DATA QUALITY IN FINANCES AND ITS IMPACT ON CREDIT RISK MANAGEMENT AND CRM INTEGRATION

Berislav Nadinić
Risk Department, OTP Banka Hrvatska d.d, Ulica Domovinskog rata 3, 23 000 Zadar, Croatia

Damir Kalpić
Department of Applied Computing, University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3 10 000 Zagreb, Croatia

Keywords: Basel II, CRM, Data Quality, data mining, standardization, validation, framework.

Abstract: Basel II Capital Accord and increased competition on the financial market are responsible for creation of repositories of aggregated, customer centric historical data used for internal credit risk model development and CRM initiatives in banks. This paper discusses the effect of data quality on development of internal rating models and customer churn models, as well as potential improvements in this regard. A comprehensive framework for data quality improvement and monitoring in financial institutions is proposed, taking into account the Basel II requirements for data quality as well as requirements of customer centric retention campaigns.

1 INTRODUCTION

In recent years, the Basel II Capital Accord has become the focal point of interest in the banking world. A successor to Basel I Capital Accord, it specifies a framework for dealing with major issues facing banks in relation to credit, operational and market risks.

This new framework aligns capital adequacy assessment more closely with the key elements of banking risks in order to provide incentives for banks to enhance their risk measurement and management capabilities. The legislature in countries throughout the world specifies timetables when banks have to comply with the guidelines of the Accord.

The new Accord consists of three mutually reinforcing pillars, which should contribute to safety and soundness in the financial system (Tapiero 2004):

- First pillar - sets out minimum capital requirements, with the added option of allowing banks to use internal estimates of borrower creditworthiness to assess credit risk in their respective portfolios. Banks can estimate the probability of default associated with each borrower based on social, demographic and financial information
- Second pillar - supervisory review process is the key aspect of the Basel II Accords which ensures that the management and control checks, such as internal control review, assessment of risks and capital assessment are thoroughly documented and communicated to the senior management.
- Third pillar – is market discipline which affects the increased disclosure by banks about the market trading practices.

Based on these abstract concepts, Basel II Accord specifies that in order to ensure that successful credit risk calculations from data contained in different source systems can be performed, certain requirements need to be met. These three main requirements for data include authenticity, accuracy and transparency (Tapiero 2004) ensuring appropriate levels of data quality for historical data about customers from which their risk measures are being calculated.

Also, due to increased market saturation, banks are increasingly turning to targeted marketing campaigns in order to retain profitable customers

Nadinić B. and Kalpić D. (2008).
DATA QUALITY IN FINANCES AND ITS IMPACT ON CREDIT RISK MANAGEMENT AND CRM INTEGRATION.
In *Proceedings of the Third International Conference on Software and Data Technologies*, pages 327-331
DOI: 10.5220/0001879103270331

and increase the revenue on existing clients by offering new products and services. The targeted campaigns usually fall in the domain of analytical Customer Relationship Management (CRM) systems.

However, similar to the data quality requirements specified by the Basel II Capital Accord, CRM initiatives require that large quantities of analytical data, containing financial and demographic information aggregated on the customer level are of sufficient data quality.

Therefore, these two major business drivers in the banking community can be severely affected by poor data quality. In this paper we will discuss the impact of most common problems encountered, and suggest a complete framework for improving data quality.

2 IMPACT OF DATA QUALITY

When dealing with data quality issues while preparing historical data for creation of internal rating models in accordance with Basel II guidelines, as well as targeted marketing campaigns and retention models for CRM initiatives, one most frequently encounters the following problems (Nadinic 2006):

- multiple names for the same entity
- missing values
- incorrect values
- duplicate records for the same customer

These problems affect the model creation and reporting processes in several ways. Multiple names for the same entity (usually a name of a customer, organization or a product type) prevent the aggregation of data on a customer level, while giving a false idea about the actual number of customer or organization/product types.

Predictive model development is affected by missing values and multiple names in the following way (Nadinic 2006):

- If missing values in relevant fields are treated as a separate category, they may affect the accuracy by grouping the characteristics of high-risk and low-risk clients in the same category. This is also valid for multiple names where specific incorrect values can be treated as a separate category in the modelling process
- Multiple names for education levels, organization and employment information about the customer can lead to overtraining of the model on the training set if the number of defaults (and therefore the training set) is low. In this way, when the created model is used for actual scoring of the customers, the model will not identify the correct probability of default (and therefore rating class) to the customer

Creation of marketing campaigns is mostly affected by incorrect address fields of the customer which prevents contracting the customer through different sale channels, and incorrect and missing information about demographic and financial data which prevent segmentation and creation of customer churn models.

