was being repaid properly or not.

Definice dobrého / špatného klienta:
Klient se někdy v průběhu prvních M měsíců po poskytnutí úvěru dostal do zpoždění se splácením aspoň o K měsíců, přitom dlužná částka byla větší než tolerance.

"**Good loans**" – good payment morale

"**Bad loans**" – bad payment morale

"**Unspecified loans**" – neither good or bad payment morale, or the repayment history is too short to decide about payment morale

Requirements for target variable:

A sufficient number of bad loans should be provided.

The sharper contrast between the definition of a good and a bad loan, the better.

Development sample definition

Development time period:
Specify if you define this period by date of ratification or date of first due.

In order to reflect actual economic conditions, the data used for development should be as recent as possible.

Application data are sufficiently homogeneous and similar to the most recent new portfolio.

The chosen period provides enough data for scorecard development.

Development and validation sample:
The data sample was divided into development (70 %) and validation (30 %).

The development and validation of the scorecard should be done on distinct samples.

To test the performance of the model on data from the same period.

Tests should be performed on an out-of-time validation sample, too.

Development sample definition

Structure of the development and validation sample

First installment prescription	Development sample				Validation sample			
	Bad	Good	TOTAL	Bad rate	Bad	Good	TOTAL	Bad rate
	N	N	N	%	N	N	N	%
JUL2007	120	367	487	24.6%	54	139	193	28.0%
AUG2007	166	566	732	22.7%	67	237	304	22.0%
SEP2007	185	587	772	24.0%	74	235	309	23.9%
OCT2007	117	470	587	19.9%	48	199	247	19.4%
NOV2007	109	473	582	18.7%	48	187	235	20.4%
DEC2007	183	868	1051	17.4%	69	383	452	15.3%
JAN2008	189	860	1049	18.0%	52	399	451	11.5%
FEB2008	150	673	823	18.2%	61	282	343	17.8%
MAR2008	121	695	816	14.8%	52	268	320	16.3%
APR2008	88	0	88	100%	47	0	47	100%
MAY2008	66	0	66	100%	32	0	32	100%
JUN2008	41	0	41	100%	11	0	11	100%
JUL2008	4	0	4	100%	0	0	0	
TOTAL	**1539**	**5559**	**7098**	**21.7%**	**615**	**2329**	**2944**	**20.9%**

Scoring - Analysis

CATEGORIZATION OF CONTINUOUS PREDICTORS

Reasons for categorization

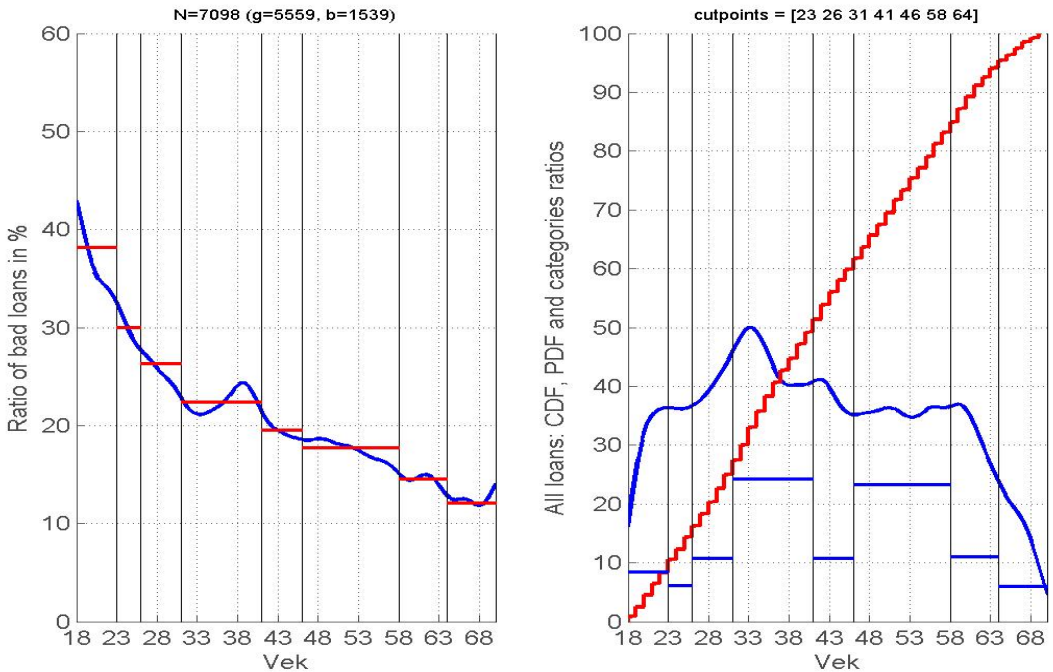
We prefer not to use continuous variables as explanatory variables in logistic regression models for scorecard development. For usage in logistic regression models, all continuous variables are categorized.

The goal of the categorization is to achieve categories which discriminates well (there are the considerable differences in badrate ratio between categories) and which are stable within the time.

Categorization algorithm

Each continuous variable is categorized separately.

CATEGORIZATION OF CONTINUOUS PREDICTORS



Categorization of the final demographic scorecard variable “age”. On the left pictures, the dependence of bad rate (smoothed using normal probability density function) on the variables is presented. On the right, the cumulative distribution function is presented. Vertical lines represent the borders between categories, horizontal red lines in the left picture represent the mean bad rate in categories, horizontal blue lines in the right picture represent the relative distribution of observations in the categories.

