CRC | Credit Research Centre

CREDIT SCORING & CREDIT CONTROL XIII

28-30 AUGUST 2013

Scorecard Development Beyond the Standard Textbook

A Case Study

Dr. Hendrik Wagner, RiskParameters.EU

Dr. Farkas Bagaméry, Lombard Lízing

Edinburgh, CSCCXIII, August 28-30 2013

Agenda

- Part1: Weight of Evidence over time (15 minutes)
- Part2: Other topics (5 minutes)
  - Variable-length performance windows
  - Integrating Score-PD mapping and Virtual CT

Part1:
Weight of Evidence Over Time

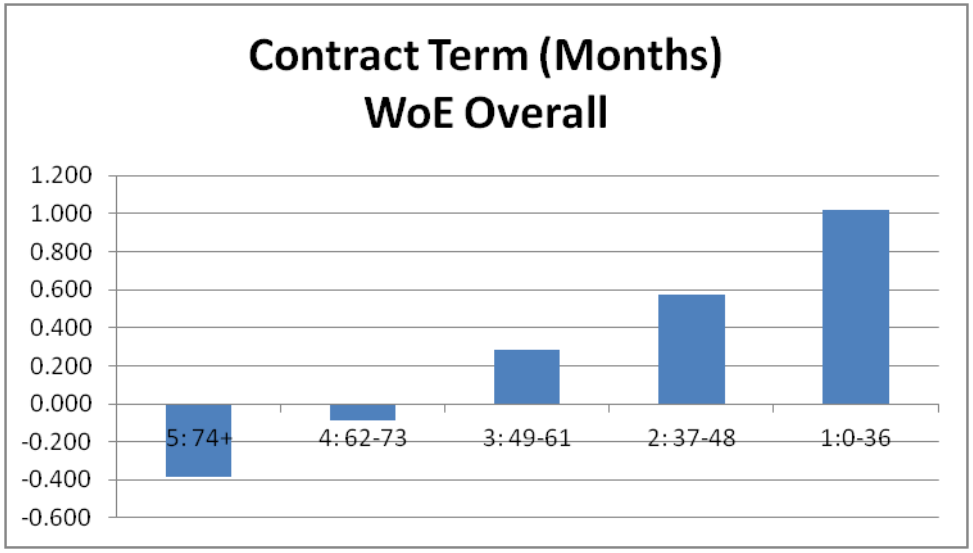
Message

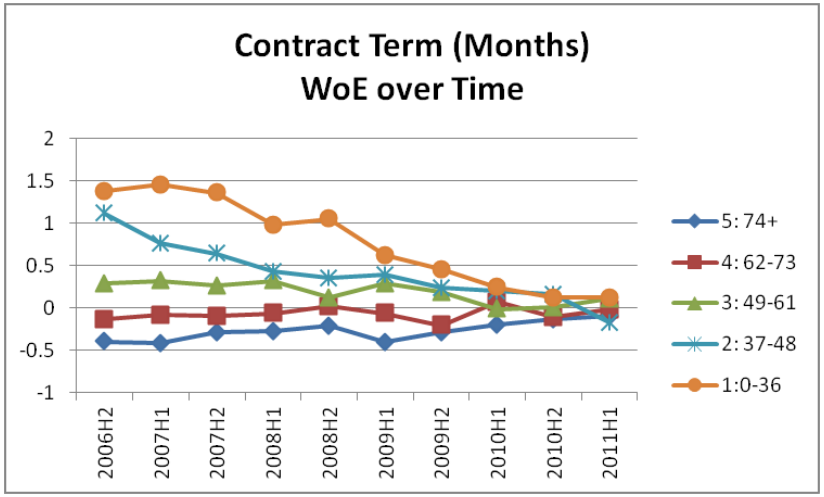
- When developing a scorecard, consider Weight of Evidence values over time and not just over the development sample as a whole
- Don’t assume the attribute/Weight of Evidence relationships would be constant

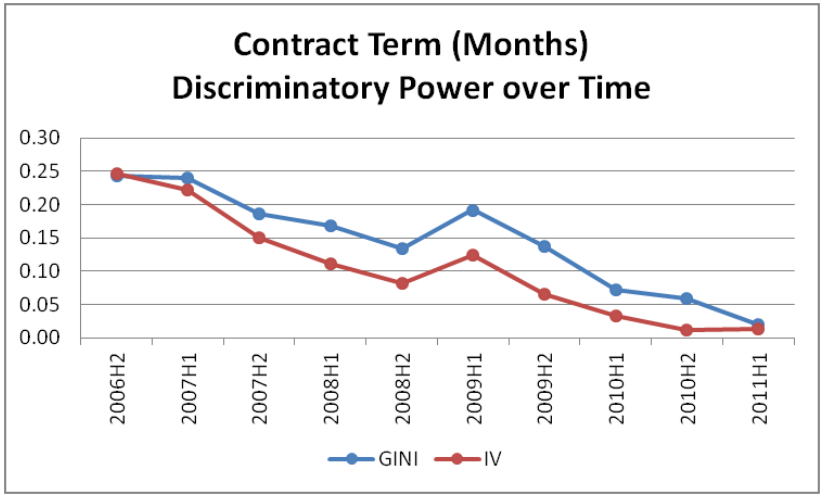
Car finance case study: Application scoring using contract properties

