Advanced statistical methods for credit risk modeling in practice

Kádár Ferenc
OTP Bank
Analysis and Modeling Department
2016.09.27

OTP Group is the biggest independent banking group in Central Eastern Europe

OTP Group is offering universal banking services to more than 13 million customers in 9 countries via 1400 branches and more than 4000 ATMs.

In 2015 the OTP Group achieved 63 billion HUF (~200 million EUR) corrected consolidated profit after tax. The profitability, liquidity and the capital adequacy of the Group is still outstanding in international comparison.

OTP Group highlights

- OTP is a dominant banking player in Hungary, founded in 1949 (privatization in 1995 – introduced to Budapest Stock Exchange).
- Currently the bank is characterized by dispersed ownership of mostly private and institutional (financial) investors.
- OTP Bank has completed several successful acquisitions in the past years, becoming a key player in the region. Besides Hungary, OTP Bank currently operates in 8 countries of the region.
- Around 43.000 employees in the region, more than 10.000 billion HUF (around 33 billion EUR, 1/3 of Hungarian GDP) total assets.
- Despite the intense competition OTP Bank market position is stable in several segments, as well as in terms of profitability and stability belongs to the European frontline.

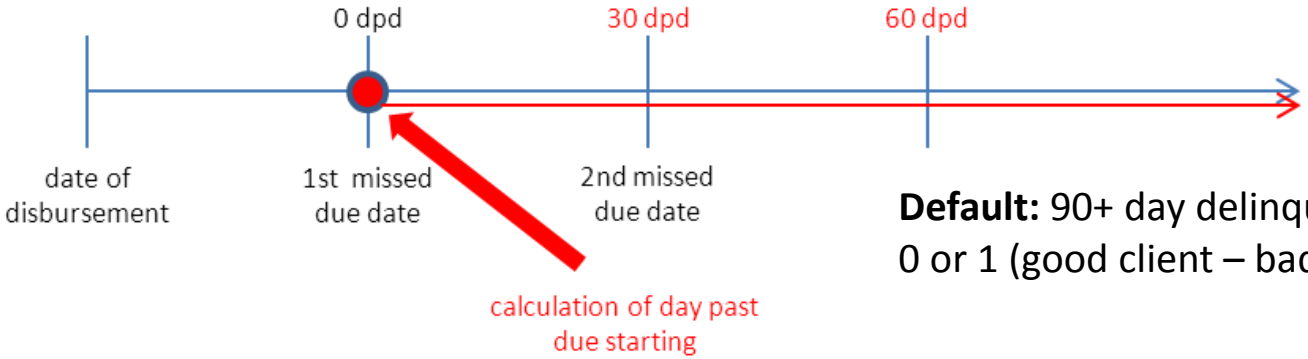
Modeling what? And why?

- Type of risks in a bank: Market risk, Operational risk, **Credit risk**
- How can we measure credit risk? Expected loss of lending:

Risk cost = PD·LGD·EAD

Risk parameters:

- **PD** – Probability of Default [0, 1]
- **LGD** – Loss Given Default [0, 1]
- **EAD** – Exposure at Default [0, ∞) - generally limited ☺



Default: 90+ day delinquency in 1 year.
0 or 1 (good client – bad client)

„Risk is not measurable (outside of casinos or the minds of people who call themselves 'risk experts')"

PD modeling (scorecards)

Modelling database

- Application scorecard for walk-in clients

socio-demographic and financial status, info on employment

- Application scorecards for OTP clients

socio-demographic and financial status, info on employment, historical client data

application data + behavior data of the credit

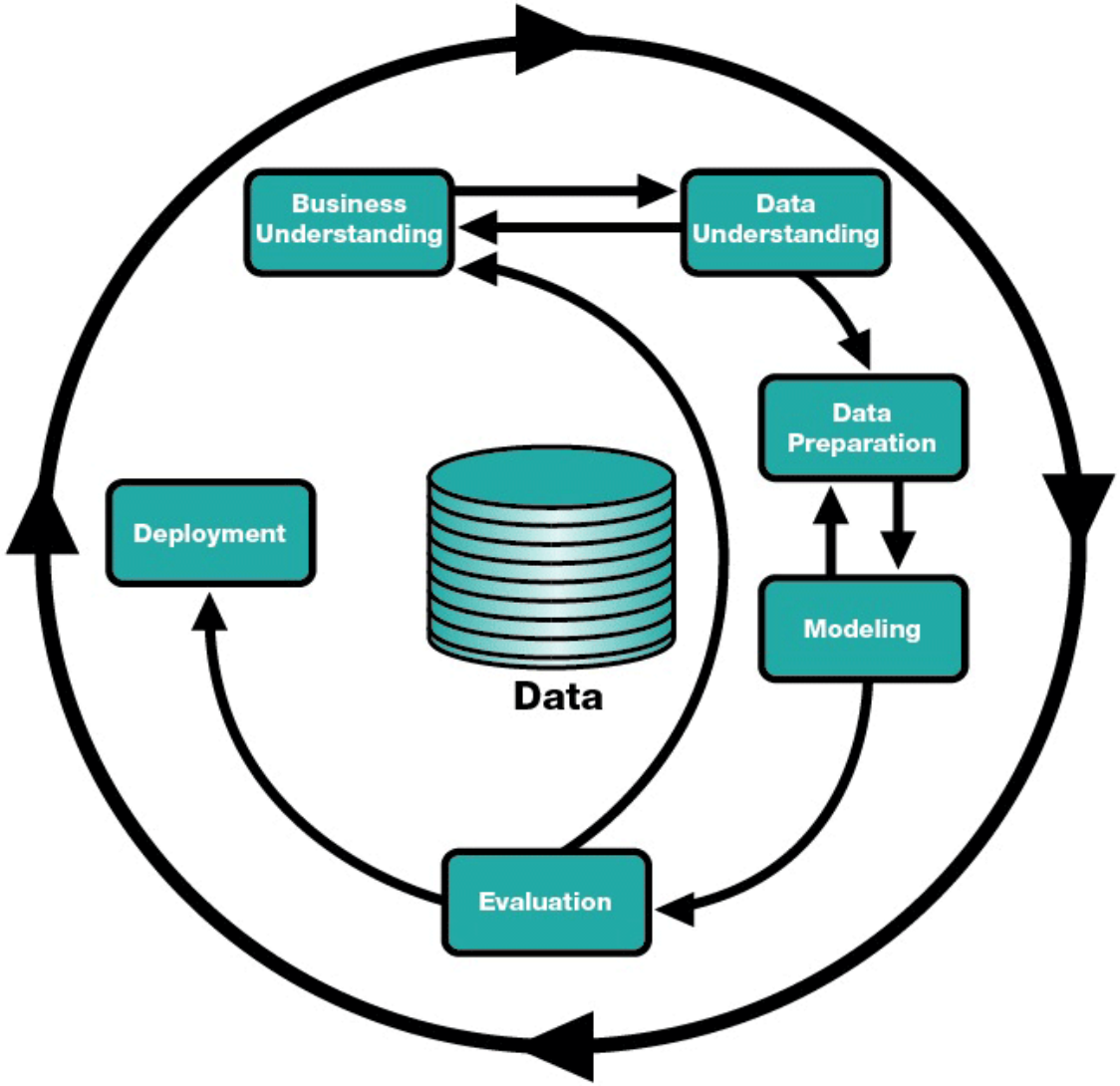
- Behaviour scorecards

application data + behavior data of the client, network data

- Early default scorecards (Fraud)