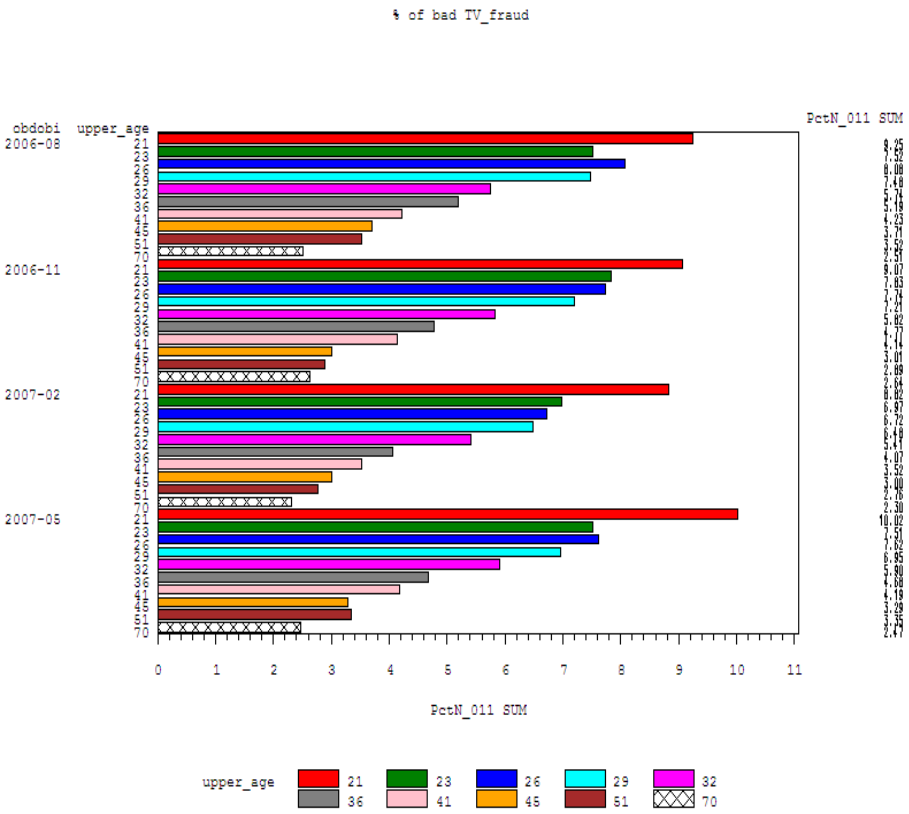
Analysis

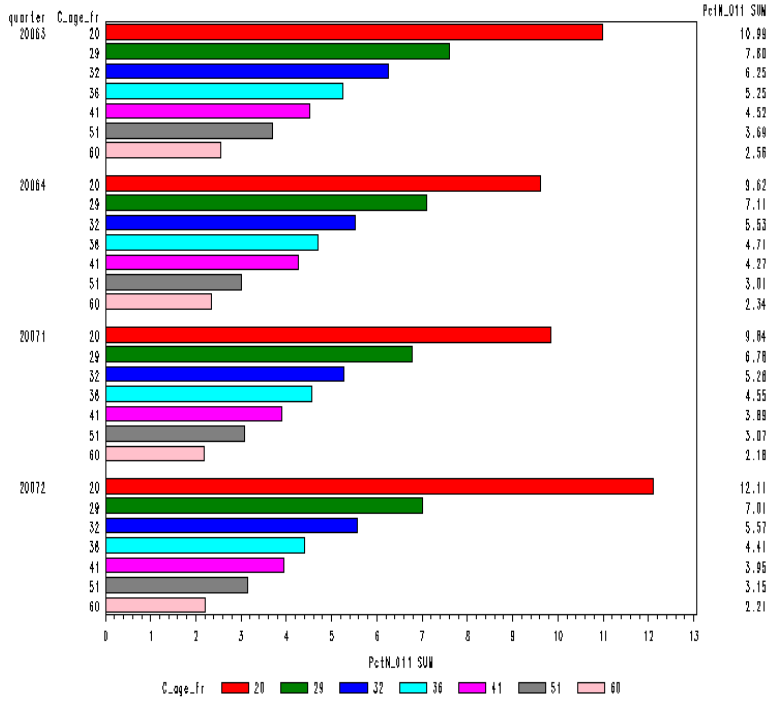
CATEGORIZATION OF CONTINUOUS PREDICTORS

We can see illogical inversion between categories 21-23 and 23-26.
In this case we rather group them in the same category.

	N	PctN	PctN	
			TV_fraud	
C_age_fr			0	1
20	35248	4.87	**89.32**	**10.68**
29	224503	31.03	**92.9**	**7.1**
32	62074	8.58	**94.36**	**5.64**
36	75261	10.4	**95.32**	**4.68**
41	82231	11.36	**95.87**	**4.13**
51	151677	20.96	**96.79**	**3.21**
60	92569	12.79	**97.7**	**2.3**
All	723563	100	**94.87**	**5.13**

age_fr pct risky all . segment





Analysis

UNIVARIATE ANALYSIS

- to think out, create and assess possible variables for the logistic regression model.

- each analysed variable is examined individually as a predictor of the target variable (good/bad loan).

The following statistics are considered:

- Weight of evidence
- Information Value
- Gini Coefficient

With help of the above mentioned statistics, it is possible to:

- Identify variables which are strong predictors for the target variable
- Create new or modify existing variables (mostly by re-categorization) to achieve even higher predicting power

Analysis

Weight of evidence, information value

r ... number of levels (categories) of the categorical 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: $woe_i = ln\,(g_i / G) - ln\,(b_i / B)$

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

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

Incorporation Date											
Raw	RegVar	Percant	B	G	TOT	G/B Odds	%Good	%Bad	Bad Rate	WoE	IV
0 & NOI	inc_1	12%	139	952	1091	7	11%	19%	12,7%	-0,557	0,046116
1	inc_2	13%	133	1073	1206	8	12%	19%	11,0%	-0,394	0,023731
2-7	miss	42%	299	3601	3900	12	42%	42%	7,7%	0,007	2,04E-05
8-15	inc_3	22%	108	1942	2050	18	23%	15%	5,3%	0,408	0,030887
16+	inc_4	11%	39	1019	1058	26	12%	5%	3,7%	0,781	0,050288
Total			718	8587	9305	12			**7,7%**		**0,151**

Scoring – model development

Model development

MODELLING APPROACH

The modelling approach used for scorecard development is **logistic regression**.

Reasons for selection:

-based on well-developed mathematical background

-world-wide market standard for scorecard development integrated in SAS software (statistical and data-mining software used in the HC Risk department)

Other approaches for scoring model development are possible, e.g. **decision trees, neural networks**, etc. These methods were not selected, because of lower transparency and worse interpretability than logistic regression.

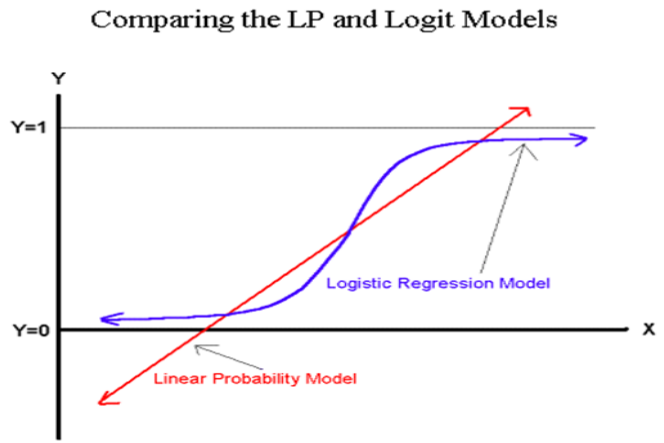
$$p(x) = 1 / [1 + exp(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_n x_n)]$$

The parameters $\beta_0, \beta_1, \ldots, \beta_n$ are the parameters of the model and represent score points. These parameters are estimated from the observed data using the so called maximum likelihood method.

Assumptions: dichotomous target variable; independence of observations (for the maximum likelihood estimates approach to be valid).

Model development

- We search coefficients for linear combination of predictors, such that bad guys have low sum of points and good guys high sum of points



$$sum\_points = intercept + \sum_{predictors} points\ from\ predictor$$

$$sum\_points = 3.45 + 0.12\,(is\ women) + 0.32\,(is\ only\ 24\ years\ old) + 0.05\,(married) \ldots$$

$$probability\_of\_default = \frac{1}{1 + \exp\,(-sum\_points)}$$

We are looking for these coefficients

HC: “score” = 1-probability_of_default (number in interval 0-1)

Model development

- **Forward**

- začíná se s prázdným modelem postupné přidávání proměnných

- **Backward**

- začíná se s plným modelem (všechny proměnné) ,postupné odebírání proměnných

- **Stepwise**

- začíná se s prázdným modelem, postupně se přidávají a odebírají proměnné

- **Enter**

- je předepsán seznam proměnných v modelu

Model development

SELECTION - consists of finding a set of variables, which will result in a “best” logistic regression model.

- The highest possible discriminating power (measured by Gini coefficient)

- Logical interpretability of all variables in model

- Stability of the Gini coefficient (the validation sample check)

Generally, the criteria could be summarized as the demand for simplicity and stability of the model.

Scoring – Stability and validation

Stability and validation

Discriminatory power

Gini coefficient, C-statistics

Gini coefficient and C-statistics are two equivalent measures of discrimination power for scoring models.

-*A* :set of loans on which we want to measure the performance of the model
-For each loan, we know whether it is a good loan (non-delinquent) or bad loan (delinquent)
- *A* consists of $N = k + l$ loans, k – number of good loans , l - number of bad loans
- *card(X)* : number of elements of a subset *X*
-*B* : subset of all possible pairs [good loan, bad loan]
-subset *B* consists of $k \cdot l$ such pairs ($card(B) = k \cdot l$)

Let' s define three subsets of the set *B*:

X_+ : all pairs [good loan, bad loan] from *B*, where *score(good) > score(bad)*
X_- : all pairs [good loan, bad loan] from *B*, where *score(good) < score(bad)*
X_0 : all pairs [good loan, bad loan] from *B*, where *score(good) = score(bad)*

It is clear that $card(B) = card(X_+) + card(X_-) + card(X_0)$.

Stability and validation

Discriminatory power

Gini coefficient is defined as follows:

$$gini := [card(X_+) - card(X_-)] / card(B)$$

C-statistics is defined as follows:

$$C := [card(X_+) + 0.5 \cdot card(X_0)] / card(B)$$

There exist the following relationships between gini coeficient and c-statistics:

$$gini = 2 \cdot C - 1$$
$$C = (gini + 1) / 2$$

Examples:
Perfect model: *gini=1, C=1*
for all pairs [good loan, bad loan] from *B score(good) > score(bad)*
Random model: *gini=0, C=0.5*
there exist significant number of pairs [good loan, bad loan] in *B* for which *score(good) < score(bad)* or *score(good) = score(bad)*
Reversed model: *gini=-1, C=0*
for all pairs [good loan, bad loan] from *B score(good) < score(bad)*. Discrimination power is as strong as for perfect model but model assigns high score to bads and low score to goods.