- Both the term of the contract and the proportion of down payment can frequently be found in application scorecards for secured lending, such as car finance
- Their Weight of Evidence (WoE) values usually show a strong monotonic trend across attributes
  - Higher contract duration resulting in higher default rates
  - Higher down payment proportion resulting in lower default rates
- However, as contract properties that can be influenced by the lender they can be a major source of instability

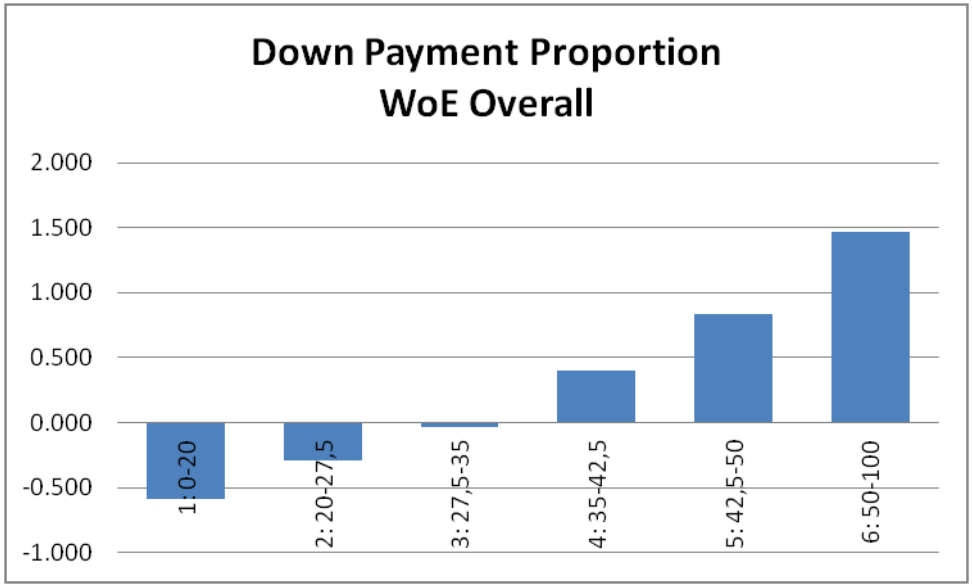
Case Study: Contract Term - WoE

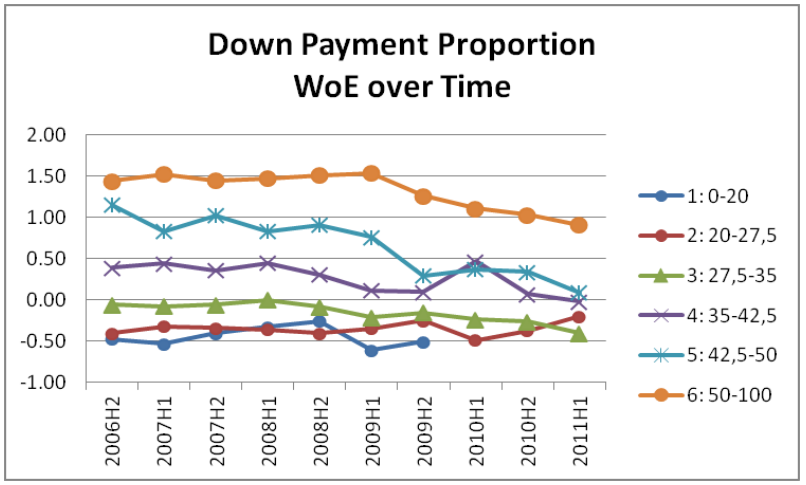


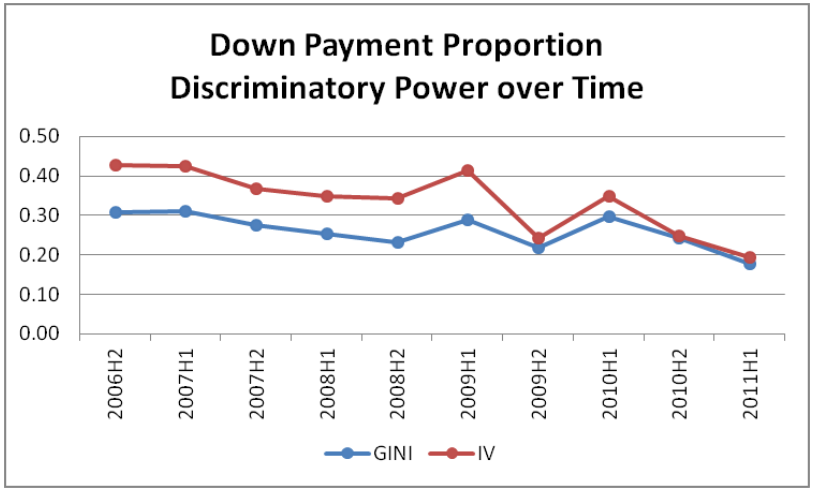




Case Study: Down Payment Proportion - WoE



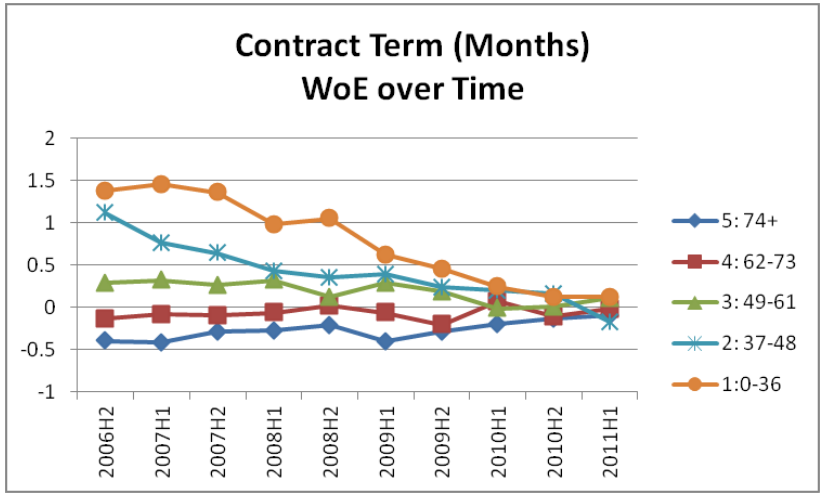


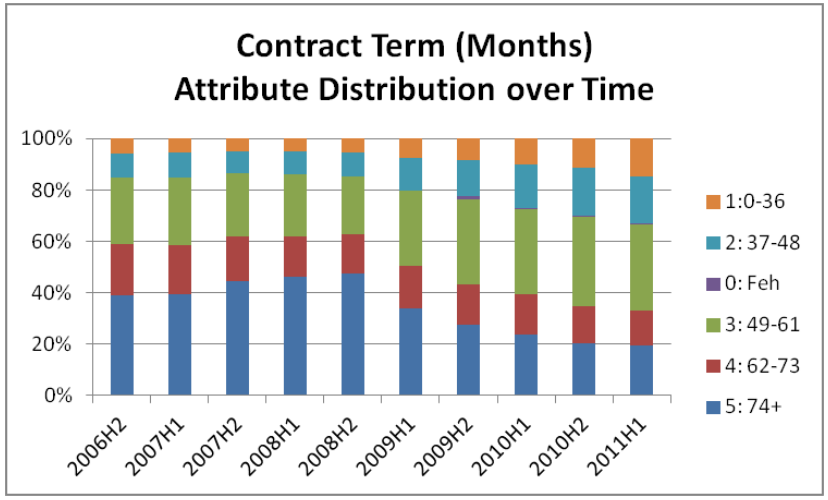


