



Data Expansion In Credit Risk Modeling

-How should to handle seasonality data at credit risk modeling

Mark An

Credit Risk Analytics, Risk Management

CIBC

May 2009

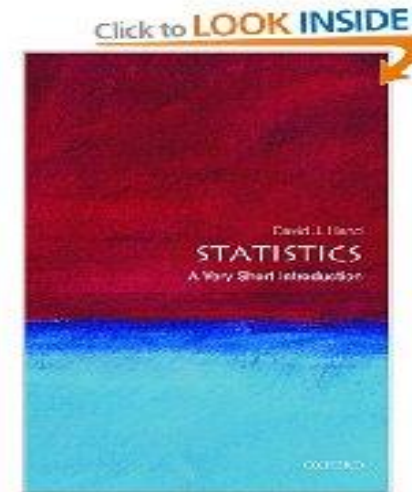
Statistics

Statistics is the discipline for predicting the future or for making inference about the unknown, or for producing convenient summaries of data.

-David Hand J. (2008) *Statistics: a very short introduction*. P3

It is easy to lie with statistics, but easier to lie without them.

-Frederick Mosteller (The chairman of [Harvard](#)'s statistics department, from 1957 to 1971)



Common issue: seasonality

Seasonality of consumer banking credit risk data: holiday seasons- December have relatively higher balance, January and February have higher delinquency rates, and loss rates

Question: How should handle seasonality data at credit risk modeling?



Introduction: Credit risk modeling processing

- Objective
- Selecting the time period (**observation period** and performance period)
- Target definition (target, non-target and indeterminate)
- Sample selection (if needed)
- Data partition
- Variable selection
- Transformation
- Missing value imputation
- Variable cluster analysis
- Variable correlation analysis
- Decision tree, Neural network, Logistic regression (as selected)



Introduction: Deciding the observation period

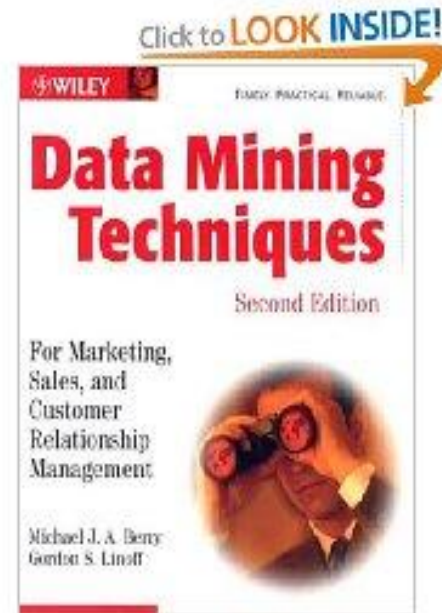
- Based on the modeling processing requirement
Example: the minimum data requirement for a scorecard is roughly 2,000 goods and 2,000 bads. In most cases, we could accumulate sufficient data using 6 months.
- Due to seasonalities of consumer banking credit risk data, an observation period of 1-year (12 months) data is recommended- otherwise sample bias will occur.



Sample Bias

'Building a model on data from a single time period increases the risk of learning things that are not generally true. One amusing example is an association rule model result all of predicted eggs-this surprise result because less so when we realized that model dataset was from the week before Easter.'

-(Michael J. A. Berry & Gordon S. Linoff Data mining techniques, P70)



Heart of every statistical model:

The sample must be representative of the population you wish to describe.

-Derek Montrichard, Reject Inference Methodologies in Credit Risk Modeling



Definition(1)

Abnormalities in banking data:

1. The first type of abnormality is seasonal data, Risk (delinquency, bankruptcy, and DWO-directly write off) of account will increase after holiday season in January and February. It will repeat or happen every year-We **focus** on this type of data.
2. The Second type is 'one time events'. These events only happen one time and never repeat again in the future. For example: certain marketing campaign.

'The technique to 'normalize' data is to filter out the source of abnormality.