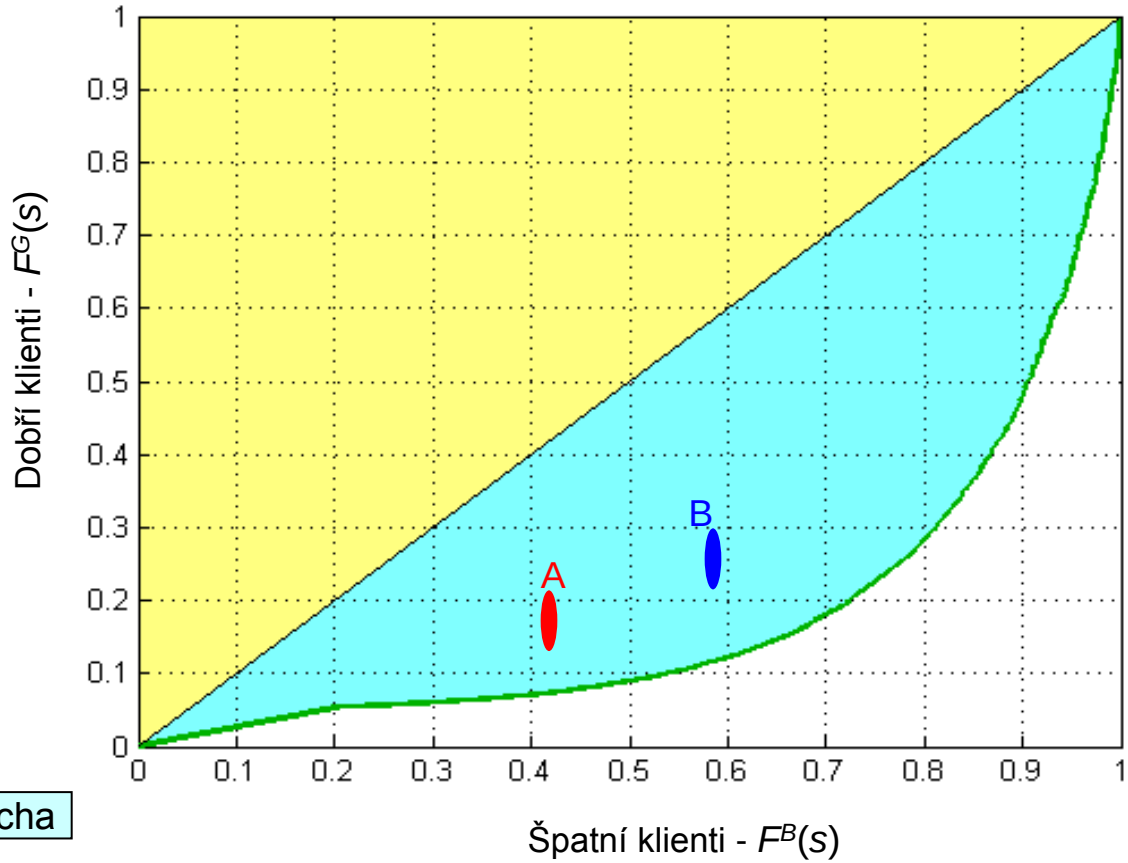
Stability and validation

Lorenzova křivka, Gini a c-statistika:

- A: se zamítnutím 10% dobrých zamítnu 55% špatných
- B se zamítnutím 20% dobrých zamítnu přes 70% špatných

$$L = \left\{ [F^B(s), F^G(s)] \in \mathbb{R}^2 : s \in S \right\}$$

$F^B(s)$ – distribuční funkce špatných klientů
$F^G(s)$ - distribuční funkce dobrých klientů



- Giniho koeficient = 2* modrá plocha
- c-statistika = modrá plocha + žlutý trojúhelník

Stability and validation

Discriminatory power

Lift *n%*

Lift *n%* coefficient is an alternate measure of discrimination power for scoring models. It describes the performance of the model with a cut-off in the *n%* quantile of the testing sample.

-Let's have a set of loans *A*; like in the previous section.
-For each loan, we know whether it is a good loan or a bad loan. Let's denote
-*card(X)* the number of elements of a set *X*
-b_X number of bad loans in the set *X*

For each loan, we calculate the score using the model we want to evaluate. Then, we sort the set *A* according to the score and define a set *B* of a *n%* quantile of *A*.

Example: For computing lift 10%, the set B is 10 % of loans from A with the lowest score.

$card(B) = floor[n\% \cdot card(A)]$

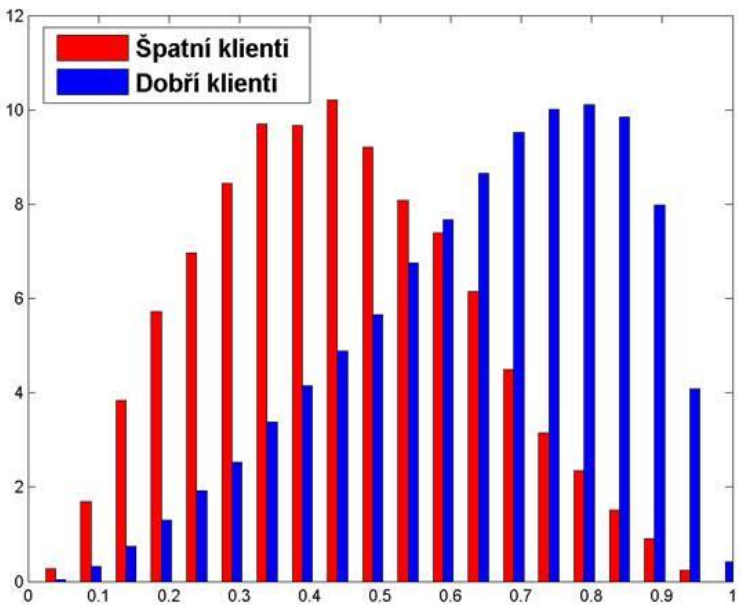
The *lift n%* coefficient is then defined as follows:

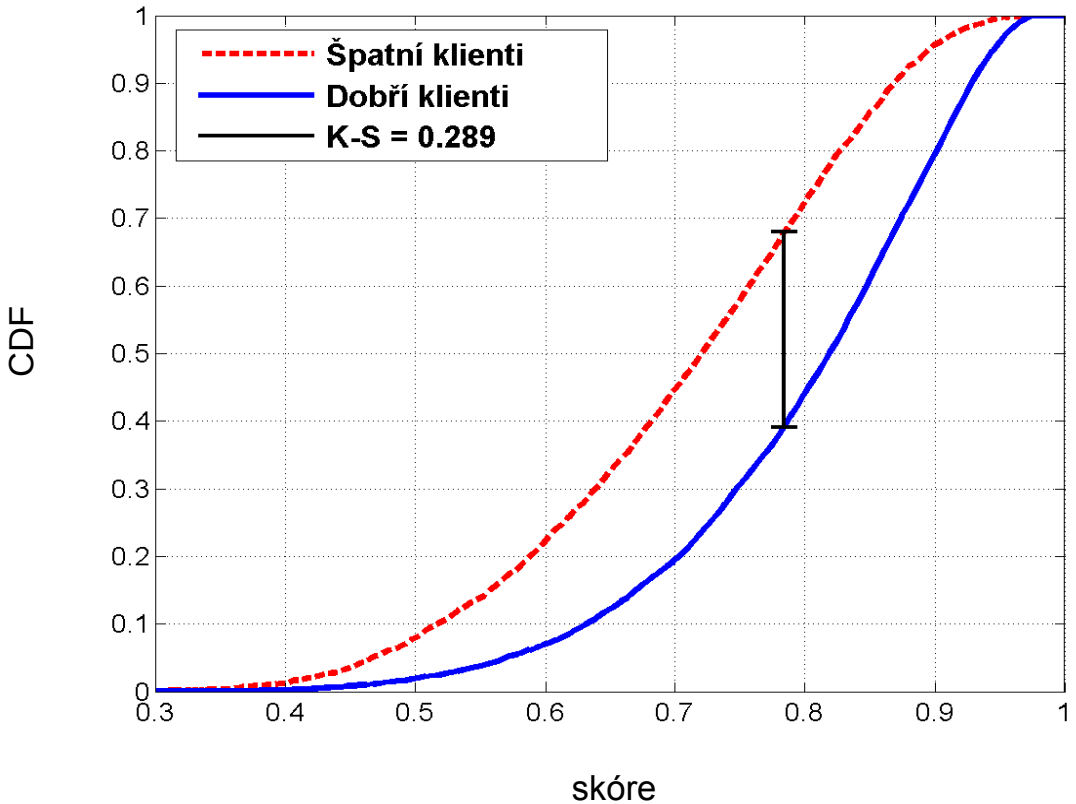
$$\textbf{Lift n\% := } [b_B / card(B)] / [b_A / card(A)].$$

Stability and validation

Distribuční funkce a K-S statistika:

- při skóre <= 0.78 je v populaci **40% dobrých** a **69% špatných**
- K-S je tedy rovno **29%**





Stability and validation

VALIDATION SAMPLE TEST

The performance of the models was checked on the validation sample and the target variable used during the model development .