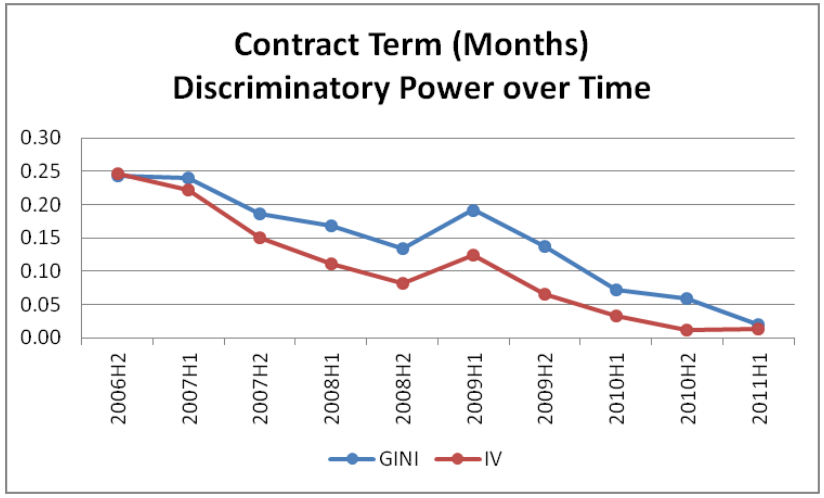
PSI is not to the point

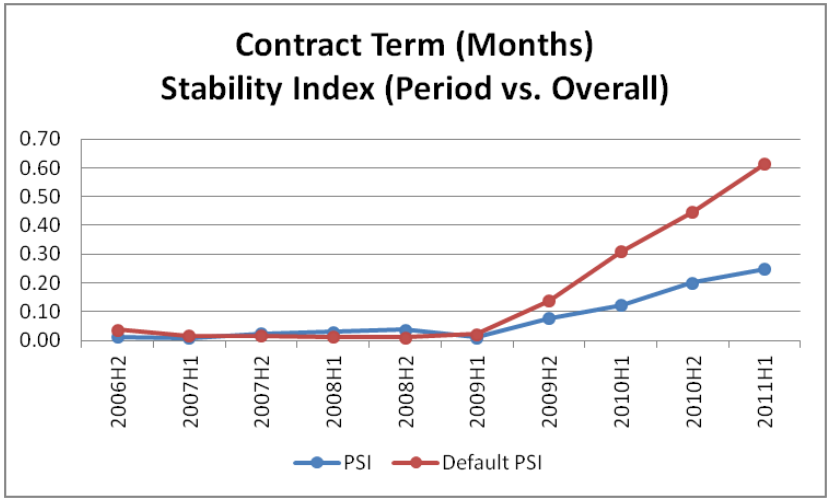
- The Population Stability Index is a commonly used stability measure, however, it
  - addresses only the stability of the relative attribute frequencies (attribute distribution)
  - does not address the stability of the relative attribute default rates (Weights of Evidence)
- A characteristic may still be equally predictive and well calibrated, even if the attribute distribution changes. By contrast, its quality will suffer (or improve) from a change in the Weights of Evidence
- The PSI on defaults-only (default stability index) is also only partially to the point. It measures a kind of combination of distribution and default rate change and therefore can be dominated by irrelevant distribution change

