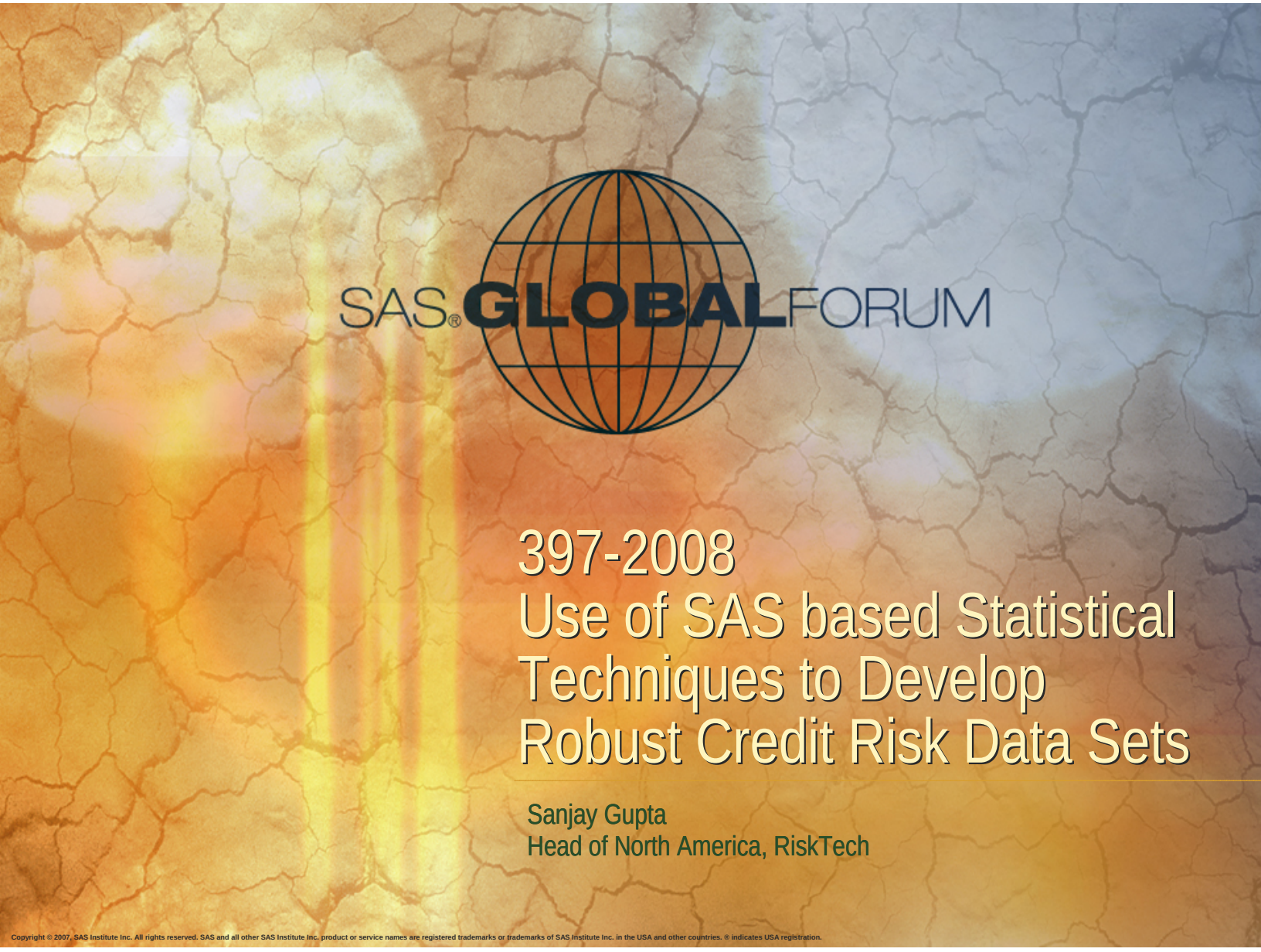


397-2008
Use of SAS based Statistical Techniques to Develop Robust Credit Risk Data Sets

Sanjay Gupta
Head of North America, RiskTech

Abstract

- We worked with a large financial institution in the identification of erroneous and problematic data fields. We developed a list of key problems with data fields based on business and credit risk rules.
- These problems were linked with different segmentation variables including industry, time period and customer attributes. Detailed analysis of these data issues with segmentation variables provided interesting insights for the business processes and where the key models will have limitations.
- Though this was not a primary objective but provided tremendous business value to the financial institution and led to number of initiatives to improve data quality. It also provided a robust data-set for model enhancement and subsequent development.