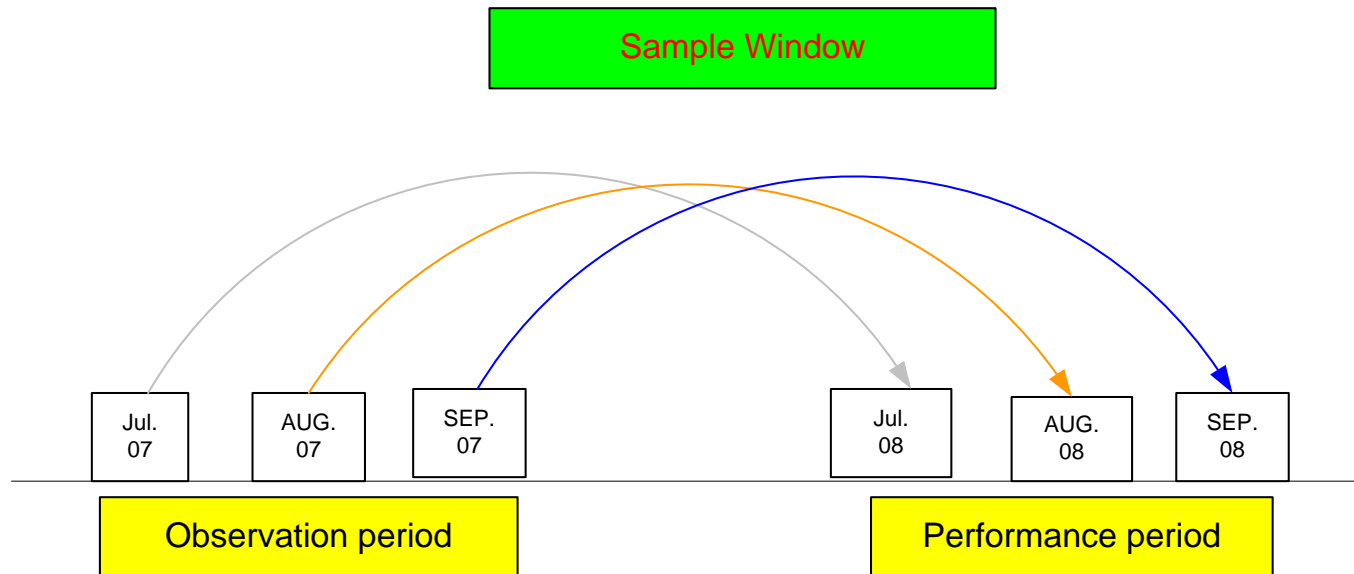
-(Naeem Siddiqi. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring, P37)'