Case Study: Contract Term – PSI





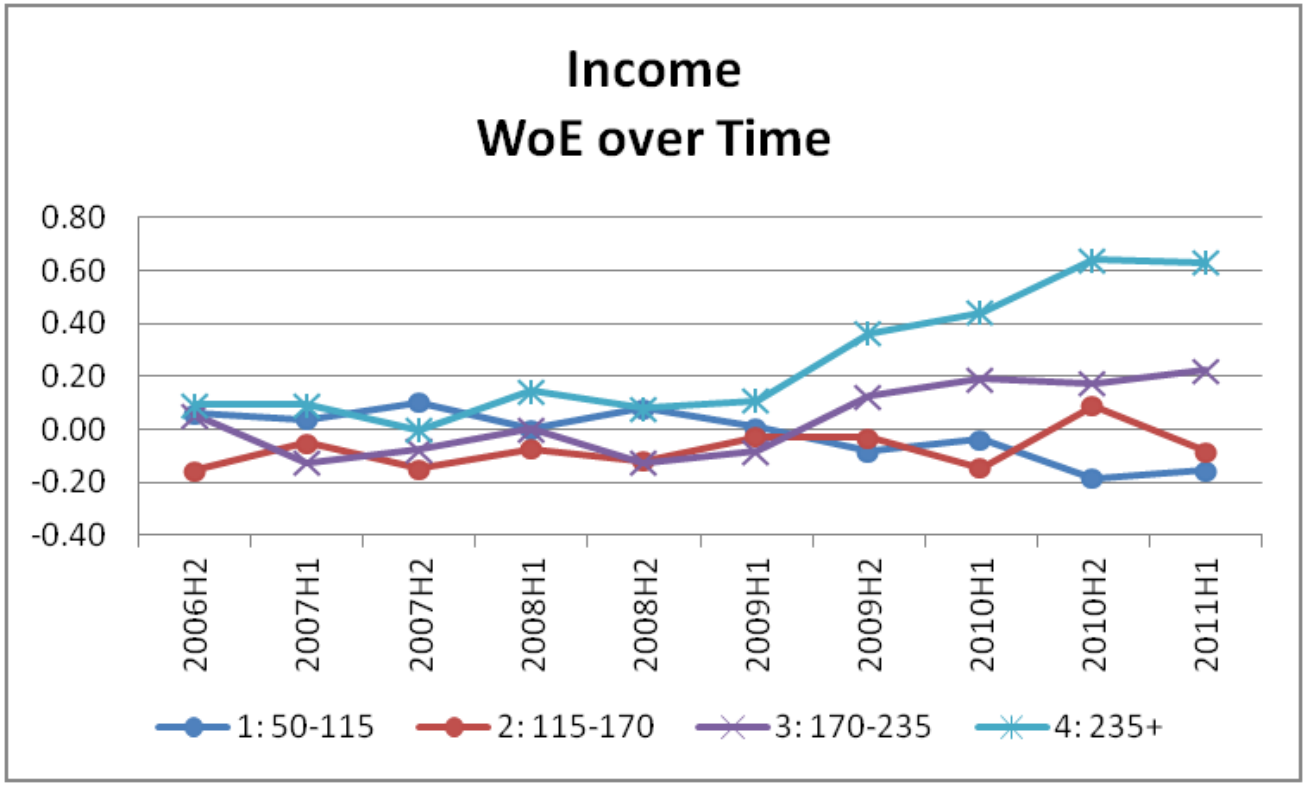




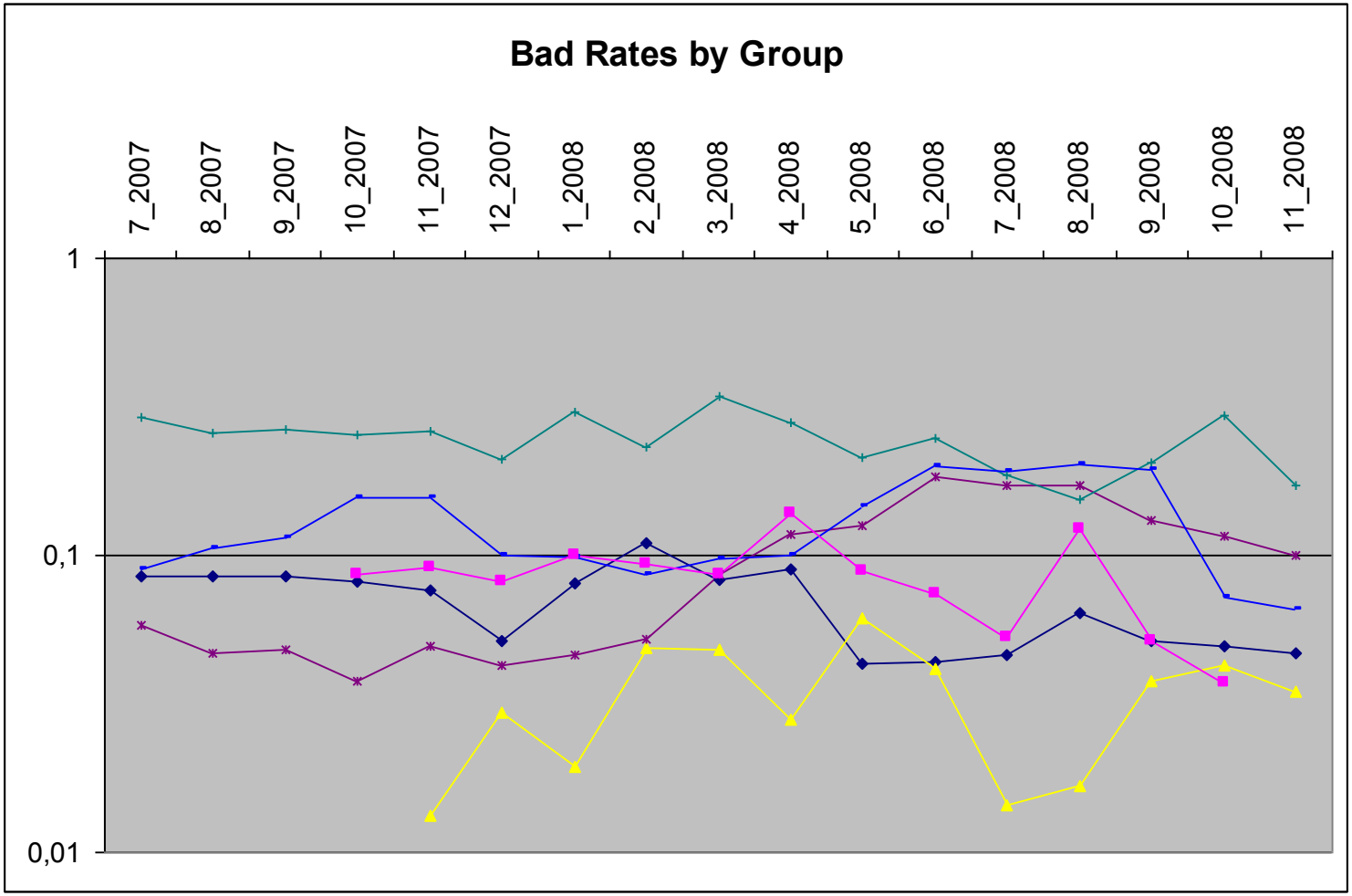
Use a long sample: Informed stability vs. power trade-off

- Verifying a characteristic's WoE-values over time on a long sample allows for an informed decision on the trade-off between its stability and power
- Different measures can then be adopted depending on different types of instability issues, including using a shorter sample window with only more recent cases
- It is not sufficient to simply pick a rather short sample from the start and rely on monitoring and re-modeling
  - It is necessary to know beforehand how often one will be forced to remodel and because of which characteristic
  - High instability may necessitate an unrealistically high remodeling frequency