Gini coefficients was compared on development and validation samples using the new and the current score.

The comparison shows that the performance of the model is exactly the same on the development and validation sample with substantial improvement from the old scorecard.

Gini	Development sample	Validation sample
New score	0.342	0.342
Old score	0.265	0.308

Comparison of the Gini coefficient on development and validation samples.

Software used for development

- SAS 9.1.3 Servise pack 4 for Windows
- MATLAB 7.1.0.246 (R14) Service pack 3
- Microsoft SQL Server Management Studio Express 9.00.2047.00
- Microsoft Office 2007

Some results for normally distributed scores

Some results for normally distributed scores

- Assume that the scores of good and bad clients are normally distributed, i.e. we can write their densities as

$$f_{GOOD}(x) = \frac{1}{\sigma_g \sqrt{2\pi}} e^{-\frac{(x-\mu_g)^2}{2\sigma_g^2}} \qquad f_{BAD}(x) = \frac{1}{\sigma_b \sqrt{2\pi}} e^{-\frac{(x-\mu_b)^2}{2\sigma_b^2}}$$

- Estimates of parameters μ_b, μ_b, σ_g and σ_b :

  M_g, M_b are means of good (bad) clients

  S_g , S_b are standard deviations of good (bad) clients

- Pooled standard deviation:

$$S = \left(\frac{nS_g{}^2 + mS_b{}^2}{n+m} \right)^{\frac{1}{2}}$$

- Estimates of mean and standard dev. of scores for all clients μ_{ALL}, σ_{ALL} :

$$M = M_{ALL} = \frac{nM_g + mM_b}{n+m} \qquad S_{ALL} = \left(\frac{nS_g{}^2 + mS_b{}^2 + n\left(M_g - M\right)^2 + m\left(M_b - M\right)^2}{\left(n+m\right)} \right)^{\frac{1}{2}}$$

Number of good clients: n
Number of bad clients: m
Proportions of good/bad clients: $p_G = \frac{n}{n+m}$, $p_B = \frac{m}{n+m}$

Some results for normally distributed scores

- **Mean difference (Mahalanobis distance):**

$$D = \frac{\mu_g - \mu_b}{\sigma}$$

- **Kolmogorov-Smirnov statistics:**

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

- **Gini coefficient:**

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

- **Lift:**

$$Lift_q = \frac{1}{q} F_{BAD}\left(F_{ALL}^{-1}(q)\right)$$

- **Information value (I_{val}) – continuous case (Divergence):**

$$I_{val} = \int_{-\infty}^{\infty} \left(f_{GOOD}(s) - f_{BAD}(s)\right) \ln\left(\frac{f_{GOOD}(s)}{f_{BAD}(s)}\right) ds$$

F_{BAD}, F_{GOOD} and F_{ALL} are cumulative distribution functions of scores for bad, good and all clients.

Some results for normally distributed scores

- Assume that standard deviations are equal to a common value σ :

$$D = \frac{\mu_g - \mu_b}{\sigma}$$

$$D = \frac{M_g - M_b}{S}$$

$$KS = \Phi\left(\frac{D}{2}\right) - \Phi\left(\frac{-D}{2}\right) = 2 \cdot \Phi\left(\frac{D}{2}\right) - 1$$

$$Gini = 2 \cdot \Phi\left(\frac{D}{\sqrt{2}}\right) - 1$$

$$Lift_q = \frac{1}{q}\Phi\left(\frac{\sigma_{ALL}}{\sigma} \cdot \Phi^{-1}(q) + p_G \cdot D\right)$$

$$Lift_q = \frac{1}{q}\Phi\left(\frac{S_{ALL}}{S}\Phi^{-1}(q) + p_G \cdot D\right)$$

$$I_{val} = D^2$$

Where $\Phi(\cdot)$ is the standardized normal distribution function, $\Phi_{\mu,\sigma^2}(\cdot)$ the normal distribution function with parameters μ, σ^2 and $\Phi^{-1}(\cdot)$ is the standard quantile function.

Some results for normally distributed scores

- Generally (i.e. without assumption of equality of standard deviations):

$$D^* = \frac{\mu_g - \mu_b}{\sqrt{\sigma_g^2 + \sigma_b^2}}$$

$$D^* = \frac{M_g - M_b}{\sqrt{S_g^2 + S_b^2}}$$

$$KS = \Phi\left(\frac{a}{b}\sigma_b \cdot D^* - \frac{1}{b}\sigma_g\sqrt{a^2 D^{*2} + 2b \cdot c}\right) - \Phi\left(\frac{a}{b}\sigma_g \cdot D^* - \frac{1}{b}\sigma_b\sqrt{a^2 D^{*2} + 2b \cdot c}\right)$$

where $a = \sqrt{\sigma_b^2 + \sigma_g^2}$, $b = \sigma_b^2 - \sigma_g^2$, $c = \ln\left(\frac{\sigma_g}{\sigma_b}\right)$

$$KS = \Phi\left(\frac{\sqrt{S_b^2 + S_g^2}}{S_b^2 - S_g^2} S_b \cdot D^* - \frac{1}{S_b^2 - S_g^2} S_g \sqrt{\left(S_b^2 + S_g^2\right) D^{*2} + 2 \cdot \left(S_b^2 - S_g^2\right) \ln\left(\frac{S_g}{S_b}\right)}\right)$$

$$- \Phi\left(\frac{\sqrt{S_b^2 + S_g^2}}{S_b^2 - S_g^2} S_g \cdot D^* - \frac{1}{S_b^2 - S_g^2} S_b \sqrt{\left(S_b^2 + S_g^2\right) D^{*2} + 2 \cdot \left(S_b^2 - S_g^2\right) \ln\left(\frac{S_g}{S_b}\right)}\right)$$

Some results for normally distributed scores

- Generally (i.e. without assumption of equality of standard deviations):

$$Gini = 2 \cdot \Phi\left(D^*\right) - 1$$

$$Lift_q = \frac{1}{q}\Phi_{\mu_b,\sigma_b^2}\left(\mu_{ALL} + \sigma_{ALL} \cdot \Phi^{-1}(q)\right) = \frac{1}{q}\Phi\left(\frac{\sigma_{ALL} \cdot \Phi^{-1}(q) + \mu_{ALL} - \mu_b}{\sigma_b}\right)$$

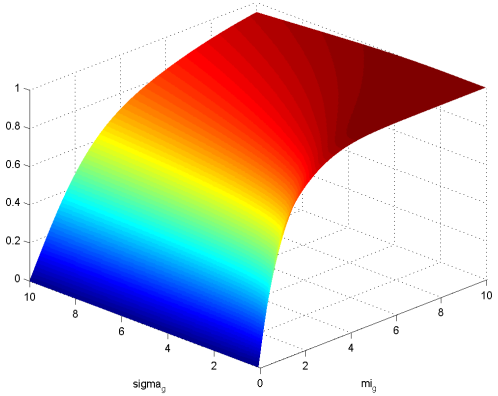
$$Lift_q = \frac{1}{q}\Phi\left(\frac{S_{ALL} \cdot \Phi^{-1}(q) + M - M_b}{S_b}\right)$$