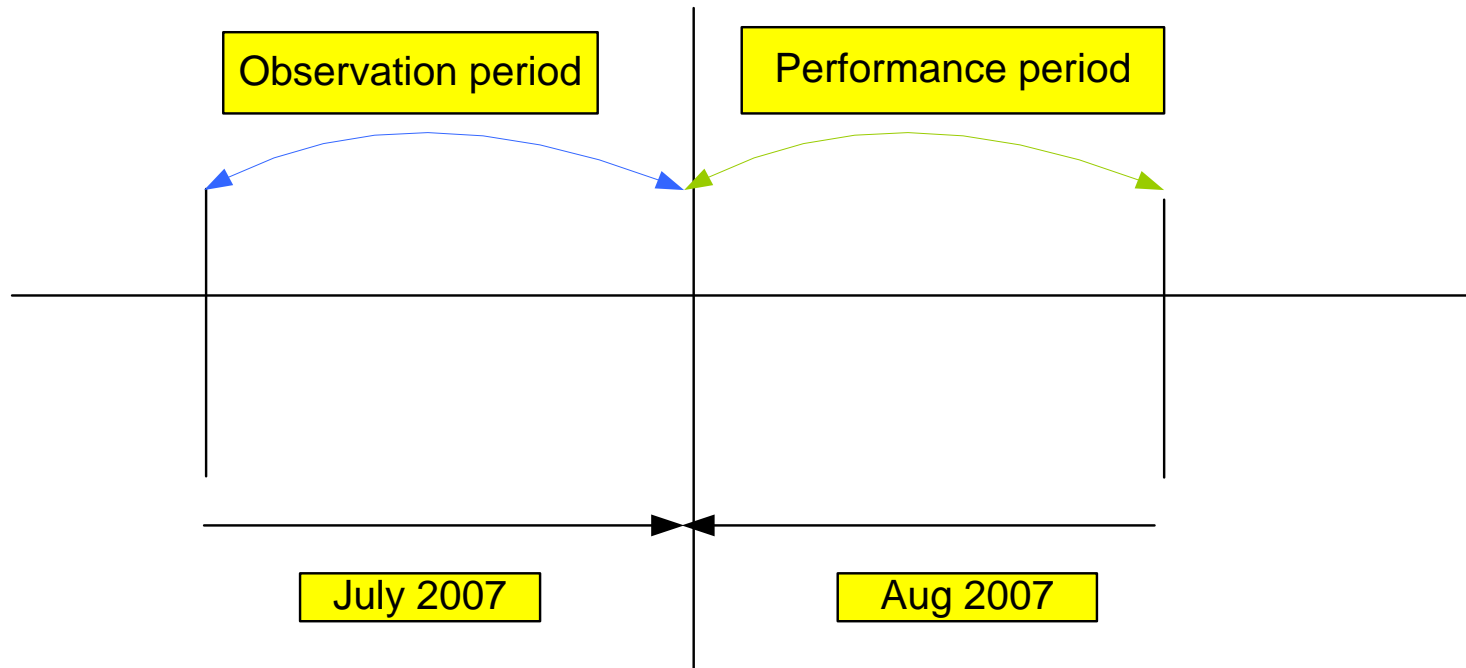
Definition(2)

Sample Window:

- **Multiple** observation periods (multiple timeframes)
- **Equal** performance periods (performance Window)



Case study



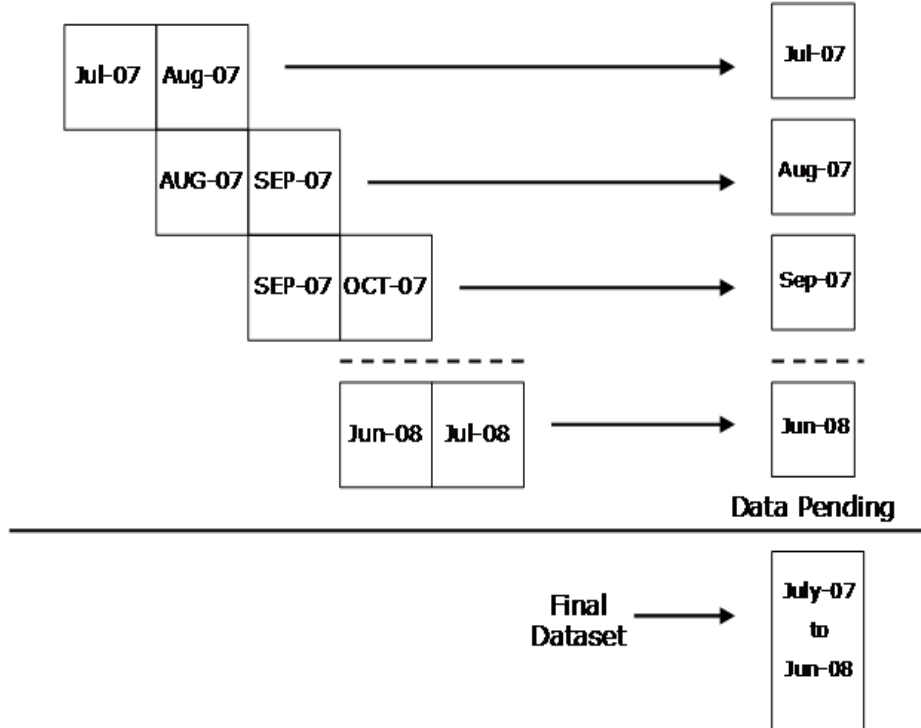
The performance period and observation period are the same at 1 month

Data Expansion



Stacking Sliding Windows – Data Expansion

Stacking Sliding Windows – Data Expansion



The objective is the maximum of cure rate on unsecured personal loan delinquency. Data expansion and data pending to get final of 12 months of dataset.