Case Study:
Example of an improving characteristic: Income



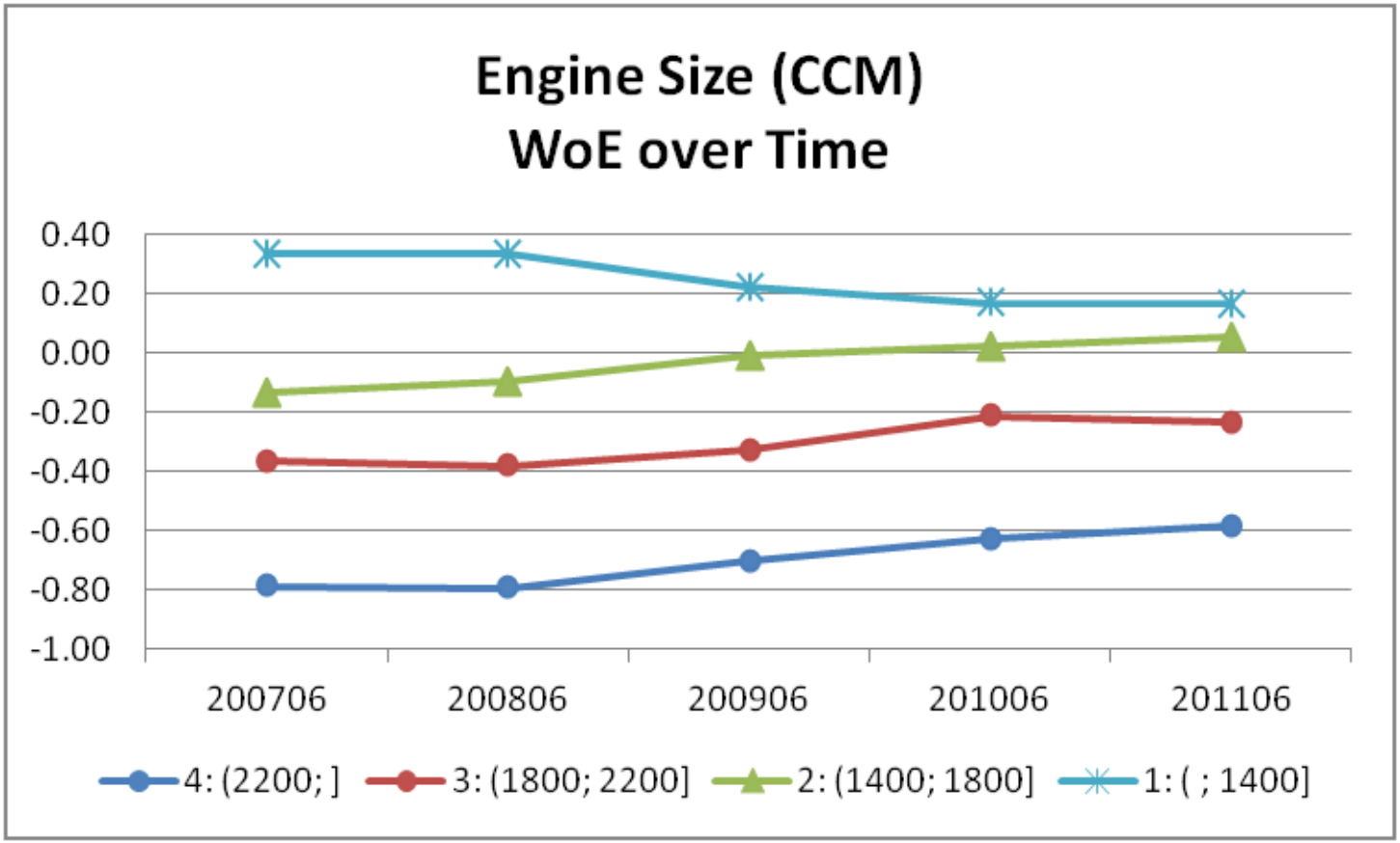
Excursion: Quickly changing telco tariffs



WoE over time in behavioural scoring

- In application scoring, the time axis displays cohort start dates
- In behavioural scoring the time axis displays snapshot dates

Case Study: Example of application data losing power in behavioural scoring



How to quantify WoE-instability: Fixed Gini

- Measure change of discriminatory power
  - Use a “fixed Gini”, where the overall sample is used for fixing the attribute sort order
  - Information Value or “time period wise Gini” (where sort order is determined by time period) are not suitable, as they can be constantly high, even when attributes change rank from one period to the next