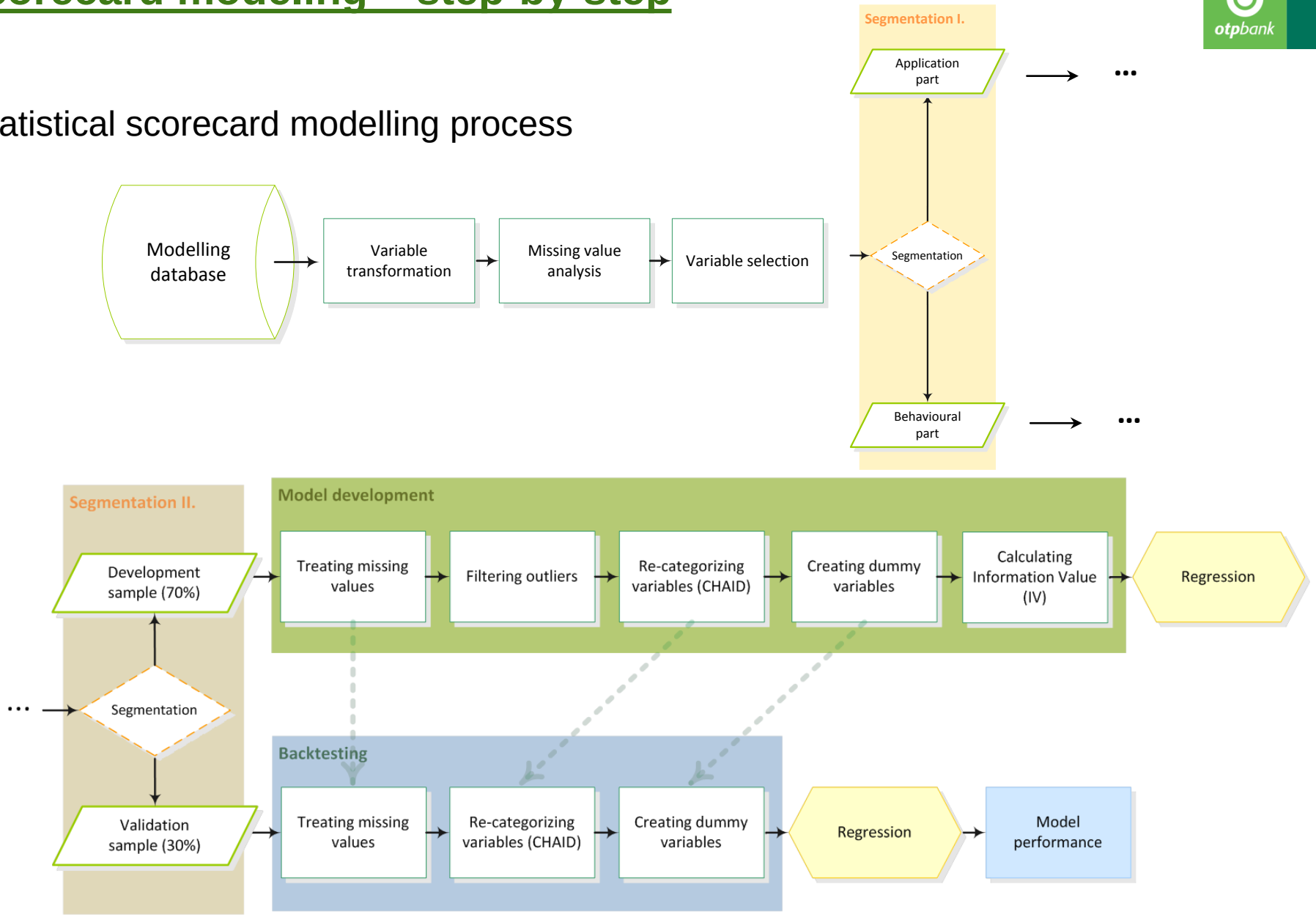
CRISP-DM methodology

CRoss
Industry
Standard
Process for
Data
Mining

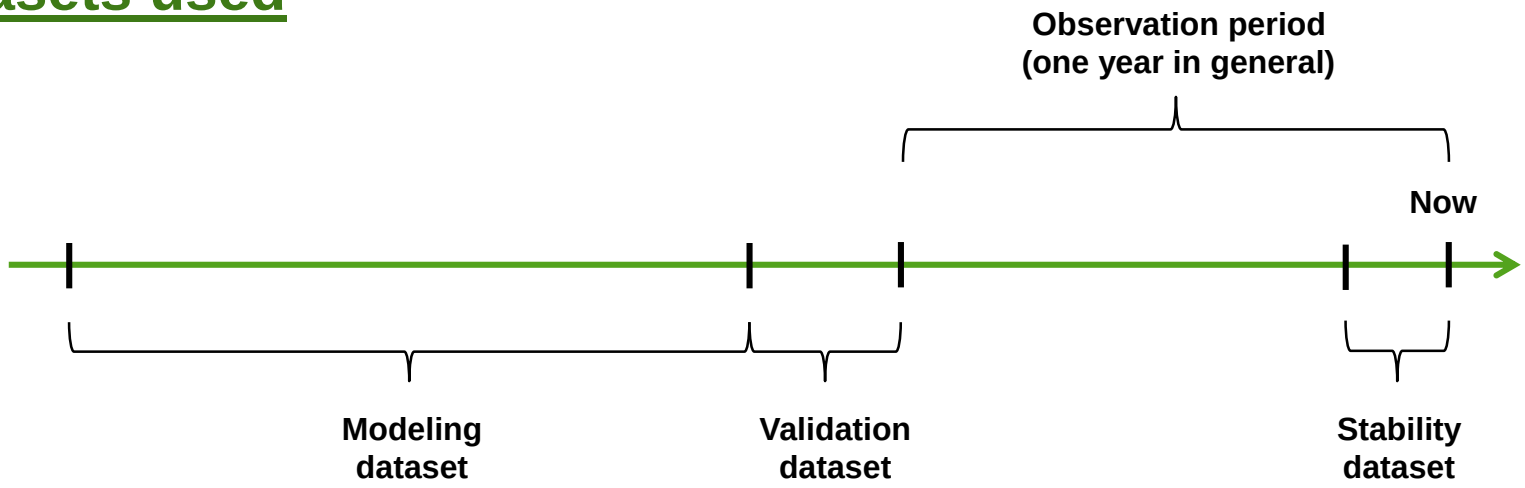


Scorecard modeling – step-by-step

Statistical scorecard modelling process



Datasets used



MODELING dataset is divided as follows:

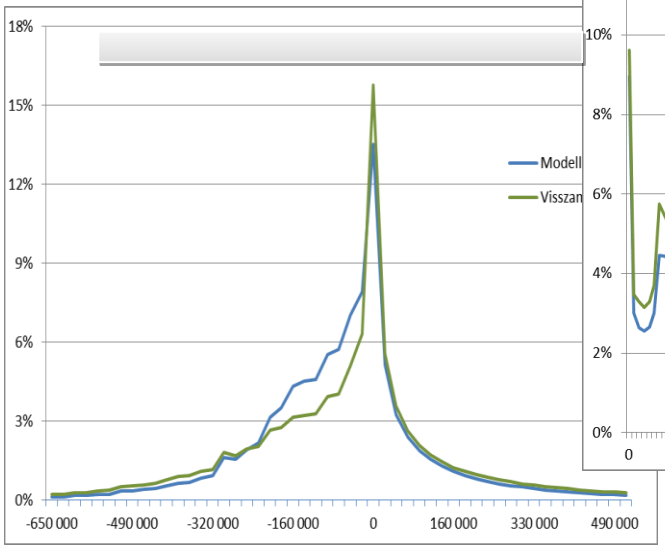
- Training dataset
- Testing dataset – out-of-sample validation
- Filtered dataset

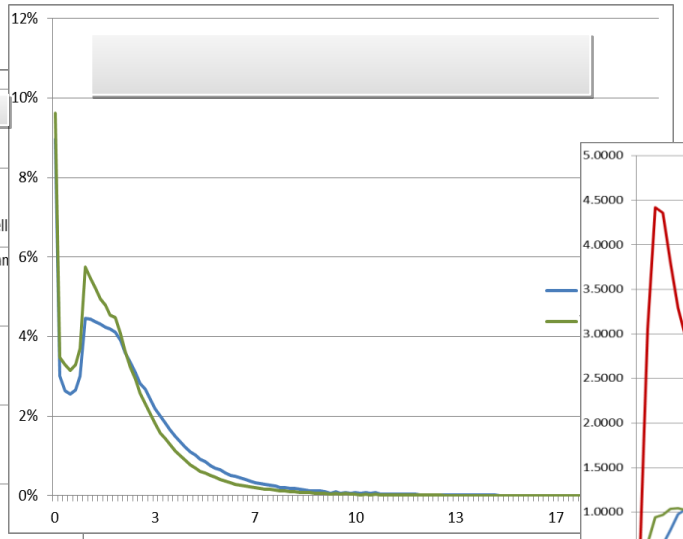
VALIDATION dataset serves as out-of-time validation (equivalent with currently used „More recent database”)

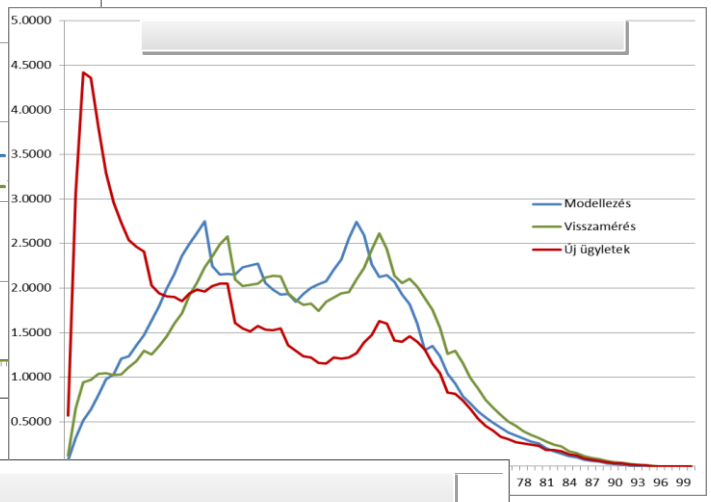
STABILITY dataset is used for checking stability (equivalent with „Most recent database”)

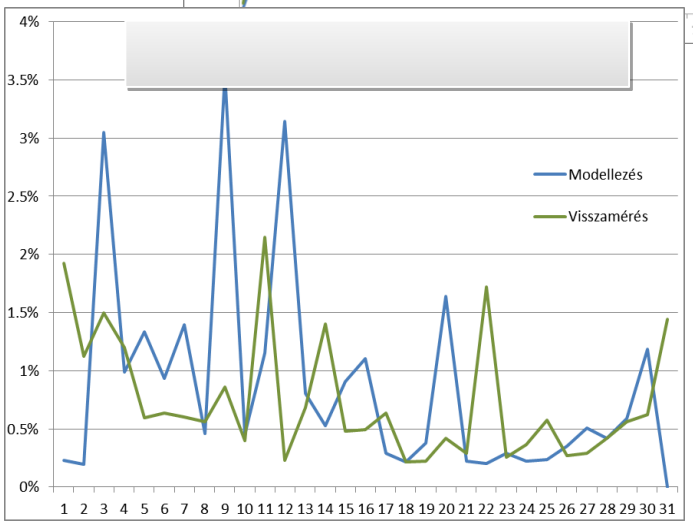
„Time series analysis is similar to sending troops after the battle”

Data preparation









- Missing value handling
- Data transformation
- Outliers
- Correlated variables

Is our modeling method sensitive to them?