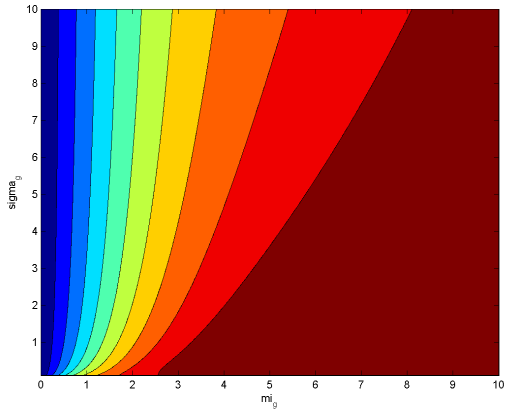
$$I_{val} = (A+1)D^{*2} + A - 1, \quad A = \frac{1}{2}\left(\frac{\sigma_b^2}{\sigma_g^2} + \frac{\sigma_g^2}{\sigma_b^2}\right)$$

$$I_{val} = (A+1)D^{*2} + A - 1, \quad A = \frac{1}{2}\left(\frac{S_b^2}{S_g^2} + \frac{S_g^2}{S_b^2}\right)$$

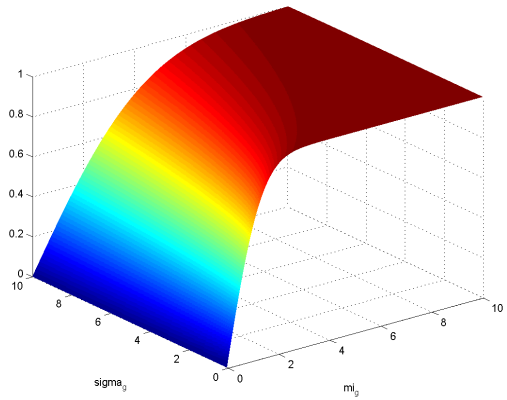
Some results for normally distributed scores

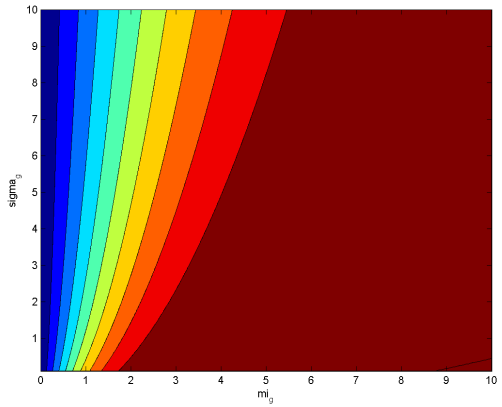
- **KS:** $\mu_b = 0, \sigma_b^2 = 1$





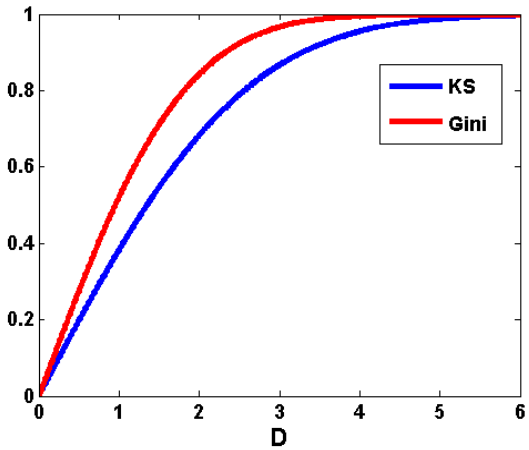
- **Gini** $\mu_b = 0, \sigma_b^2 = 1$





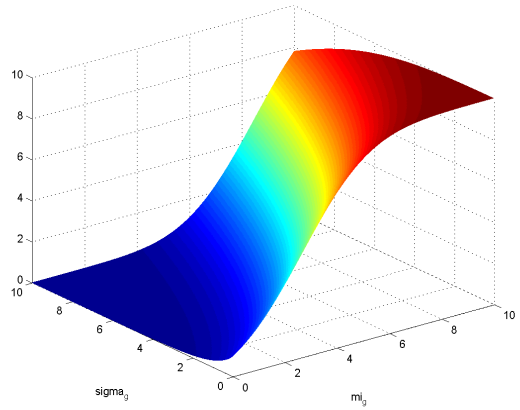
- KS and the Gini react much more to change of μ_g and are almost unchanged in the direction of σ_g^2 .

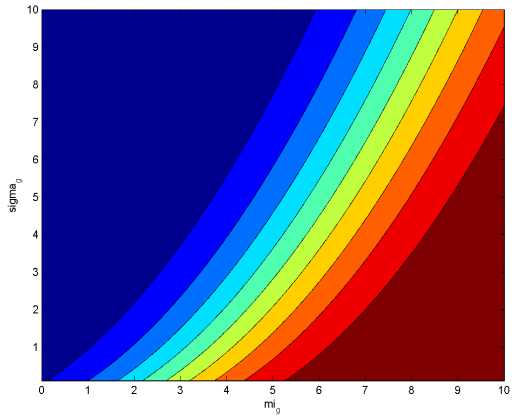
- **Gini > KS**



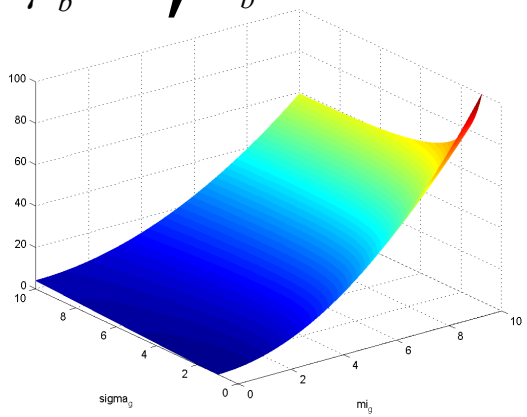
Some results for normally distributed scores

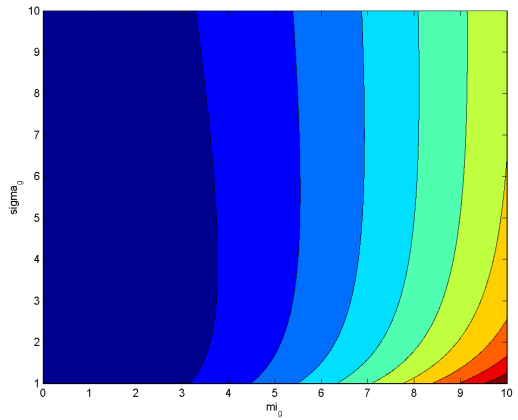
- **$Lift_{10\%}$:** $\mu_b = 0$, $\sigma_b^2 = 1$





- **I_{val}:** $\mu_b = 0$, $\sigma_b^2 = 1$





- In case of $Lift_{10\%}$ it is evident strong dependence on μ_g and significantly higher dependence on σ_g^2 than in case of KS and Gini.

- Again strong dependence on μ_g. Furthermore value of I_{val} rises very quickly to infinity when σ_g^2 tends to zero.