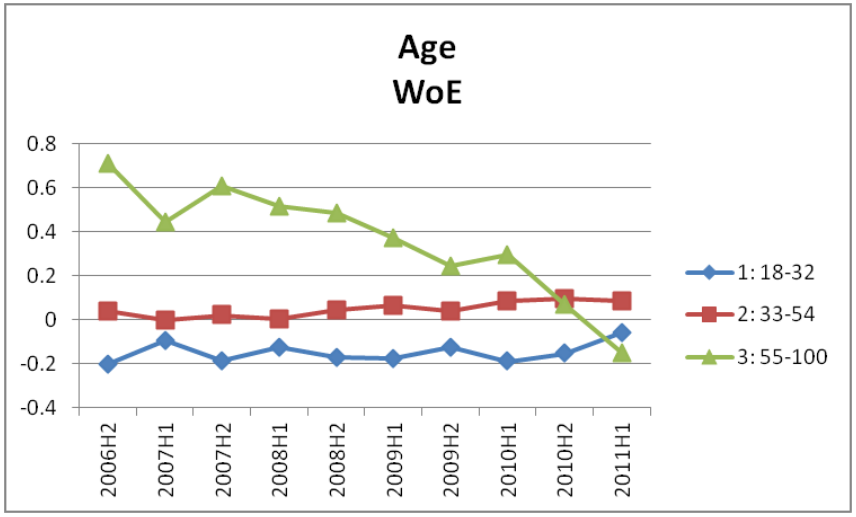
How to quantify WoE-instability: Scaled Calibration

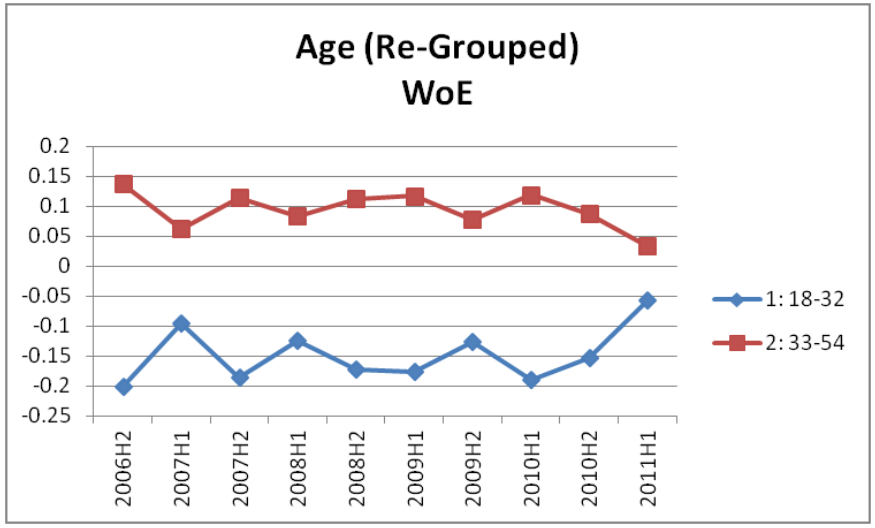
- Scaled default rate calibration tests
  - First transform each time period's WoE values into scaled attribute default rates
    - Use the overall default rate over both time and attributes as the value for WoE=0. This is equivalent to a Bayes scaling of each period to the overall default rate and necessary to eliminate the effect of changes to the default rate over time
  - Then compare scaled time period attribute default rates to overall average attribute default rates
  - Use for example standard default rate calibration tests, such as
    - Binomial test on each attribute
    - Hosmer-Lemeshow over all attributes

Scaled calibration tests: Footnotes

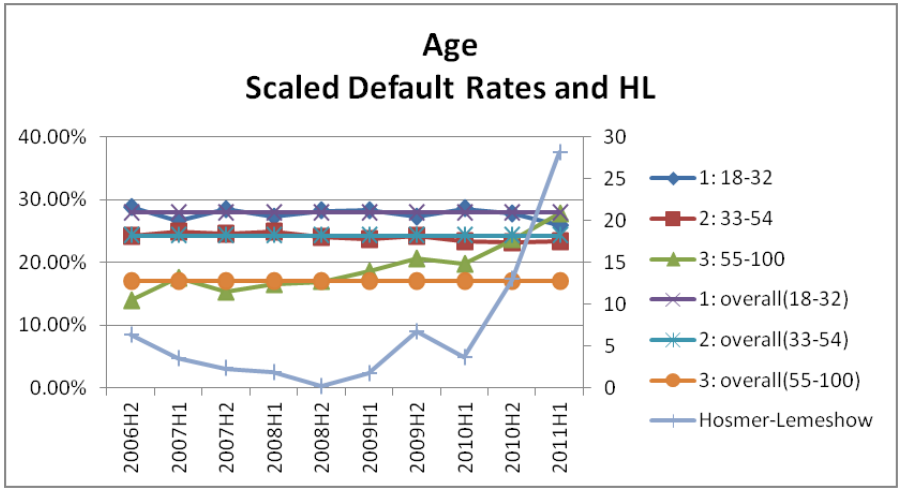
- The following is equivalent:
  - Transforming attribute default rates into Weight of Evidence values, then re-transforming the Weight of Evidence values into new default rates using a different overall default rate ODR* as the value for WoE=0
  - Using the Bayes rule to transform the attribute default rates using ODR* as the target default rate
- In validation, as opposed to development, compare each attribute's actual value to either:
  - The average multivariate model prediction per attribute and time period
  - The constant univariate prediction as expressed through the attribute's scorecard points
- Generally, instead of doing calibration tests on default rates, one could also directly define confidence intervals around WoE values