Variable selection

Variable Gini

Ginis - DataFrame

Index	Gini
MAG_ATL_TULHIVAS_AVG	70.9
MAG_EFT_FOLY_MAX_NAPOK_SZ_AVG	68.2
V2_TE_MAX_FOGYH_KIHA_AVG	60.3
MAG_ATL_HK_KIHASZN_AVG	58
MAG_MAX_EGYENLEG_AVG	56.9
V2_TE_MAX_A_HK_KIHASZ_AVG	56.9
V2_TX_KIADAS_SUM__AVG	53.6
V2_TX_MAX_JOVAIRAS_AVG	53
V2_PU_JOV_2_AVG	50.6
V2_TE_EGY10_LAK_FOLY_AVG	49.6
V2_TE_EGY20_LAK_FOLY_AVG	47.8
V2_PU_OPER_EGY_AVG	47.2
V2_TE_EGY_LAK_FOLY_AVG	47.2
V2_TE_LA_FO_EGY_AVG_AVG	45.1
MAG_ATL_HK_FELHASZN_AVG	44.6
account_history_in_months	44.4
V2_CS_VAS_DB_AVG	42.7

Format | Resize | Backgroun | Column mir | OK | ance

Correlation matrix

corr_matx_orig - DataFrame

Index	ATL_HK_KIHASZN	MAX_FOGYH_KIH	ATL_TULHIVAS_	MAX_A_HK_KIHA	FOLY_MAX_NAPO	_MAX_EGYENLEG	_MAX_JOVAIRAS	X_KIADAS_SUM__	V2_PU_JOV_2_AVG	_EGY10_LAK_FOL	_PU_O
MAG_ATL_HK_K…	1	0.937	0.775	0.949	0.74	0.86	0.302	0.298	0.268	0.867	0.865
V2_TE_MAX_FO…	0.937	1	0.76	0.978	0.725	0.831	0.307	0.294	0.268	0.839	0.847
MAG_ATL_TULH…	0.775	0.76	1	0.772	0.945	0.629	0.259	0.316	0.235	0.62	0.618
V2_TE_MAX_A_…	0.949	0.978	0.772	1	0.738	0.848	0.305	0.296	0.266	0.857	0.859
MAG_EFT_FOLY…	0.74	0.725	0.945	0.738	1	0.594	0.243	0.307	0.219	0.586	0.585
MAG_MAX_EGYE…	0.86	0.831	0.629	0.848	0.594	1	0.433	0.344	0.39	0.963	0.911
V2_TX_MAX_JO…	0.302	0.307	0.259	0.305	0.243	0.433	1	0.617	0.885	0.315	0.219
V2_TX_KIADAS…	0.298	0.294	0.316	0.296	0.307	0.344	0.617	1	0.635	0.281	0.2
V2_PU_JOV_2_…	0.268	0.268	0.235	0.266	0.219	0.39	0.885	0.635	1	0.275	0.181
V2_TE_EGY10_…	0.867	0.839	0.62	0.857	0.586	0.963	0.315	0.281	0.275	1	0.92
V2_PU_OPER_E…	0.865	0.847	0.618	0.859	0.585	0.911	0.219	0.2	0.181	0.92	1
MAG_ATL_HK_F…	0.929	0.851	0.637	0.866	0.603	0.875	0.131	0.128	0.0947	0.91	0.913
V2_TE_EGY20_…	0.874	0.852	0.626	0.872	0.591	0.953	0.278	0.25	0.235	0.968	0.938
V2_TE_EGY_LA…	0.877	0.857	0.619	0.875	0.585	0.95	0.249	0.224	0.21	0.96	0.961
V2_TE_LA_FO_…	0.867	0.849	0.617	0.862	0.583	0.936	0.23	0.213	0.192	0.943	0.981
account_hist…	0.24	0.233	0.224	0.231	0.208	0.254	0.318	0.299	0.323	0.21	0.193
MAG_MIN_EGYE…	0.875	0.851	0.589	0.869	0.556	0.905	0.151	0.132	0.111	0.942	0.938
V2_TE_EGY_LA…	0.875	0.851	0.589	0.869	0.556	0.905	0.151	0.132	0.111	0.942	0.938
V2_TE_MAX_A_…	0.895	0.82	0.59	0.836	0.558	0.844	0.075	0.0689	0.038	0.888	0.895
V2_CS_VAS_DB…	0.242	0.228	0.259	0.225	0.257	0.256	0.409	0.715	0.447	0.22	0.155
V2_CS_VAS_SU…	0.276	0.273	0.275	0.272	0.268	0.34	0.508	0.72	0.535	0.281	0.205
V2_PU_JOV_1_…	0.193	0.193	0.19	0.19	0.177	0.215	0.381	0.291	0.445	0.134	0.128
V2_PU_JOV_DB…	0.112	0.114	0.123	0.112	0.113	0.116	0.151	0.161	0.274	0.057	0.07
age	0.181	0.183	0.185	0.182	0.172	0.158	0.13	0.0193	0.0899	0.114	0.148
V2_PU_ETTRX_…	0.0832	0.0871	0.0892	0.0825	0.0815	0.127	0.292	0.191	0.368	0.0518	0.034

Format | Resize | Background color | Column min/max | OK | Cancel

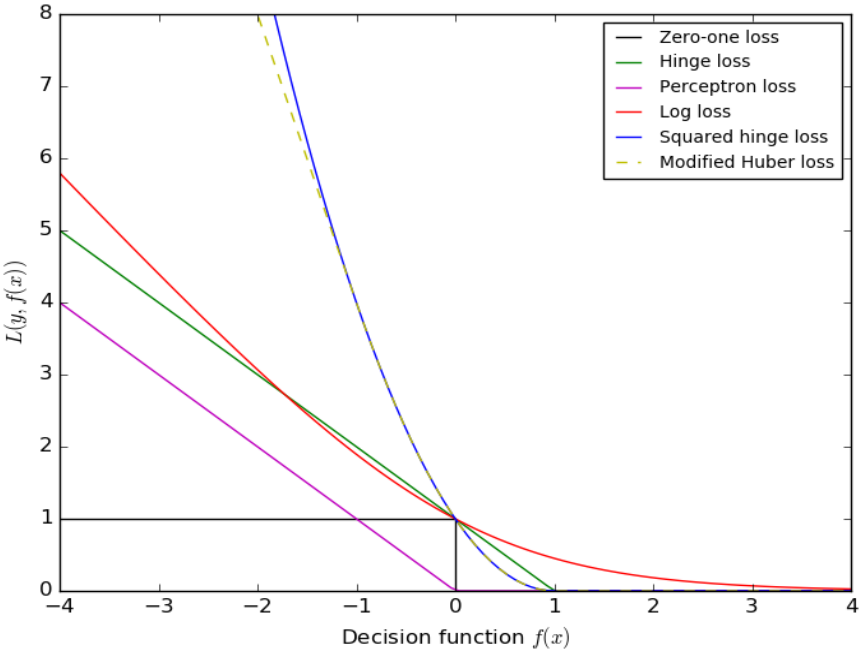
Loss functions

X – the vector space of all inputs,
y– the space of binary targets,
$f: X \to \mathbb{R}$, theestimatorto$y$

