Credit Risk Models Cross-Validation – Is There Any Added Value?

Croatian Quants Day

Zagreb, June 6, 2014

Vili Krainz

vili.krainz@rba.hr

The views expressed during this presentation are solely those of the author

Introduction

- Credit risk – The risk that one party to a financial contract will not perform the obligation partially or entirely (default)
- Example – Bank loans
- The need to assess the level of credit risk – credit risk rating models (credit scorecards)
- Problem – to determine the functional relationship between obligor or loan characteristics *X1, X2,* ... , *Xn* (risk drivers) and binary event of default (0/1), in a form of latent variable of probability of default (PD)

Scorecard Development Process

- Potential risk drivers – retail application example
  - Sociodemographic characteristics:
    - Age, marital status, residential status...
  - Economic characteristics:
    - Level of education, profession, years of work experience...
  - Financial characteristics:
    - Monthly income, monthly income averages...
  - Stability characteristics:
    - Time on current address, current job...
  - Loan characteristics:
    - Installment amount, approved limit amount, loan maturity...

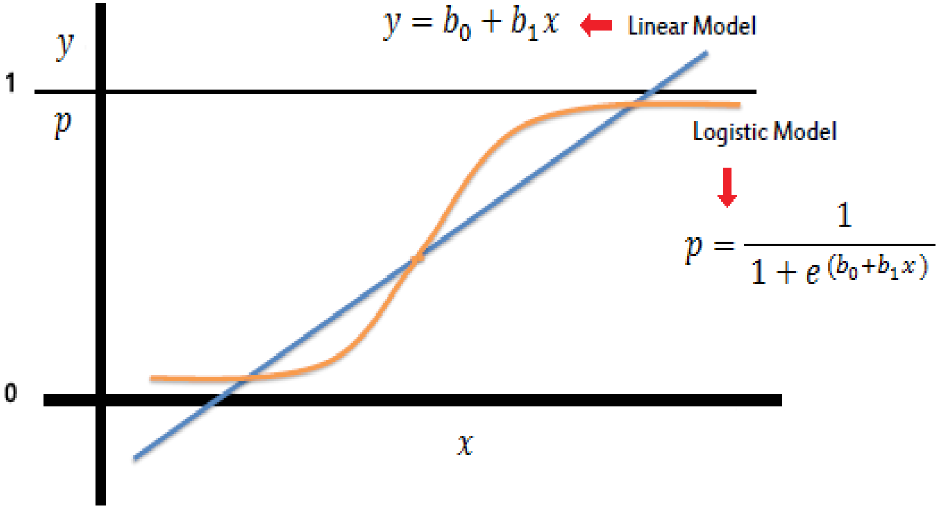
Scorecard Development Process

- Univariate analysis – analysis of each individual characteristic
  - Fine classing – division of numeric variables into a number (e.g. 20) of subgroups, analysis of general trend
  - Coarse classing – grouping into (2-5) larger classes to optimize predictiveness, with certain conditions (logical, monotonic trend, robust enough...)

Age	Bad rate
<30	3.47%
[30, 55]	2.86%
>55	1.73%

Scorecard Development Process

- Multivariate analysis
  - Correlation between characteristics
  - Logit model – most widely used
  - Logistic regression (with selection process)