Case study

Datasets:

1. Data A-observation period over 12 months
2. Data B-observation period over 4 months (July to Oct.)

Variables:

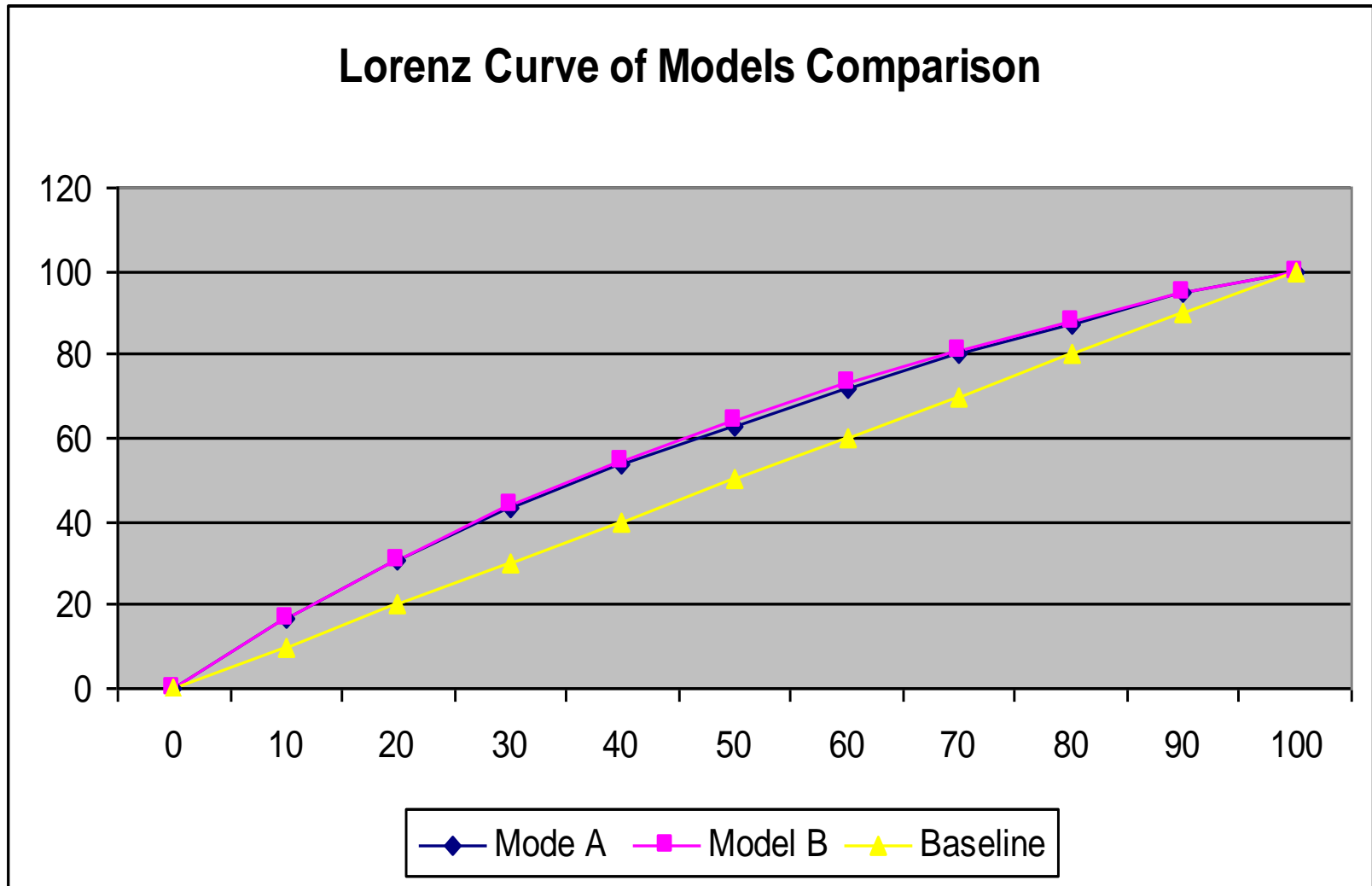
8 variables (Behavior score, Bureau Score, Balance at risk, last delinquency, Amt pass due, Months on book, Automatic debt indicator, cycle 1 delinquency times at 12 months)