We seek to minimize empirical risk:

$$I[f] = \frac{1}{n}\sum L(f(X_i), y_i)$$

$L$is the loss function



Hingeloss: $L(f(x), y)) = |1 - yf(x)|_+$

Squareloss: $L(f(x), y)) = (1 - yf(x))^2$ (extremelypenalizes the outliers)

Huber loss: Quadraticfor$|x| < r$ and linearfor$|x| > r$

Logistic loss:$L(f(x), y)) = \log(e^{-yf(x)} + 1)$

…

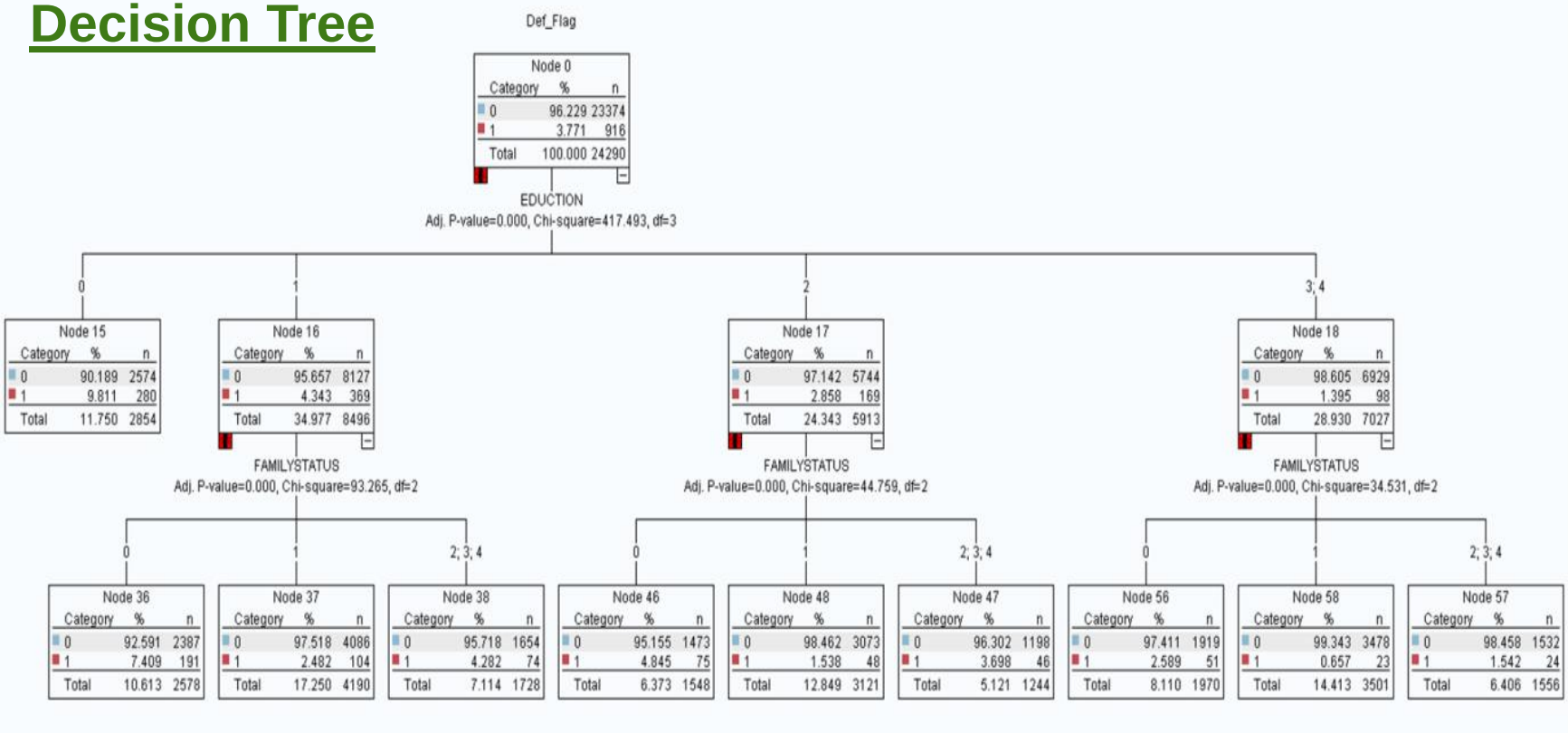
Logistic Regression

We look for the probability of default in form:

$$p = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i X_i)}}$$ or equivalent: $$logit(p) = log\frac{p}{1-p} = \beta_0 + \sum \beta_i X_i$$

Advanced log-loss: $\frac{1}{2}\sum_{i=1}^{n} \beta_i^2 + \mathrm{C} \cdot \log(e^{-y(\beta_0 + \sum \beta_i x_i)} + 1)$