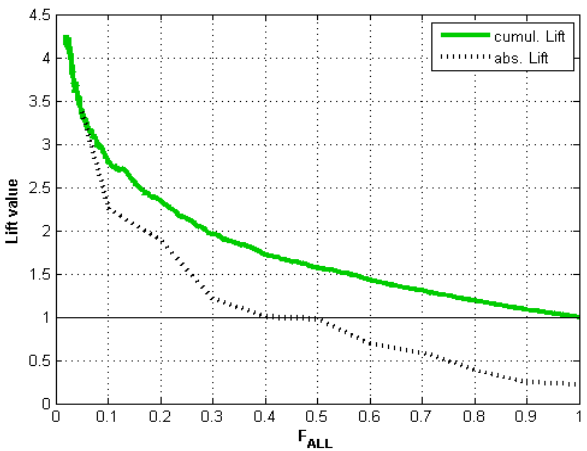
Some results for Lift

Lift

- *cumulative Lift* says how many times, at a given level of rejection, is the scoring model better than random selection (random model). More precisely, the ratio indicates the proportion of bad clients with less than a score *a*, $a \in [L, H]$, to the proportion of bad clients in the whole population. Formally, it can be expressed by:

$$Lift(a) = \frac{CumBadRate(a)}{BadRate} = \frac{\frac{\sum_{i=1}^{n+m} I(s_i \le a \wedge Y = 0)}{\sum_{i=1}^{n+m} I(s_i \le a)}}{\frac{\sum_{i=1}^{n+m} I(Y = 0)}{\sum_{i=1}^{n+m} I(Y = 0 \vee Y = 1)}} = \frac{\frac{\sum_{i=1}^{n+m} I(s_i \le a \wedge Y = 0)}{\sum_{i=1}^{n+m} I(s_i \le a)}}{\frac{n}{N}}$$

$$absLift(a) = \frac{BadRate(a)}{BadRate}$$



Lift

- Lift can be expressed and computed by formulae:

$$Lift(a) = \frac{F_{n.BAD}(a)}{F_{N.ALL}(a)} \quad a \in [L, H]$$

$$QLift(q) = \frac{F_{n.BAD}(F^{-1}_{N.ALL}(q))}{F_{N.ALL}(F^{-1}_{N.ALL}(q))} = \frac{1}{q} F_{n.BAD}\left(F^{-1}_{N.ALL}(q)\right)$$

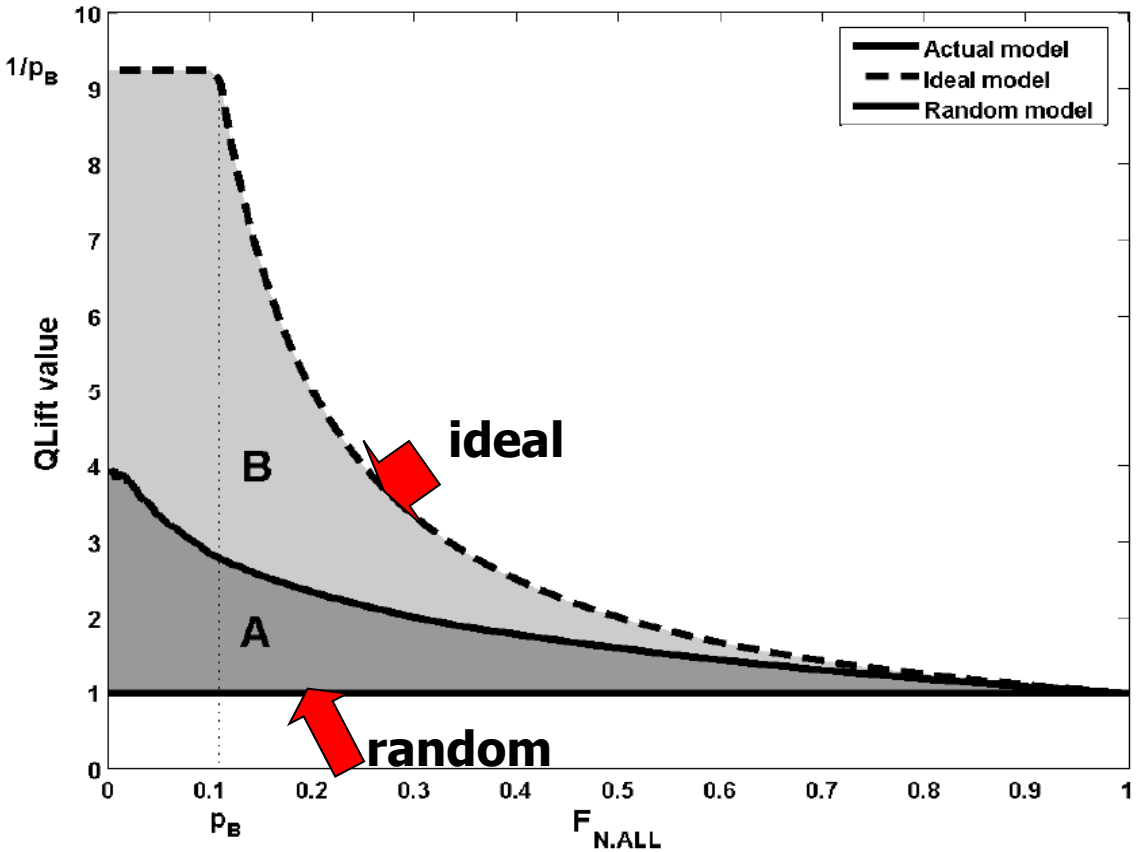
$$F^{-1}_{N.ALL}(q) = \min\{a \in [L, H], F_{N.ALL}(a) \geq q\}$$

$$QLift(0.1) = 10 \cdot F_{n.BAD}\left(F^{-1}_{N.ALL}(0.1)\right)$$

Lift

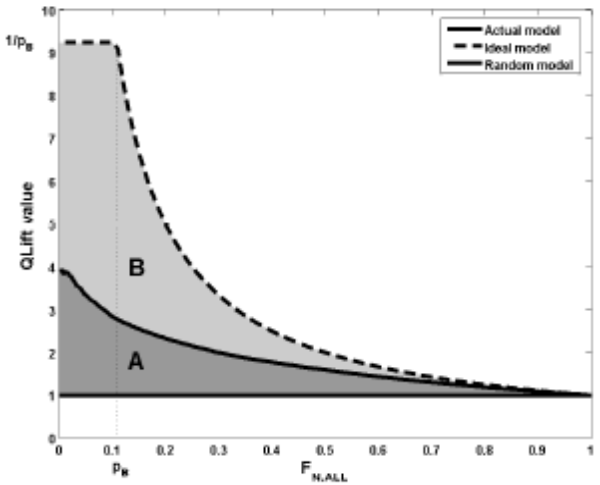
- Lift for ideal model:

$$Lift_{ideal}(a) = \begin{cases} \frac{1}{p_B}, & a \leq c \\ \frac{1}{F_{N.ALL}(a)}, & a > c. \end{cases}$$



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

Lift



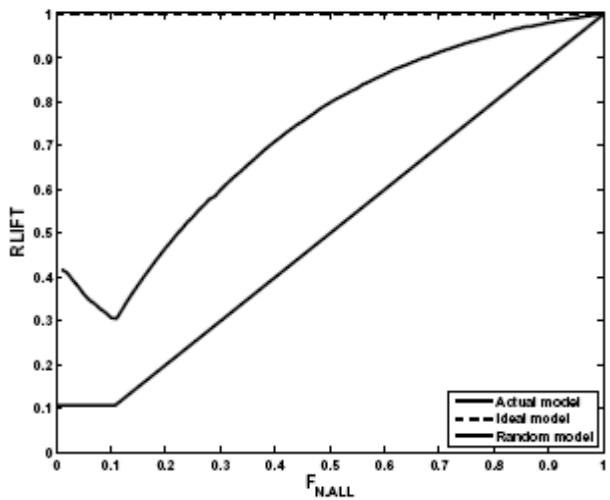
- Lift ratio as analogy to Gini coefficient:

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

Podstatnou výhodou tohoto indexu je fakt, že umožňuje korektní porovnání modelů vyvinutých na různých datech, což není možné pomocí hodnot funkce QLift.