Use SAS Enterprise Miner 5.3 -Logistic regression default criterion.

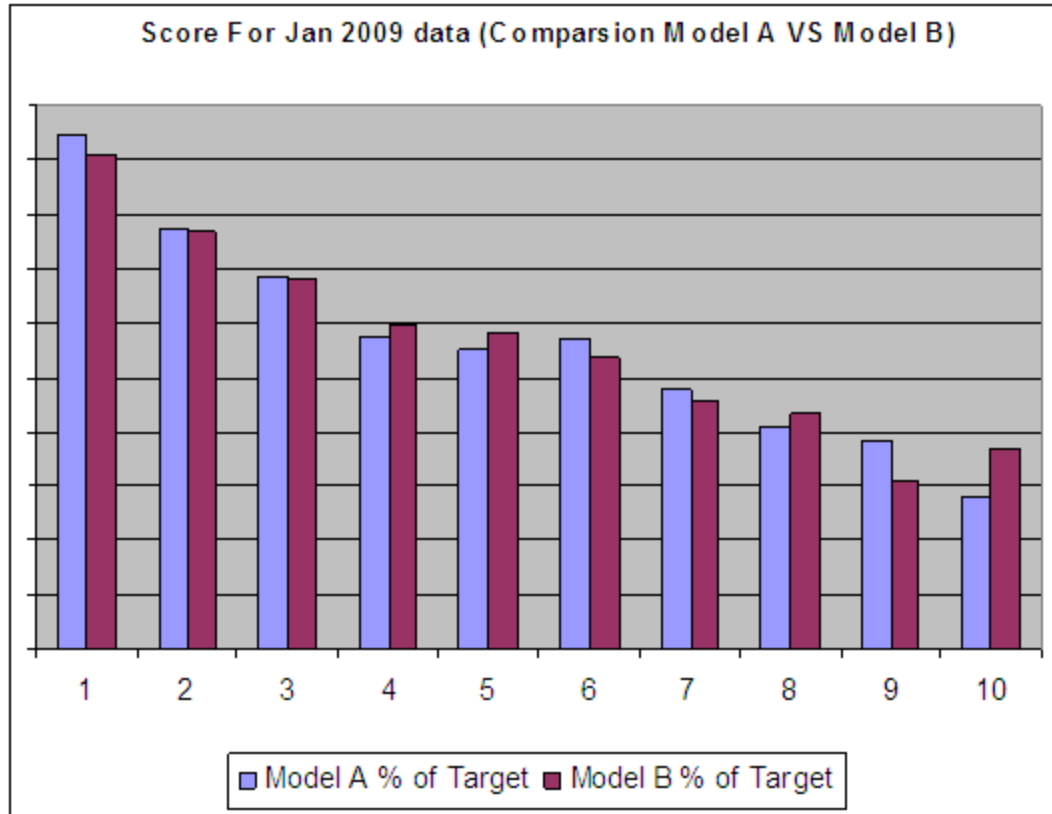
1. Use data set A-Model A
2. Use data set B-Model B