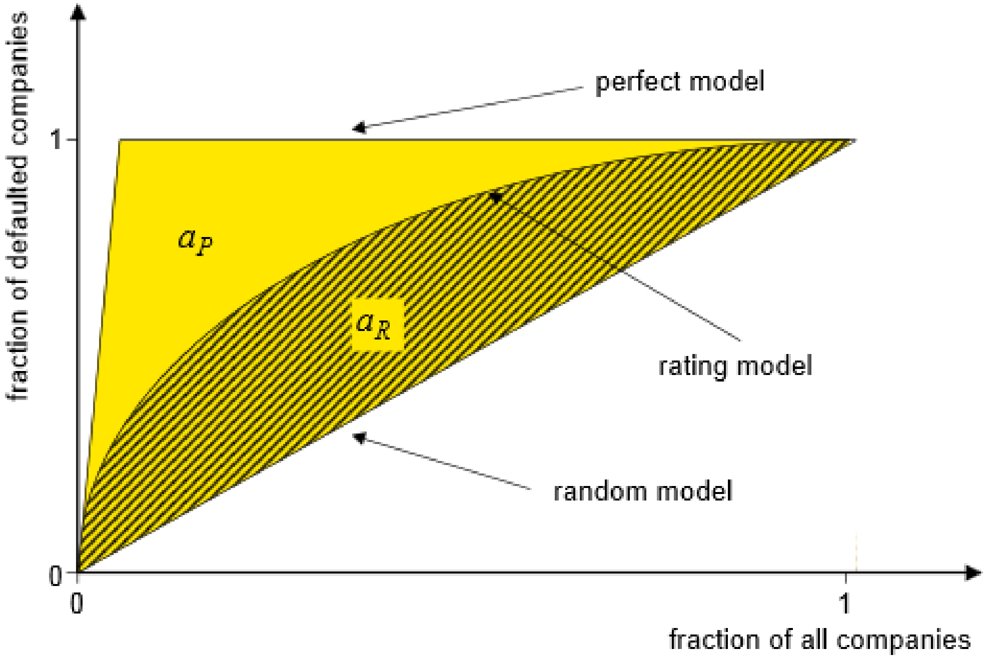


$$score_i = \ln\left(\frac{1 - PD_i}{PD_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

$$\ln\left(\frac{1 - PD_i}{PD_i}\right) = score_i \quad \Leftrightarrow \quad PD_i = \frac{1}{1 + e^{score_i}}$$

Scorecard Model Predictiveness

- The goal of a scorecard model is to discriminate between the good and the bad applications
- Predictivity is most commonly measured by Gini index (a.k.a Accuracy Ratio, Somers' D)



$$Gini = \frac{a_R}{a_R + a_P}$$

Scorecard Model Cross-Validation

- At model development start, the whole data sample is split randomly (70/30, 75/25, 80/20...)
- The bigger sample is used for model development, while the smaller sample is used for cross-validation
- Model's predictive power (Gini index) is measured on the independent, validation sample
- Done to avoid overfitting
- The predictive power shouldn't be much lower on the validation sample than it is on the development – that's when the validation is considered successful