Duplicate records about the same customer may result in repeated offerings/contacts to the customer and can prevent identification of the profitable ones.

These data quality issues can reduce the effectiveness of created internal rating models, while in the case of CRM initiatives, they cause substantial financial losses.

3 PROPOSED SOLUTION

The proposed framework for data quality monitoring and improvement is divided into several distinct phases:

- Data quality analysis through operational and strategic data quality indicators
- Standardization
- Data cleaning according to business rules and constraints
- De-duplication and creation of “best records” for each customer
- Data quality assurance on data entry

It has to be noticed that all data quality activities should be performed on the organizational level by a group of dedicated employees, thus providing a framework for total data management.

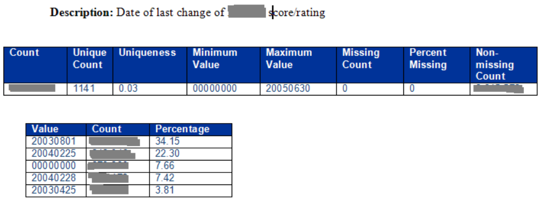
3.1 Data Quality Analysis

Data quality analysis (data profiling) is the analysis of fields and tables in search of interesting information (Ericson 2003). Therefore, we propose to formalize this approach by creating two types of information indicators:

- Operational data quality indicators
- Strategic data quality indicators

Operational indicators are used to determine data quality on an operation level – fields and tables from the transactional systems containing financial and demographic information, while strategic indicators are used to assess data quality on the customer level after the data have been aggregated.

Operational indicators, as shown in figures 1 and 2 can be comprised of information such as number of records/unique/missing records, outliers, statistical values for numerical variables and data patterns and redundant data analysis for character variables.



Description: Date of last change of [redacted] score/rating

Count	Unique Count	Uniqueness	Minimum Value	Maximum Value	Missing Count	Percent Missing	Non-missing Count
[redacted]	1141	0.03	00000000	20050630	0	0	[redacted]

Value	Count	Percentage
20030801	[redacted]	34.15
20040225	[redacted]	22.30
00000000	[redacted]	7.66
20040228	[redacted]	7.42
20030425	[redacted]	3.81

Figure 1: Example of data quality indicators.

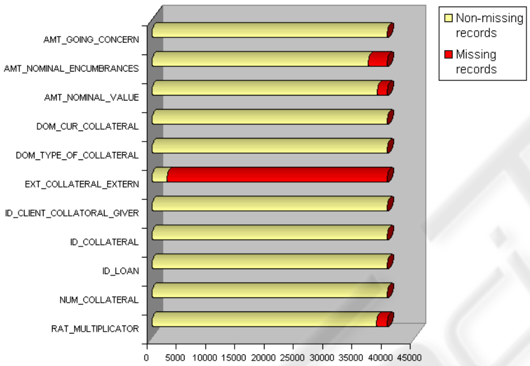


Figure 2: Example of data quality indicators.

3.2 Standardization

The standardization process encompasses all activities that result in a group of alphanumerical strings being transformed into a single string as shown in (1).

$$(A_1, A_2, \ldots A_n) \rightarrow B \qquad (1)$$

Therefore, we propose the use of standardization reference lists, coupled with the version of an algorithm for calculation of the distance between strings as a foundation for all standardization efforts.

In this way, the dedicated data quality experts may develop the standardization reference lists containing the value to be standardized, together with the mapped standardized value, as shown in Table 1 example.

Table 1: Reference standardization lists denoting three different names for the same entity and the standardized value in the right column.

Value for standardization	Standardized value
FER	Faculty of Electrical Engineering and Computing
Faculty of E. Engineering and C.	Faculty of Electrical Engineering and Computing
FACULTY OF ENG. AND COMPUTING	Faculty of Electrical Engineering and Computing

We propose a variation of the Monge-Elkan algorithm (Monge, Elkan 1995) where the distance between the strings is calculated not only for the entry string and the standardized value, but for the entry string and each of the values in the cluster for the same standardization values. In this way, special cases which substantially differ from the standardized value are included in the standardization process, as shown in the first row of Table 1.