Chart: Lorenz Curve comparison-same result



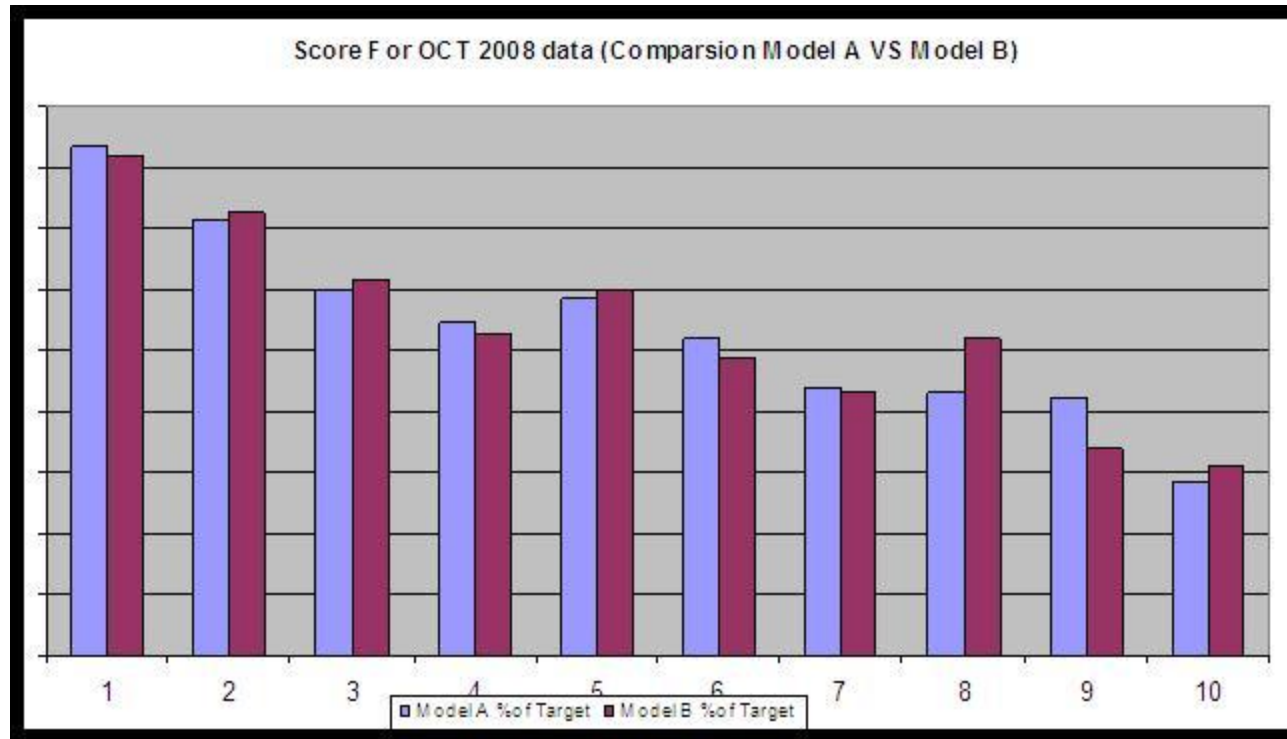
Model A and Model B score Jan 2009 data



- Model A is better for Jan. 2009 data prediction



Model A and Model B score Oct 2008 data



- Use Model A and Model B not too much different



Result explanation:

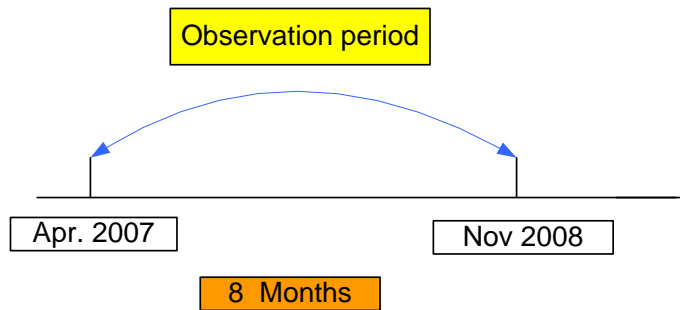
Sample over 12 months observation period, modeling will get more accurate prediction at seasonal month.