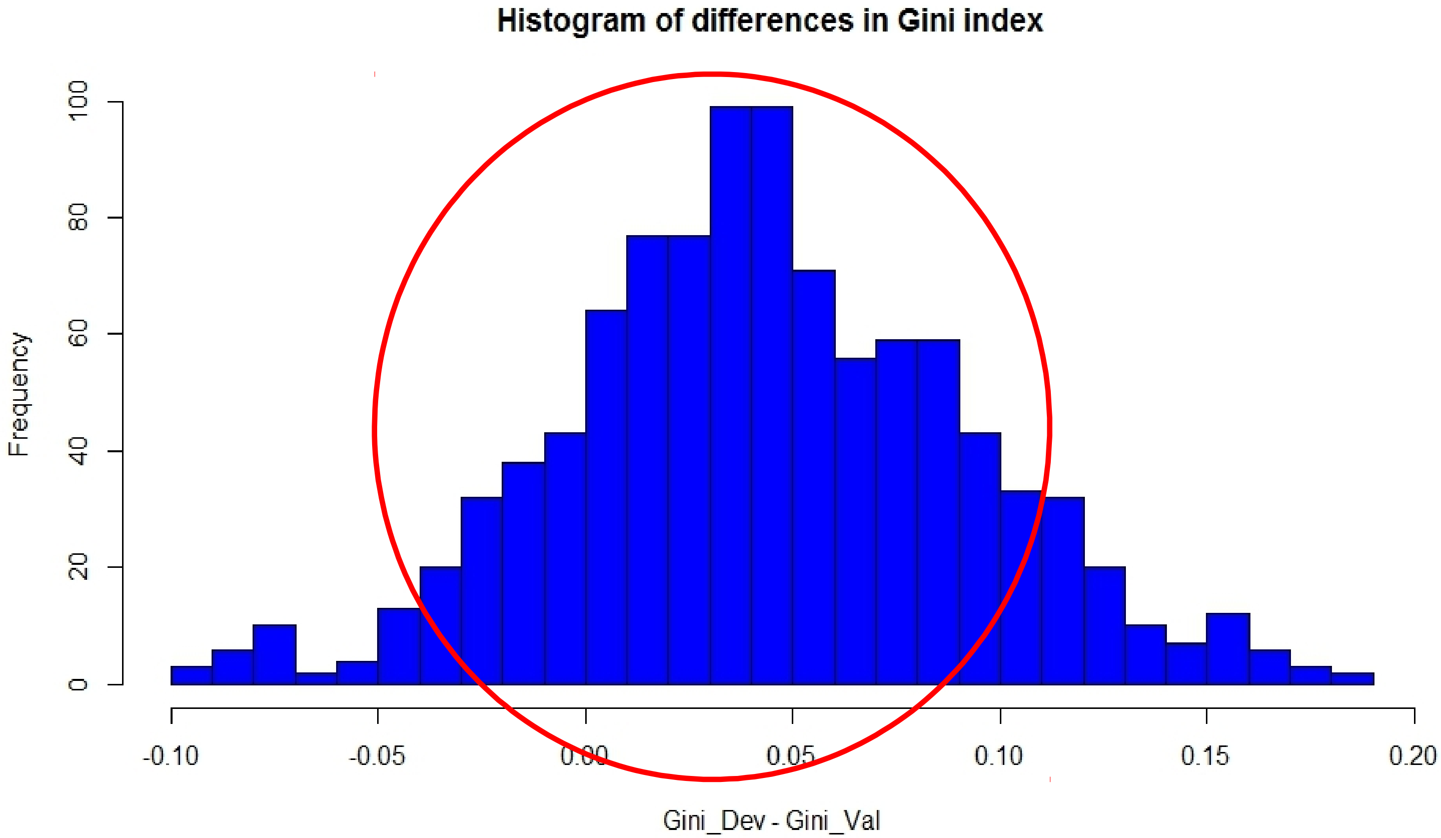
What if validation fails?

- Is it possible if everything is done „by the book”?
- Does that mean that:
  - Something was done wrong in model development process?
  - The sample is not suitable for modeling at all?
  - The process needs to be repeated?

Monte Carlo Simulations

- Real (masked) publicly available retail application data (Thomas, L., Edelman, D. and Crook, J., 2002. *Credit Scoring and Its Applications*. Philadelphia: SIAM.)
- 1000 simulations of model development process in R
  - Each time stratified random sampling (75/25) was done (on several characteristics, including the target variable – default indicator)
  - Fine classing for the numeric variables
  - Coarse classing all the variables using the code that simulates modeler's decisions
  - Stepwise logistic regression using AIC
  - Measuring Gini index on development and validation sample
- Pre-selection of characteristics for the business logic and correlation
- One reference model was built on whole data sample

Results

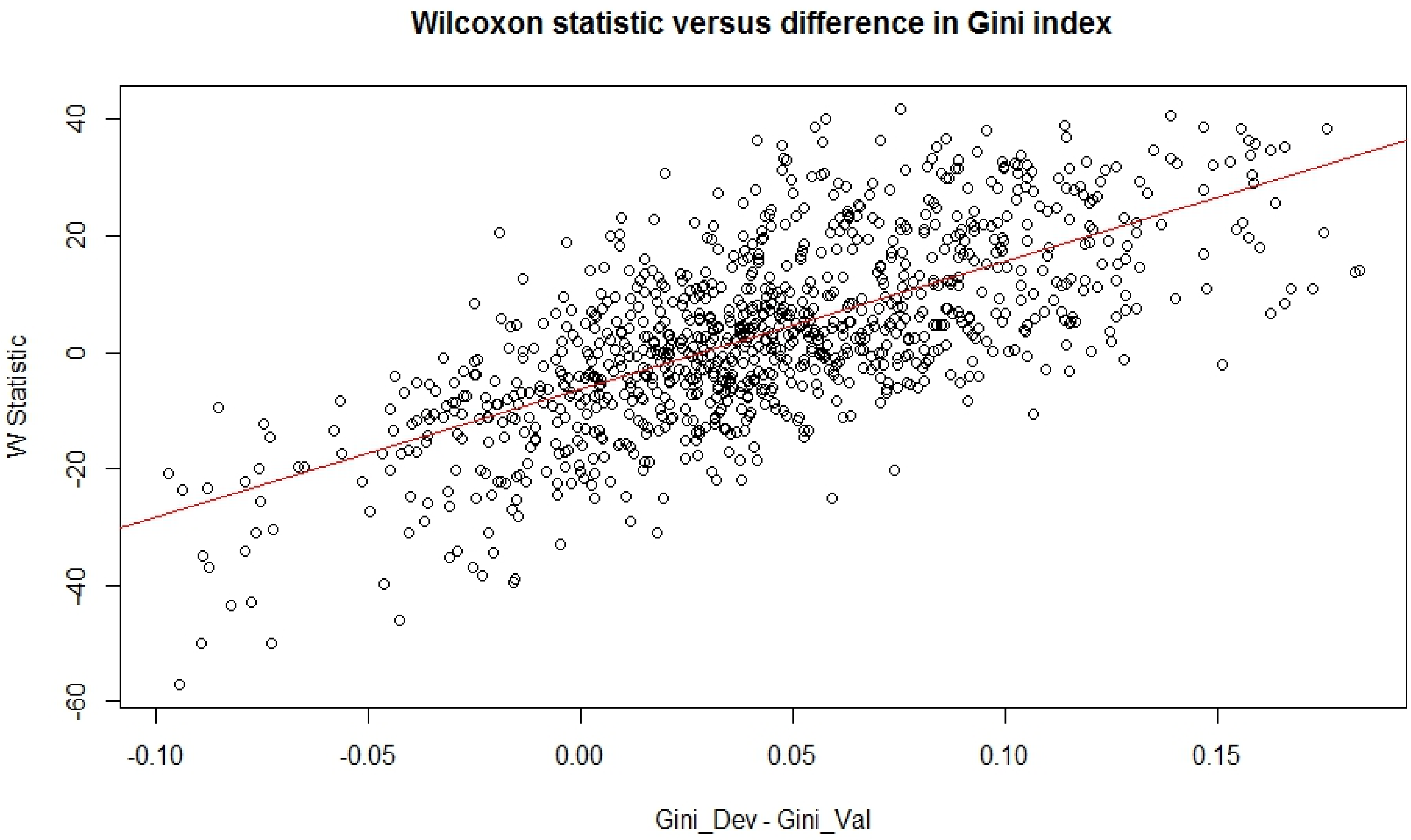


- In 12.5% of cases we get a difference bigger than 0.1
- Pearson’s chi-square test – all characteristics of all 1000 samples representative at 5% significance level