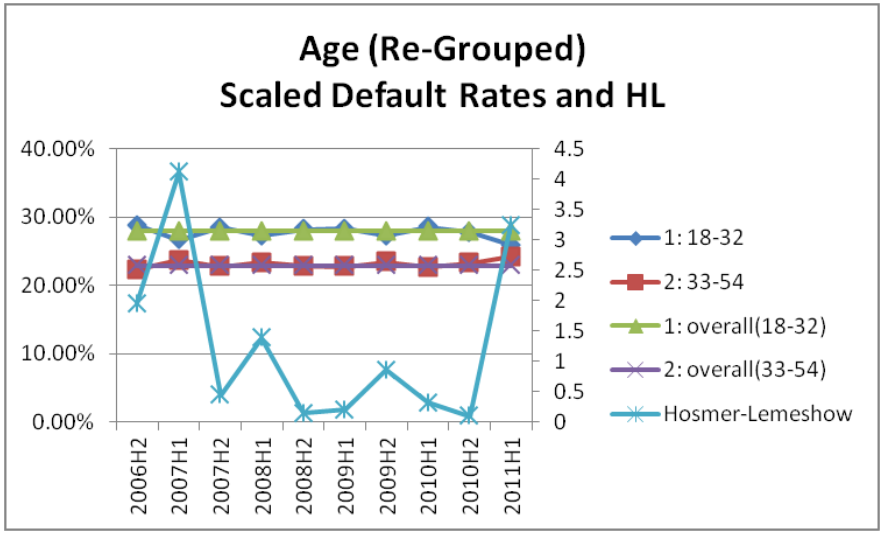
Case Study: Re-grouping for Stability - WoE





Case Study: Scaled Calibration Test





max(HL)	28.14657
p(min(HL))	0.00000
avg(HL)	6.75232
p(avg(HL)	0.03418

max(HL)	4.11729
p(min(HL))	0.04245
avg(HL)	1.27651
p(avg(HL))	0.25855

Part2: Other Topics

Variable-length performance window

- Evaluate each application's performance from its application date until today
- More recent applications hence have a shorter performance window and more recent cohorts have a lower default rate
- Pick a sample window end date sufficiently long ago to provide at least a minimum time for the default rates to mature, typically 1 year
- Quasi standard method in Germany, because it was used by SCHUFA when they introduced scoring models

Variable-length performance window - Benefits

- Allows inclusion of rather immature applications in the development sample, providing more recent information
- A fixed length performance window in combination with a product where default rates take a long time to mature would otherwise result in a sample window that ends a long time ago, neglecting possibly vital recent information

Variable-length performance window - Drawbacks

- The overall default rate cannot be interpreted. The scores need to be calibrated to a separately measured default rate, typically a long term central tendency
- Attribute distributions need to be stable over time, otherwise attributes that are more frequent in older cohorts would appear overly risky in the overall sample
- The Gini is expected to be slightly higher for more recent cohorts, because shorter horizons are easier to predict
- Calibration tests need to be based on weights of evidence or scaled default rates

Integrating Score -PD mapping and Virtual CT (1/2)

- Using the Bayes rule to transform the measured default rates of rating classes or score quantiles, so that the weighted average of the transformed values equals a given target value, for example a long term central tendency (CT), is common practice.
- It has been noted that using the Bayes rule on classes of cases with fixed class frequencies will result in a weighted average PD that is generally slightly off the target CT. This can be fixed by using instead a “virtual CT” as target value, the value of which is found iteratively, for example using Excel’s “goal seek” function or other tools.
- One can extend this idea by integrating the use of a virtual CT into the score-PD calibration process

Integrating Score -PD mapping and Virtual CT (2/2)

- We designed the following calibration process
  - Create a suitable number of score quantiles
  - Measure default rates per quantile
  - Bayes-scale default rates using a virtual CT
  - Regress the scaled default rates against score midpoints
  - Define a score-PD function using the intercept and slope of the regression line and apply to each individual score value
  - Map the resulting PDs to the Master Scale
  - Calculate the weighted average Master Scale PD
- We then determined a virtual CT iteratively, so that the weighted average Master Scale PD matches the actual Central Tendency. Note that this involves iterating the regression estimation.
- One may or may not include the Master Scale mapping into the iteration

Summary

- Always look at Weight of Evidence values over time
- Don’t rely on PSI as your only stability assurance
- Start out with a long development sample
- Re-grouping attributes may heal WoE instability

- Also:
  - Consider using variable-length performance windows for very slowly maturing products
  - In order to be spot on when calibrating, make use of a virtual CT