- *'The most important point to remember is this: "**your model is only as good as your data!**"*
 - *Olivia Parr Rud Data Mining Cookbook P48'*
- *'In general, when collecting data with the aim of answering or exploring certain question, the more data that are collected the more accurate an answer that can be obtained. This is consequence of the **Laws of large numbers**'*
 - *Hand D.J. (2008) Statistics: a very short introduction P47*

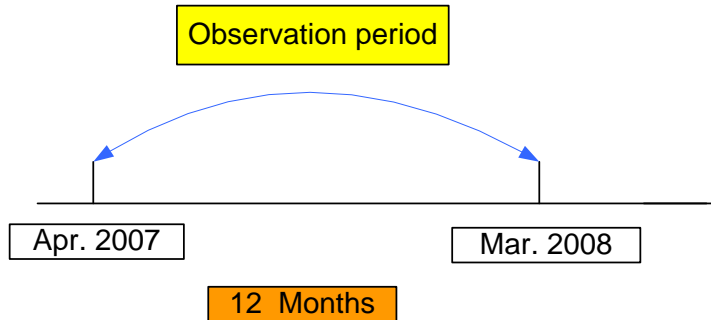


Suggestions to build a dataset

- Avoid building



- Suggestions



How to deal with large datasets? (1)

➤ Many books and working papers published before 2002, suggest using varied sampling methods to deal with large datasets. This was due to the fact that people did not have the computer power at that time.

It is unnecessary to use simple random sampling now when we have the computer power to process the original entire data.

*Derek Montrichard asked David Hand about this whole issue. David's response was, "**if you have the entire data available, why aren't you using it?**"*

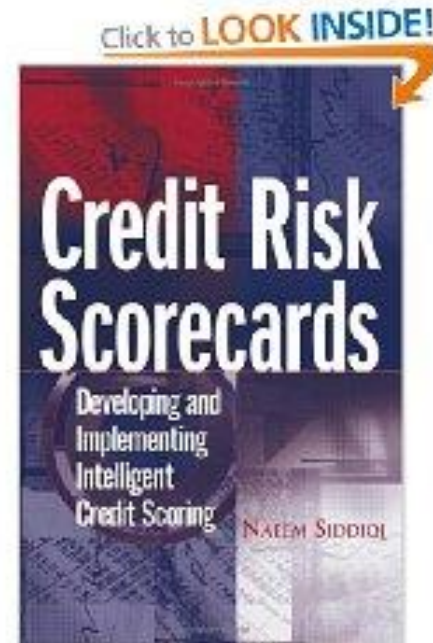


How to deal with large datasets? (2)

➤ If the target rate at **rare event**, for instance, fraud model where you have 2,000 bads and 40,000,000 goods, oversampling approach, will be recommendation, and adjustments for **over sampling** are later applied to get real forecasts.

'This method (oversampling) wildly used in the industry'

-(Naeem Siddiqi. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring, P63)



Conclusion

➤ Due to seasonality of credit risk data, we recommended using a 12 months observation period instead of a shorter observation period.

' The primary goal of multiple timeframes is creating stable modeling, that means model will work at any time of year and well into the future'

-(Michael J. A. Berry & Gordon S. Linoff Data mining techniques, P70)

➤ Another benefit of using a large sample is that it reduces the impact of multicollinearity and make the result of logistic regression statistically significant (C.H. Achen.1982.).



Questions?



References

- *Hand D.J. (2008) Statistics: a very short introduction. Oxford University Press*
- *C.H. Achen.1982. Interpreting and using Regression*
- *Edward M. Lewis An Introduction to Credit Scoring*
- *Christopher M. Bishop Pattern recognition and Machine learning*
- *Michael J. A. Berry & Gordon S. Linoff Data mining techniques*
- *Derek Montrichard, Reject Inference Methodologies in Credit Risk Modeling*
- *Olivia Parr Rud Data Mining Cookbook*
- *Naeem Siddiqi. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*
- *Jerome Friedman Trevor Hastie Robert Tibshirani The Elements of Statistical Learning*
- *Charles T. Clark & Lawrence L. Schkade Statistical analysis for Administrative Decision*