- Zatímco LR porovnává plochy pod funkcí Liftu pro daný model a model ideální, následující myšlenka je založena na porovnání přímo těchto funkcí samotných. Definujme relativní Lift funkci pomocí

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

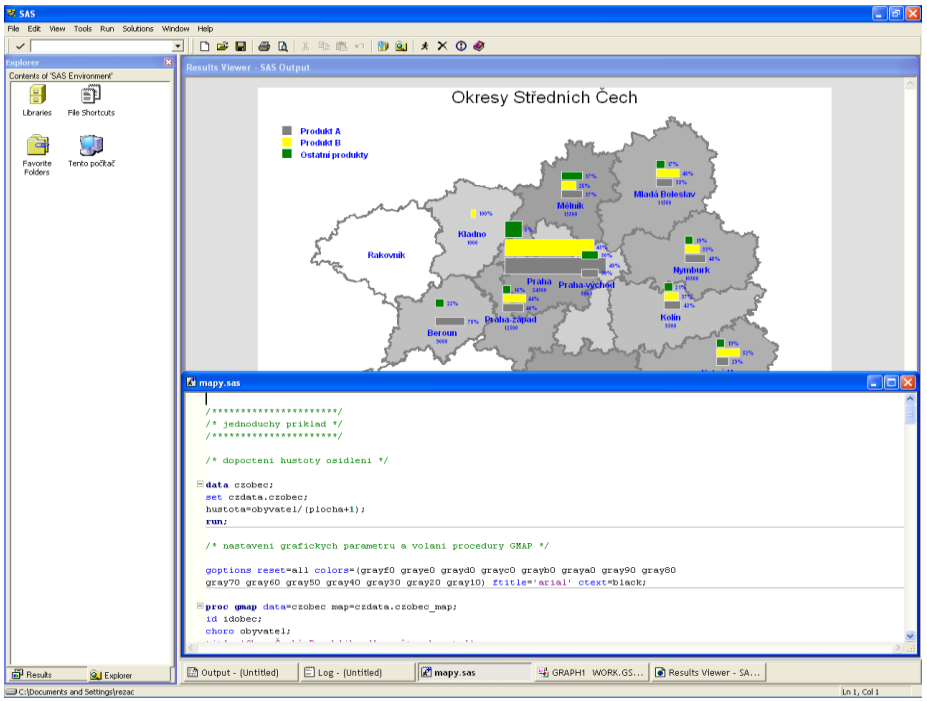


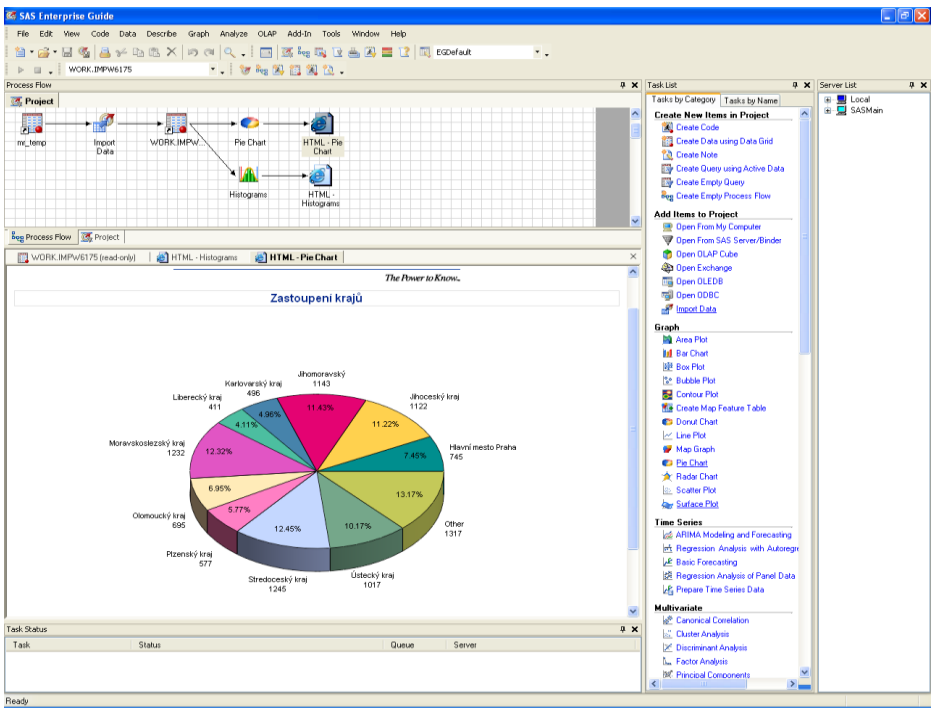

HOME
CREDIT


FACULTAS RERUM NATURALIUM
SCIENTIA EST POTENTIA
UNIVERSITAS MASARYKIANA BRUNENSIS

SAS

: www.sas.com


SAS
File Edit View Tools Run Solutions Window Help
Explorer
Contents of 'SAS Environment'
Libraries
File Shortcuts
Favorite Folders
Tento počítač
Results Viewer - SAS Output
Okresy Středních Čech
Produkt A
Produkt B
Ostatní produkty
Mělník
Mladá Boleslav
Kladno
Rakovník
Nymburk
Praha
Praha-východ
Praha-Západ
Kolín
Beroun
mapy.sas
/**********************/
/* jednoduchy priklad */
/**********************/
/* dopocteni hustoty osidleni */
data czobec;
set czdata.czobec;
hustota=obyvatel/(plocha+1);
run;
/* nastaveni grafickych parametru a volani procedury GMAP */
goptions reset=all colors=(grayf0 graye0 grayd0 grayc0 grayb0 graya0 gray90 gray80
gray70 gray60 gray50 gray40 gray30 gray20 gray10) ftitle='arial' ctext=black;
proc gmap data=czobec map=czdata.czobec_map;
id idobec;
choro obyvatel;
Output - (Untitled)
Log - (Untitled)
mapy.sas
GRAPH1 WORK.GS...
Results Viewer - SA...
C:\Documents and Settings\rezac
Ln 1, Col 1


SAS Enterprise Guide
File Edit View Code Data Describe Graph Analyze OLAP Add-In Tools Window Help
Process Flow
Project
mr_temp
Import Data
WORK.IMPW...
Pie Chart
HTML - Pie Chart
Histograms
HTML - Histograms
Task List
Tasks by Category
Tasks by Name
Create New Items in Project
Create Code
Create Data using Data Grid
Create Note
Create Query using Active Data
Create Empty Query
Create Empty Process Flow
Add Items to Project
Open From My Computer
Open From SAS Server/Binder
Open OLAP Cube
Open Exchange
Open OLEDB
Open ODBC
Import Data
Graph
Area Plot
Bar Chart
Box Plot
Bubble Plot
Contour Plot
Create Map Feature Table
Donut Chart
Line Plot
Map Graph
Pie Chart
Radar Chart
Scatter Plot
Surface Plot
Time Series
ARIMA Modeling and Forecasting
Regression Analysis with Autoregr
Basic Forecasting
Regression Analysis of Panel Data
Prepare Time Series Data
Multivariate
Canonical Correlation
Cluster Analysis
Discriminant Analysis
Factor Analysis
Principal Components
Server List
Local
SASMain
WORK.IMPW6175 (read-only)
HTML - Histograms
HTML - Pie Chart
The Power to Know.
Zastoupení krajů
Jihomoravský 1143
Karlovarský kraj 496
Liberecký kraj 411
Jihocesky kraj 1122
Moravskoslezský kraj 1232
Hlavní mesto Praha 745
Olomoucký kraj 695
Other 1317
Plzenský kraj 577
Ústecký kraj 1017
Stredocesky kraj 1245
11.43%
11.22%
4.96%
4.11%
12.32%
7.45%
13.17%
6.95%
5.77%
12.45%
10.17%
Task Status
Task
Status
Queue
Server
Ready

SAS

Společnost SAS Institute:

- Vznik 1976 v univerzitním prostředí
- Dnes:největší soukromá softwarová společnost na světě (více než 11.000 zaměstnanců)
- přes 45.000 instalací
- cca 9 milionů uživatelů ve 118 zemích
- v USA okolo 1.000 akademických zákazníků (SAS používá většina vyšších a vysokých škol a výzkumných pracovišť)

Studentské soutěže o hodnotné ceny

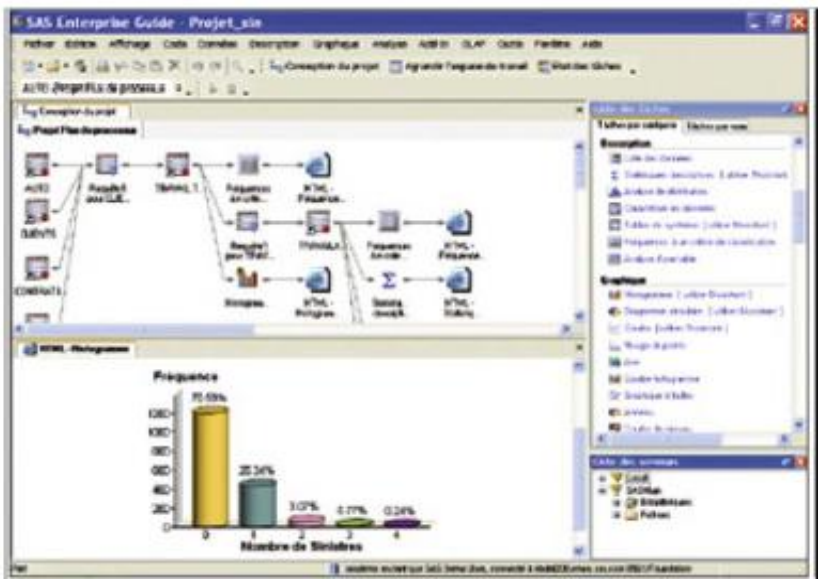
- Lokální (ČR) soutěž o nejlepší studentskou práci
- SAS Student Ambassador - celosvětová soutěž o nejlepší práce s využitím SAS
- Možnost účasti a prezentace na SAS konferenci v Seattlu