Decision Tree

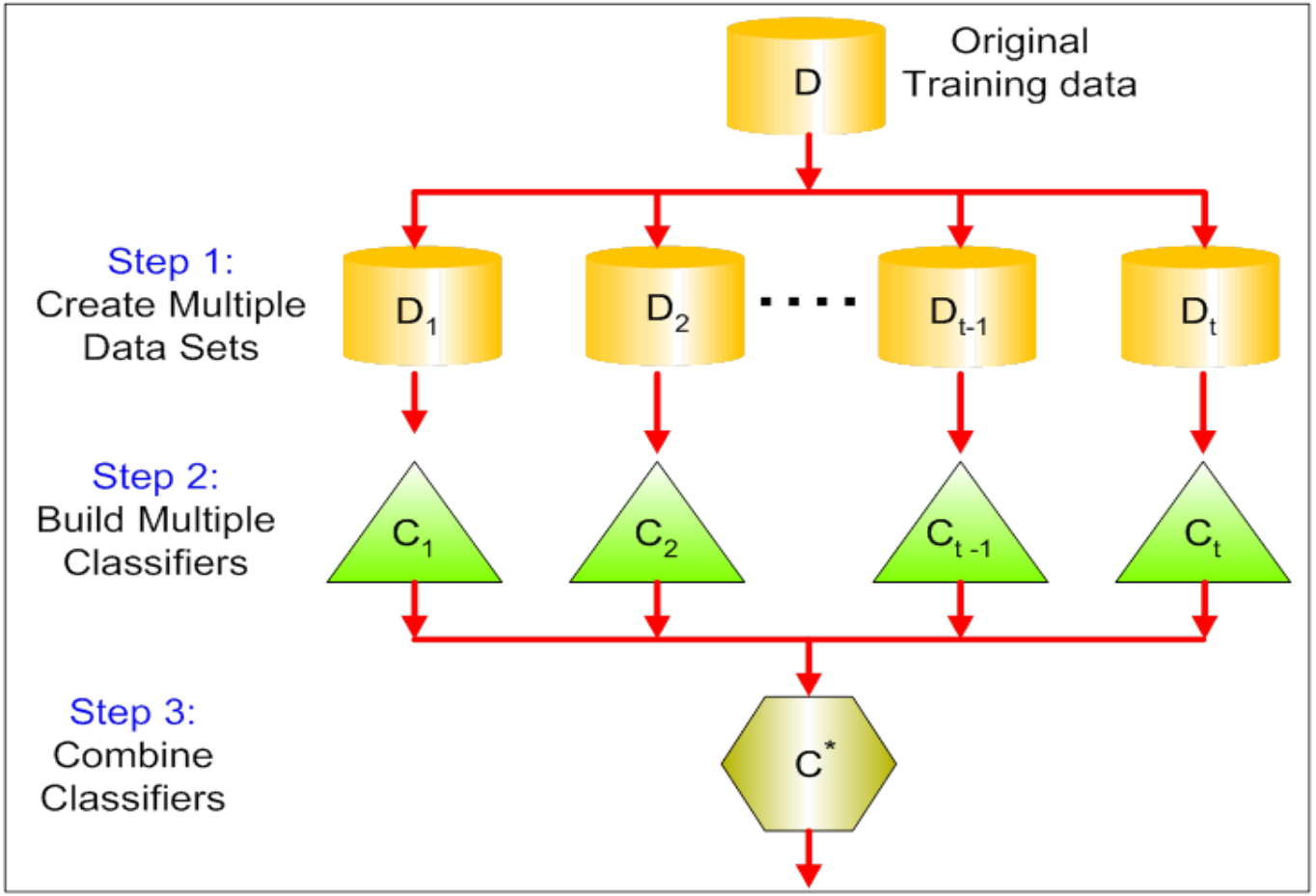


Classic methods are simple and easily interpretable.
Logistic regression is sensitive to missing values, outliers and correlated variables, decision trees are not.

We generally use the combination of the two above methods.
Most often with one-deep trees (decision stumps).

Ensemble methods

Combination of many weak learners.



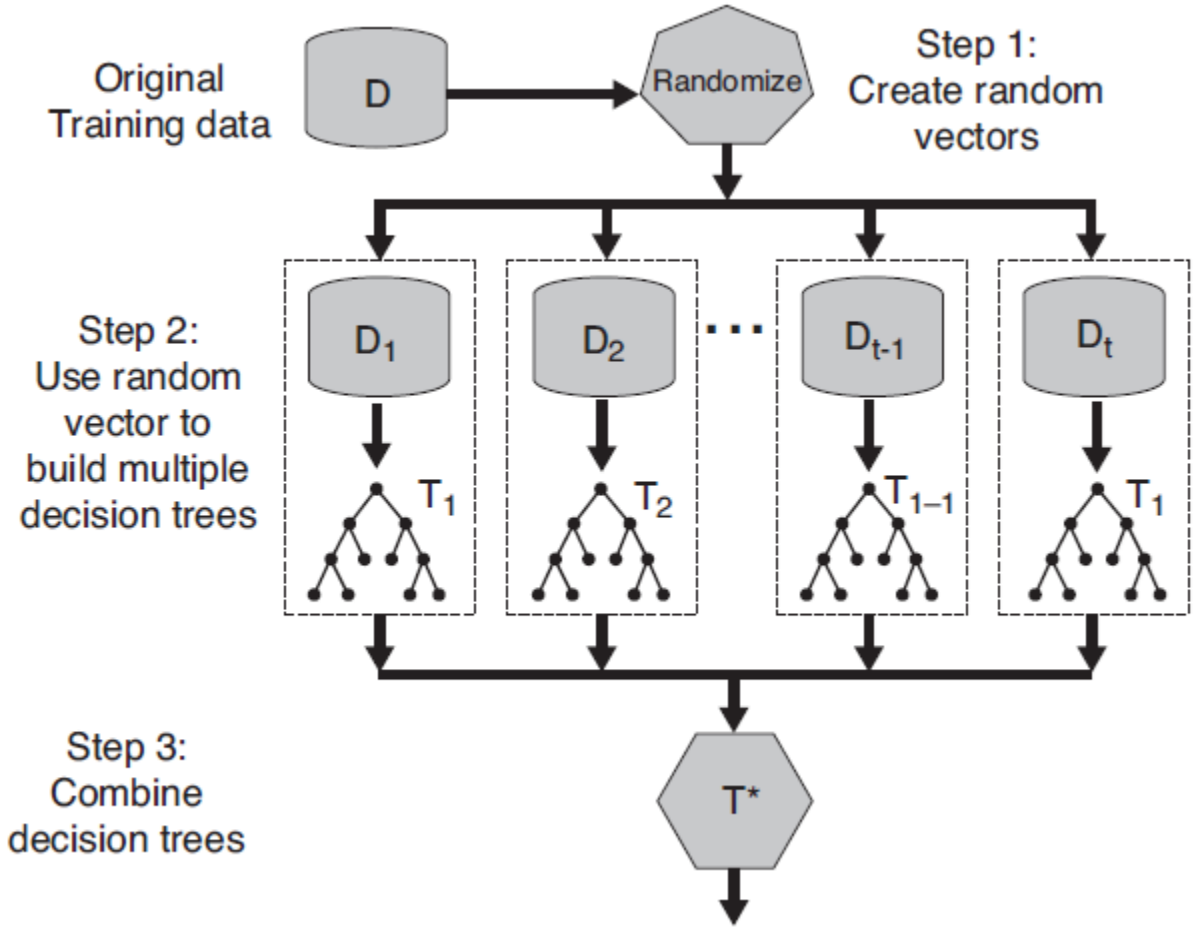
- Random Forest
- LogitBoost
- AdaBoost
- Gradient Boosting Machine

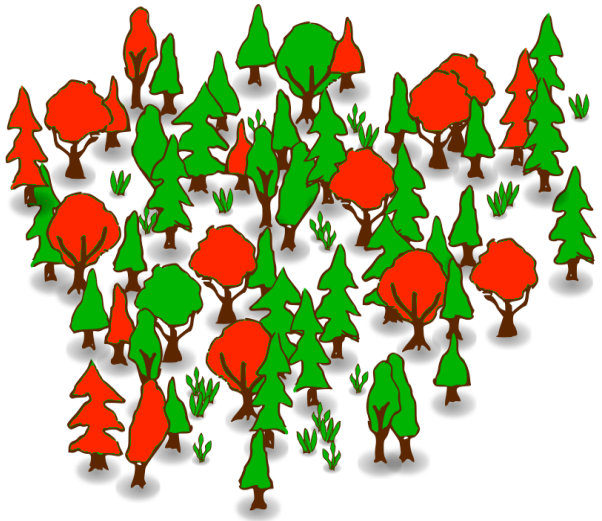
Cooperation with **SZTAKI** and **BME dmlab**

MTA SZTAKI

MŰEGYETEM 1782

Random Forest





GBM algorithm

Input: training set $\{(x_i, y_i)\}_{i=1}^n$, a differentiable loss function $L(y, F(x))$, number of iterations M.