Results

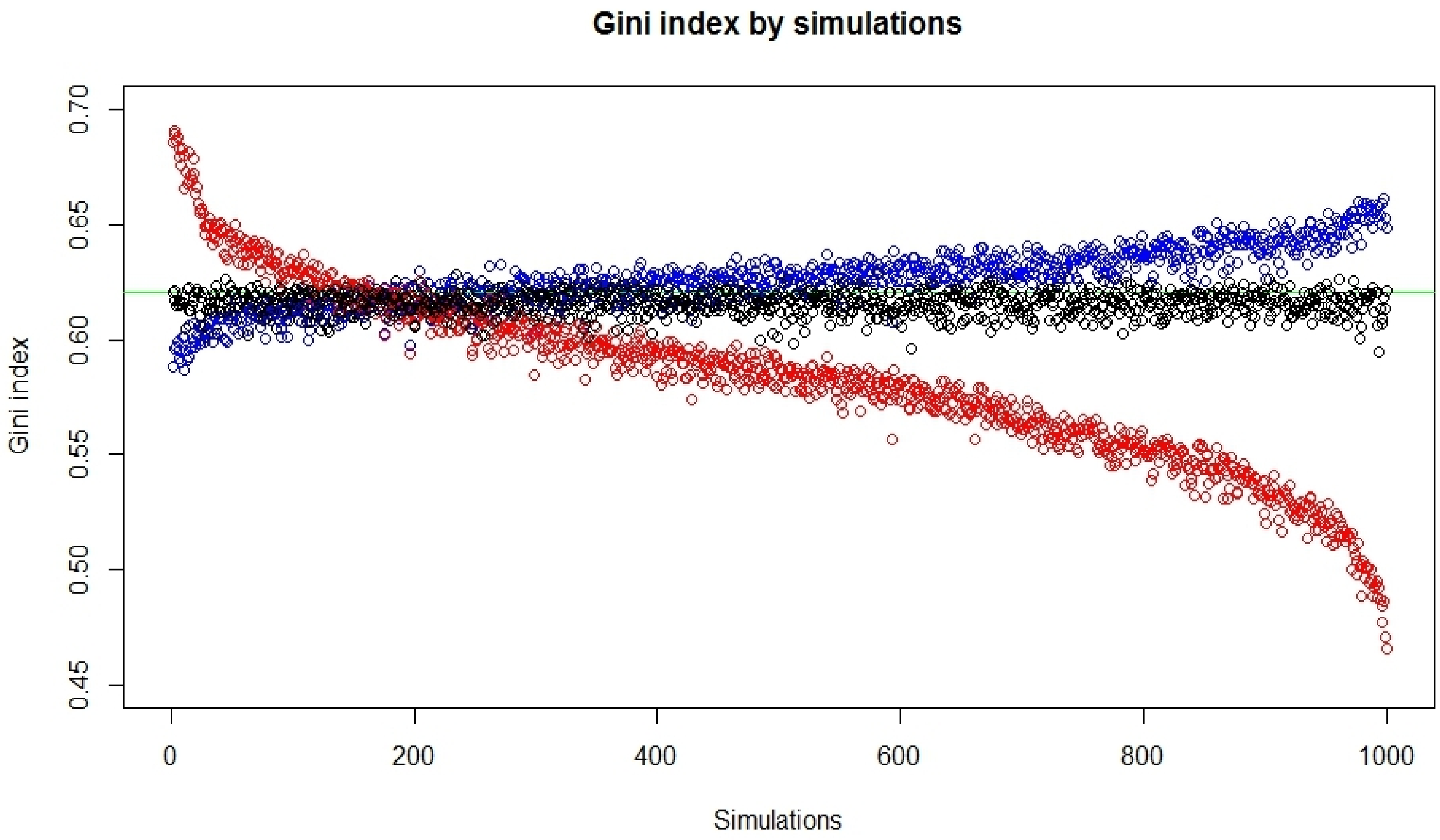
- Idea: Compare the scores from each simulation model to reference model (on the whole sample) and relate to differences in Gini
- If there is a strong connection – we strive to get a model similar to the reference model
- Wilcoxon paired (signed rank) test
  - H0: median difference between the pairs is zero
  - H1: median difference is not zero.
- Basically, the alternative hypothesis states that one model results in systematically different (higher or lower) scores than the other

Results



- Correlation: 0.68

Results


Gini index by simulations
Gini index
0.45
0.50
0.55
0.60
0.65
0.70
0
200
400
600
800
1000
Simulations

From The Simulations...

- Regardless of a modeling job done right, validation can fail by chance
- We like to have Gini index on the development sample “similar” to the one on the validation sample – we tend to get the model that is more similar to the reference model – why not develop on the whole sample in the first place?
- Regardless of validation results and difference in Gini, predictive power on the whole data sample does not vary too much

Instead Of A Conclusion...

- Does this method of cross-validation bring any added value?
- It may be more important to check whether all the modeling steps have been performed carefully and properly, and that best practices are used, in order to avoid overfitting
- Can any cross-validation method can offer real assurance or does the only real test come with future data?

Thank You!

vili.krainz@rba.hr