Podpora studentů

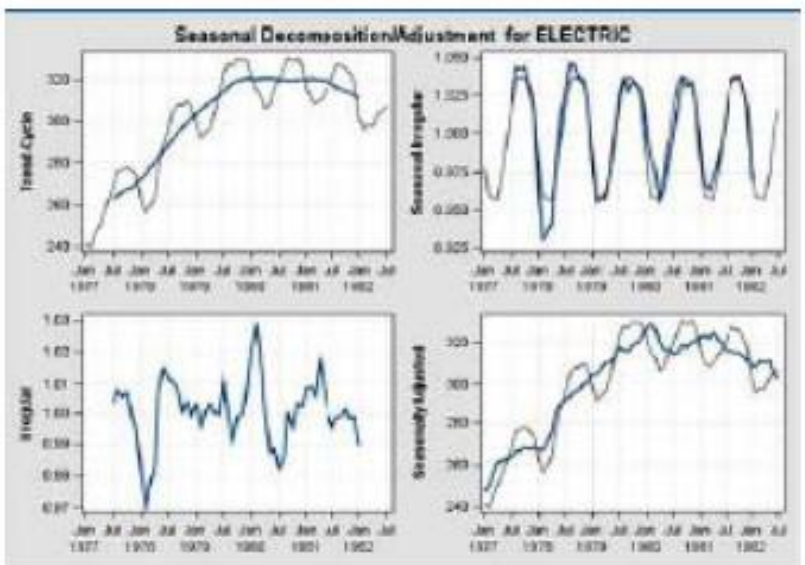
- Možnost rozšíření licence na domácí instalace pro studenty
- SAS Fellowship Program – software zdarma pro diplomku či dizertaci
- Zadávání a vedení diplomových prací
- Sdílení informací, zkušeností či příkladů v uživatelských skupinách

- Interaktivní moduly nebo programovací prostředí
  - Statistická analýza
  - Matice
  - Časové řady
  - Operační výzkum
  - Kontrola kvality

- Statistická analýza:
  - Popisná statistika
  - Analýza kontingenčních (frekvenčních) tabulek
  - Regresní, korelační, kovarianční analýza
  - Logistická regrese
  - Analýza rozptylu
  - Testování hypotéz
  - Diskriminační analýza
  - Shluková analýza
  - Analýza přežití
  - …



- Analýza časových řad:
  - Regresní modely
  - Modely se sezónními faktory
  - Autoregresní modely
  - ARIMA
  - Metody exponenciálního vyrovnání
  - …



SAS

- Více o SASu: http://www.sas.com/offices/europe/czech/

- (neúplný) seznam komerčních společností využívající SAS: http://www.sas.com/offices/europe/czech/reference/list.html

- o akademickém programu: http://www.sas.com/offices/europe/czech/academic/index.html

- o konferenci SAS forum: http://www.sas.com/reg/offer/cz/2010_sas_forum_2010

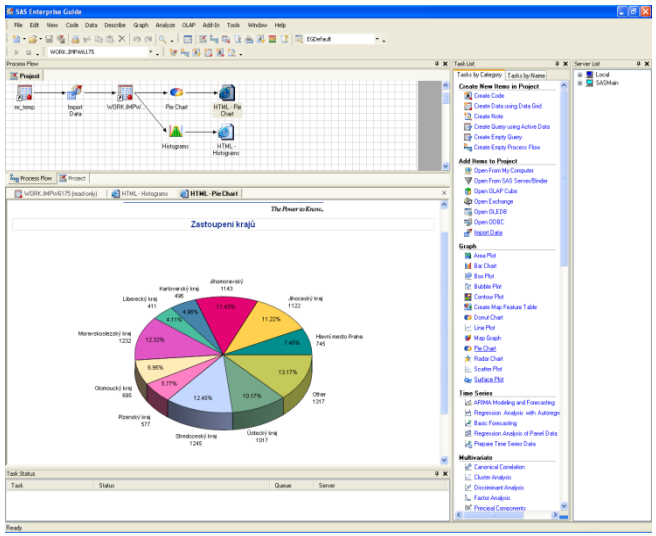
SAS v HC

SAS Base

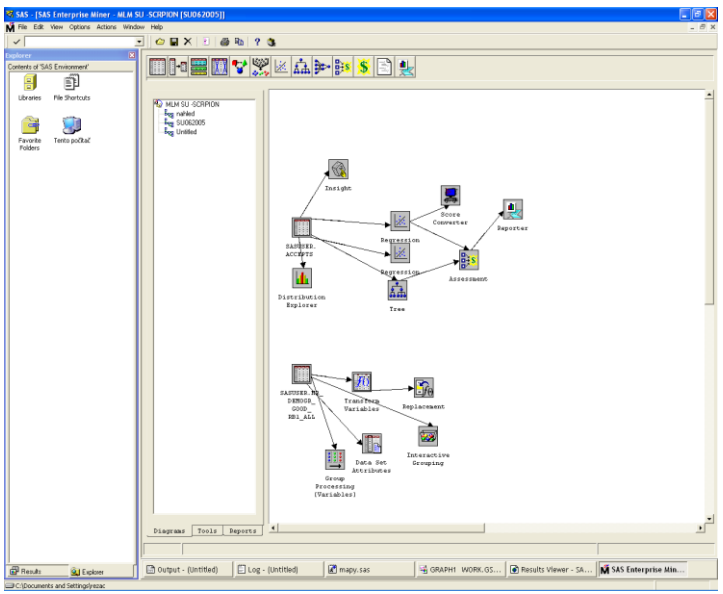
SAS/STAT

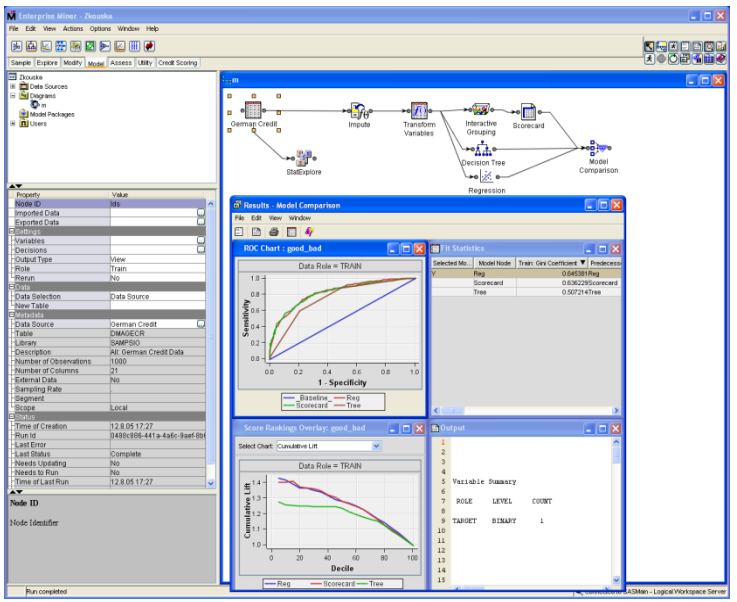
SAS/GRAPH

SAS/ETS



SAS Enterprise Miner:





SAS v HC

SAS používáme na: (Risk + CRM)

-import, přelití a transformaci dat

-tvorbu grafických výstupů

-prediktivní modelování (scoring)

-segmentaci dat (clustering – shlukování)

SAS

- SAS používají např.: