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:

\[
\text{Risk cost} = PD \cdot LGD \cdot EAD
\]

**Risk parameters:**
- **PD** – Probability of Default \([0, 1]\)
- **LGD** – Loss Given Default \([0, 1]\)
- **EAD** – Exposure at Default \([0, \infty)\) - 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)

- Application scorecard for walk-in clients
  - Socio-demographic and financial status, info on employment
  - Application data + behavior data of the credit

- Behaviour scorecards

- Application scorecards for OTP clients
  - Socio-demographic and financial status, info on employment, historical client data
  - Application data + behavior data of the client, network data

- Early default scorecards (Fraud)

Modelling database
CRISP-DM methodology

CRoss
Industry
Standard
Process for
Data
Mining
Scorecard modeling – step-by-step

Statistical scorecard modelling process

Modeling database
  ▶ Variable transformation
  ▶ Missing value analysis
  ▶ Variable selection

Segmentation I.
  ▶ Application part
  ▶ Behavioural part

Segmentation II.
  ▶ Development sample (70%)
  ▶ Validation sample (30%)

Model development
  ▶ Treating missing values
  ▶ Filtering outliers
  ▶ Re-categorizing variables (CHAID)
  ▶ Creating dummy variables
  ▶ Calculating Information Value (IV)

Regression

Backtesting
  ▶ Treating missing values
  ▶ Re-categorizing variables (CHAID)
  ▶ Creating dummy variables

Model performance
Datasets used

Observation period (one year in general)

Modeling dataset
Validation dataset
Stability dataset

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

Correlation matrix
Loss functions

\( X \) – the vector space of all inputs, 
\( y \) – the space of binary targets, 
\( f : X \rightarrow \mathbb{R} \), the estimator to \( y \)

We seek to minimize empirical risk:

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

\( L \) is the loss function

Hinge loss: \( L(f(x), y) = |1 - yf(x)|_+ \)
Square loss: \( L(f(x), y) = (1 - yf(x))^2 \) (extremely penalizes the outliers)
Huber loss: Quadratic for \( |x| < r \) and linear for \( |x| > r \)
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:

\[ \text{logit}(p) = \log \frac{p}{1 - p} = \beta_0 + \sum \beta_i x_i \]

Advanced log-loss:

\[ \frac{1}{2} \sum_{i=1}^{n} \beta_i^2 + C \cdot \log(e^{-y(\beta_0 + \sum \beta_i x_i)} + 1) \]
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
Random Forest

Original Training data

Randomize

Step 1: Create random vectors

Step 2: Use random vector to build multiple decision trees

D_1 -> T_1
D_2 -> T_2
D_{t-1} -> T_{t-1}
D_t -> T_t

Step 3: Combine decision trees

T^*
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:

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

   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 \( \gamma_m \) by solving the following one-dimensional optimization problem:

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

   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.

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.

„Less is more, and usually more effective”
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 \( \beta_i \) estimators for the logistic regression is simulated through the inverse of its respective distribution functions, \( \beta_i^k = F_i^{-1}(x_k) \).
- The entire portfolio is scored with each set of estimators.

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

Risk cost = PD \cdot LGD \cdot EAD

„Understand model error before you use a model”
## Software

<table>
<thead>
<tr>
<th>Programming language</th>
<th>Graphical</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Open source</strong></td>
<td></td>
</tr>
<tr>
<td>Python</td>
<td>RapidMiner</td>
</tr>
<tr>
<td>R</td>
<td>Orange</td>
</tr>
<tr>
<td>KNIME</td>
<td></td>
</tr>
<tr>
<td><strong>Commercial</strong></td>
<td></td>
</tr>
<tr>
<td>MATLAB</td>
<td>SAS</td>
</tr>
<tr>
<td></td>
<td>IBM</td>
</tr>
</tbody>
</table>
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!