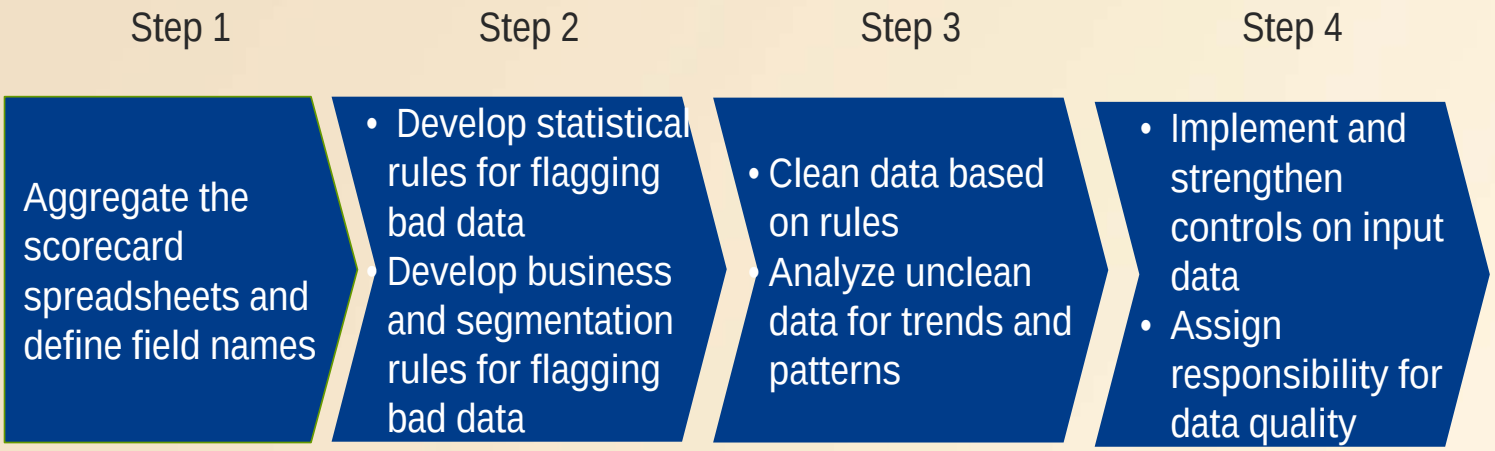
Business Problem

- A large bank had implemented Excel Worksheet Based solution for calculating and storing Credit Ratings of its middle market customers
- These Excel spreadsheets needed to be aggregated for data analysis and validation
- Without basic “data hygiene” these spreadsheets had numerous data integrity issues
- Additionally there were lots of “default” fields that were populated skewing results and contaminating validation results

Flagging Bad Data

- Bad data can be flagged as:
  - Missing fields
  - Duplicate entries
  - Erroneous fields
  - Use of wrong units (percentages vs. decimal values, millions versus thousands, etc)
  - Problem cases (0/0, 0/input, input/0, etc)
- Outliers can be reported based on the following:
  - Data ranges based on thresholds
  - Data ranges based on distribution
  - Business rules (i.e. Management quality is excellent while the counterparty's rating is below a defined threshold)

Key Project Steps



Examples of Business Rules – Financial Data

- Total Debt > Total Liabilities
- (Short Term Debt + Long Term Debt) > Total Debt
- (Net Worth + Total Debt) > (Shareholders Equity + Total Liabilities)
- (Total Liabilities + Net Worth) <> Total Assets
- EBITDA > EBIT + Depreciation + Amortization
- Gross Sales < Net Sales
- Total Assets = 0
- COGS < 0
- Sales > 0
- Current Liabilities > 0
- Total Liabilities > 0

Similar set of rules were developed for key segmentation variables like region, product type and lending centers

Use of SAS in Data Analysis

- We used basic SAS procs:
  - Proc Correlation
  - Proc Means
  - Proc Tabulate
  - Proc Format
  - Proc Gchart
- Reports were generated on a number of scorecards

Advantages to the Client

- Use of the existing infrastructure without additional costs
- Reports can be easily generated on a repetitive basis
- Analysis of the “bad data” identified issues in business processes by region and product type
- Additional statistical tests like Proc Logistic can be run on the clean data for validation of models
- Regulatory and SOX compliance

Contact Details

Sanjay Gupta

Head of North America

RiskTech

One Penn Plaza, Suite 3910

New York, NY 10119

212 912 3907

sanjayg@risk-technology.com


SAS GLOBAL FORUM
2008 SAN ANTONIO TEXAS
16-19 MARCH 2008