Algorithm:

1. Initialize model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

2. For m = 1 to M:

   1. Compute so-called *pseudo-residuals*:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \qquad \text{for } i = 1, \ldots, n.$$

   2. Fit a base learner $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^n$.
   3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

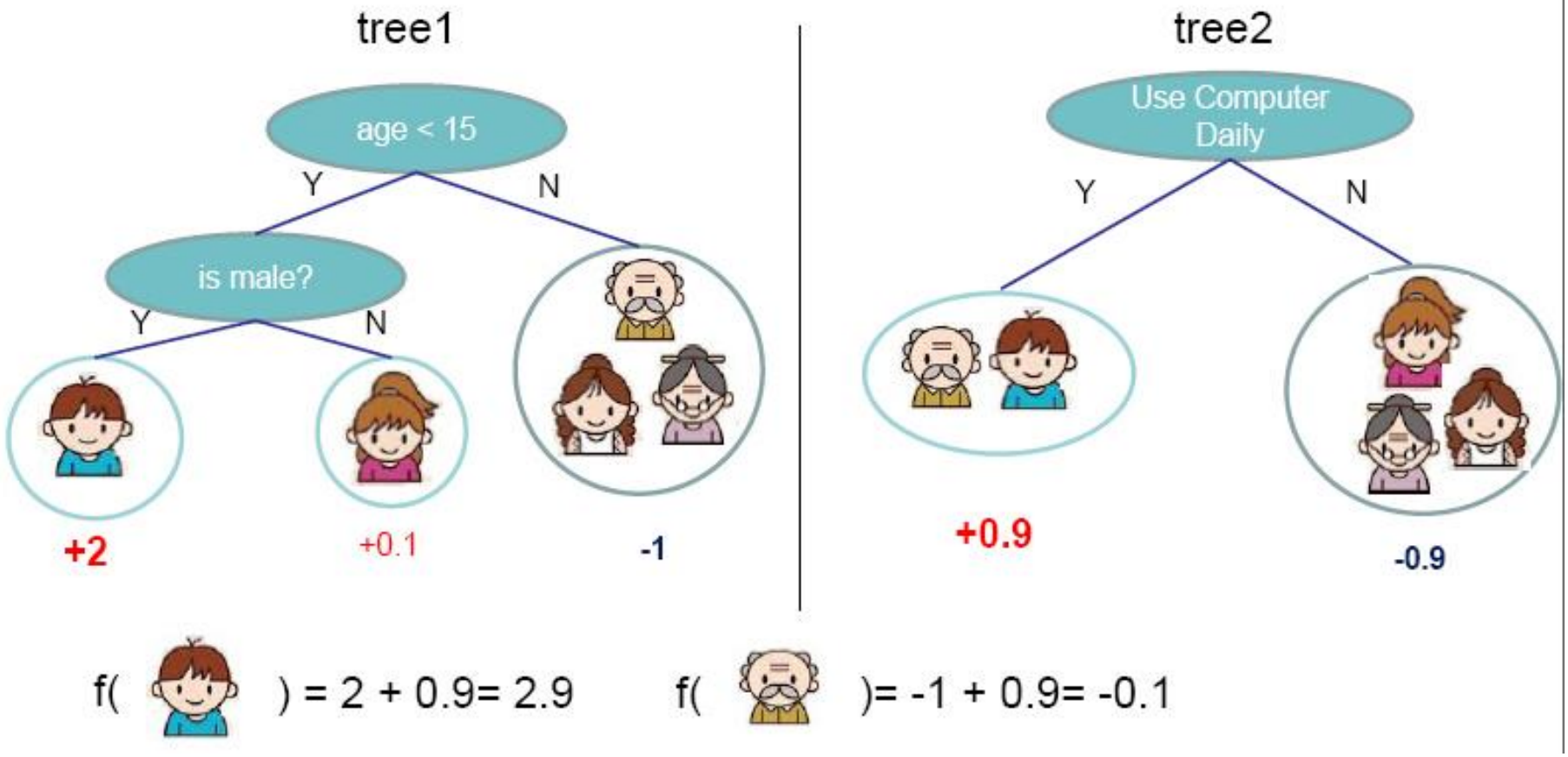
   4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

Gradient Boosting Tree

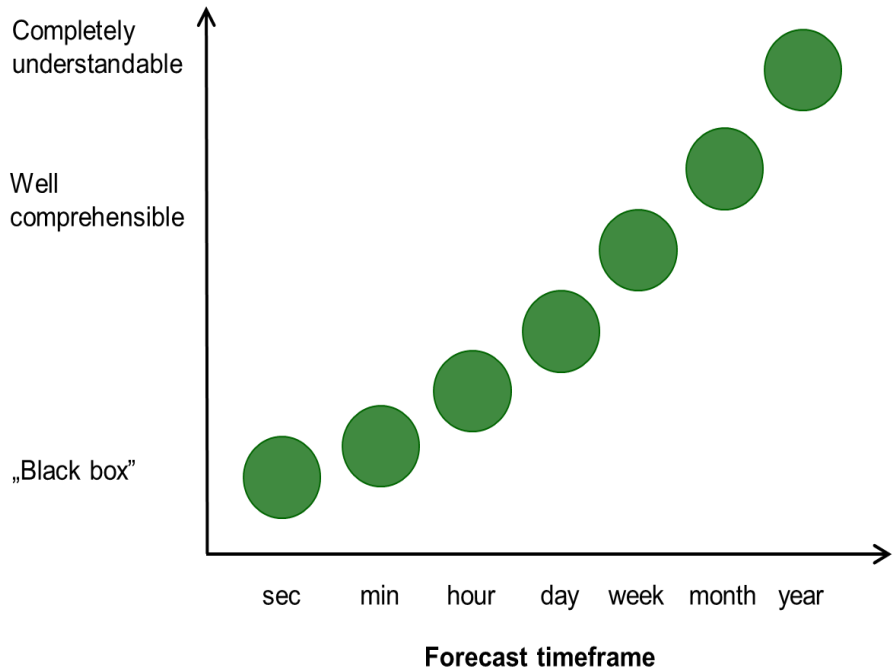
Gradient Boosting Trees = Gradient Boosting Machine, weak learners are simple decision trees (deep 1-4)

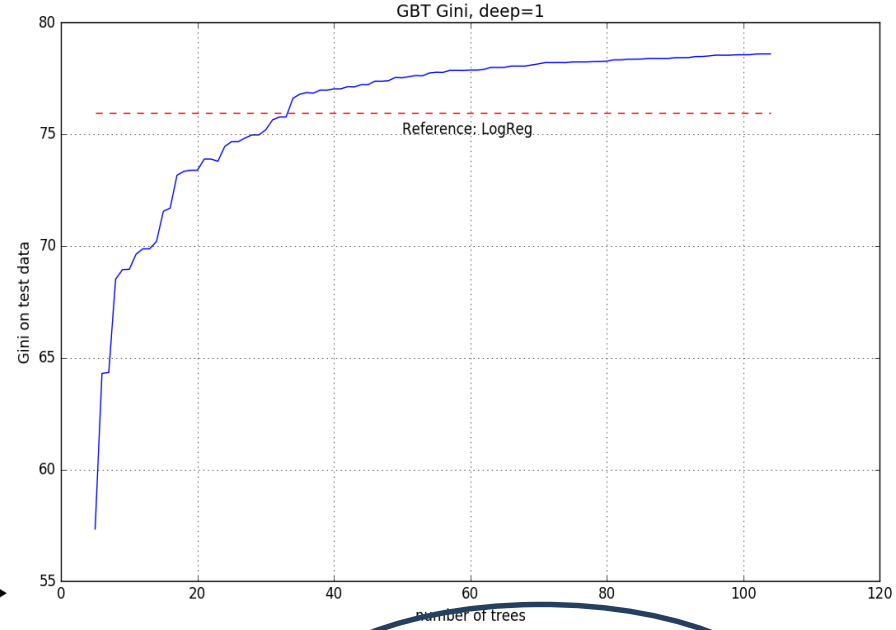


http://zhanpengfang.github.io/418home.html

Classics vs. Ensembles

Keep balance between power, stability, interpretability, simplicity.





„Less is more, and usually more effective”

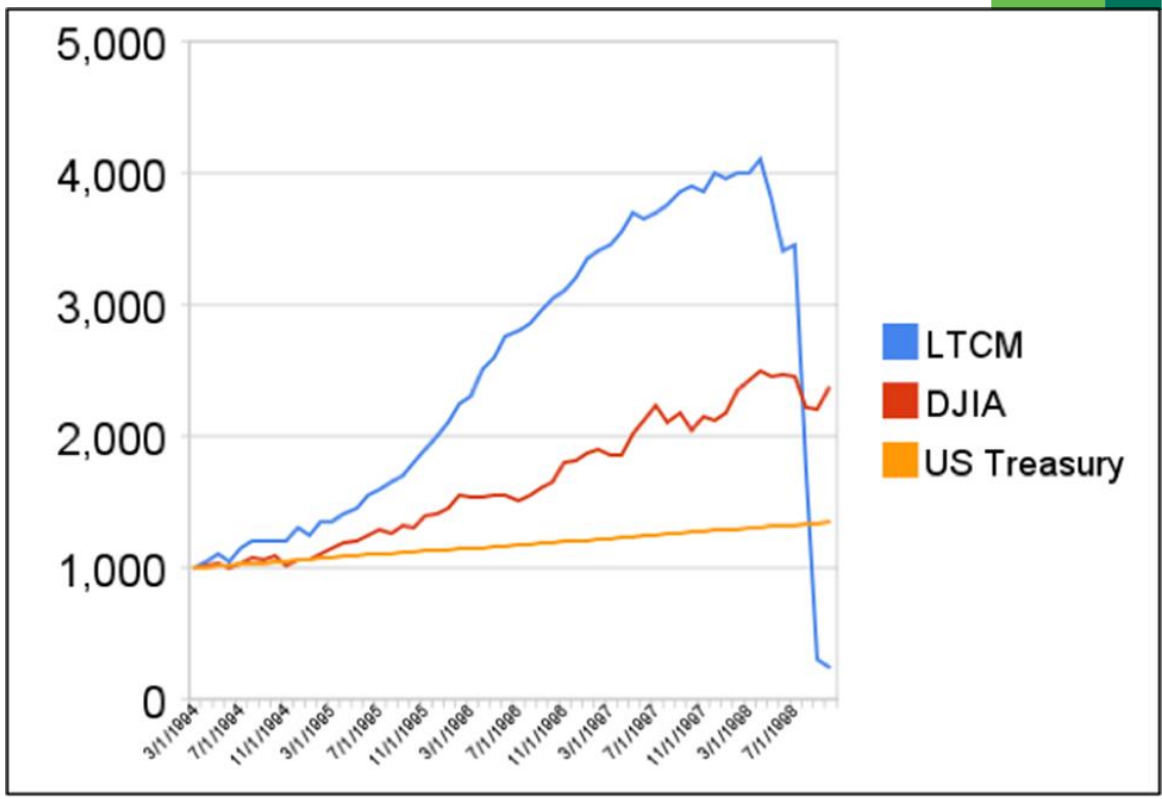
What we do is not „high-frequency trading”, takes months to reveal whether our estimation is good or bad.
We should avoid black-boxes.
We have to understand our models – the knowledge of business experts is essential.

Model error, Model risk

The model itself also entail risk!

Sources of model risk:

- Data errors
- Parameter uncertainty
- Misuse of the model
- …



Long Term Capital Management profit curve
(Scholes, Merton – Nobel prize 1997)

„We go from reality to models not from models to reality"

Propagation of error

Quantification of parameter uncertainty: confidence intervals $I_\alpha = \beta_i \pm z_{1-\alpha/2}\sigma_i$

- A large amount (n) of random numbers x is simulated,according to an even distribution between 0 and 1,$X \sim U(0,1)$.
- For each $k \in \{1..n\}$, a full set of β_i estimators for the logisticregression is simulated through the inverse of its respectivedistribution functions, $\beta_i^k = F_i^{-1}(x_k)$
- The entire portfolio is scored with each set of estimators.

$$f = A \cdot B \quad \Rightarrow \quad \sigma_f^2 \approx f^2 \left[\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2 + 2\frac{cov_{AB}}{AB} \right]$$

Risk cost = PD·LGD·EAD

„Understand model error before you use a model"

Software

	Programming language	*Graphical*
Open source	python™ R	rapidminer orange DATA MINING FRUITFUL&FUN Open for Innovation KNIME
Commercial	MATLAB	sas SPSS IBM

Thanks to…

- My colleagues in OTP Bank
- Benczúr András, SZTAKI
- Gáspár Csaba, BME dmlab
- Nassim Nicholas Taleb (quotes came from *Antifragile* and *Silent Risk*)
- … and many more

Thank you for your attention!