With the variation of the Monge-Elkan algorithm for string comparison between strings A and B (2), we propose to calculate two measures (distance to standardized value and distance to value for standardization) to formalize the difference between the entry string and the values in the standardization table.

$$match(A,B) = \frac{1}{|A|}\sum_{i=1}^{|A|} \max_{j=1}^{|B|} match(A_i, B_j) \qquad (2)$$

A new algorithm is used for assigning these measures to each new entry string in the data cleaning process.

3.3 Data Cleaning According to Business Rules and Constraints

We suggest that in this part of the process, business rules and data constraints developed externally or through the process of data analysis through operational and strategic data quality indicators are applied on the data.

The developed standardization techniques from the previous phase are used to standardize the heterogeneous values in the attributes and ensure adequate data quality.

3.4 De-duplication and Creation of "Best Records" for the Customer

In this phase, we propose to use a system of "keys" (Hernandez, Solfo 1998), for each record aggregated on a customer level which captures the essential information about the customer that can be used in comparing the two customer records and identifying potential duplicates, as shown in Table 2. These keys are created by concatenating information from different fields that is least susceptible to data entry errors.

Table 2: Creation of keys for different records denoting the same customer.

Name	Address	Key
John Smith	Elk Road 23	JSmEl
J. Smith	Elk r. 23	JSmEl
J. Smythe	Elkk road xx	JSmEl

When dealing with the creation of "best record" containing the most correct information about the customer from multiple duplicate records, we propose to include the criteria based on non-missing information, most recent data and data that exist in standardization reference lists.

3.5 Data Quality Assurance on Data Entry

After steps have been taken to increase data quality on historical data used for modelling and campaign creation, we propose steps to ensure increased data quality on data entry points (notably information entered through forms in branches and Internet banking Web applications) to ensure sufficient quality in future historical data. These steps include:

- Use of categorical information in credit application forms for increased data quality for application scoring
- Selection of character string attributes from standardized reference lists and existing information in the database to reduce multiple names
- Use of obligatory fields in branch and Web applications to reduce the number of missing values
- Use of data constraints and business rules for exclusion of outliers

Implementing these steps may result in reduction of data quality issues on data entry points, while simultaneously reducing the necessary data quality tasks performed while loading the data into a data warehouse or a data repository used as a source for modelling.

4 FRAMEWORK

In this section we will discuss the implementation of the proposed solution into a complete Basel II framework. This implementation is an organization-level effort, concentrating on a team of data quality analysts that should ensure enterprise-wide standardization and data cleaning conventions. Data quality should be controlled on two major points:

- data entry into transactional systems
- data entry into data warehouse/data repository

Steps should be taken to ensure that the data used for both internal rating model creation and CRM activities should be subjected to the same data quality assurance processes as discussed in the previous chapter

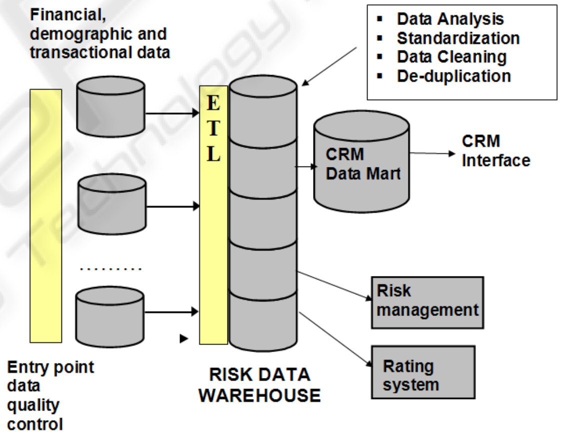


Figure 3: An integrated Basel II framework for ensuring data quality.

5 CONCLUSIONS

In this paper we have discussed the data quality issues concerning banks that are faced with two major business drivers – Basel II Accord compliancy and the development of CRM initiatives, notably marketing campaigns and the development of churn prediction models.

We have outlined major points in our approach that should increase quality of the historical data used for modelling, while at the same time giving an overview of methods used for controlling data entry.

It should be noted that this framework is currently being implemented in several Croatian banks, which are faced with the increased pressure from their respective owners and the local regulator to comply with the Basel II guidelines.

By using these methods for data quality improvement, they are trying to maximize the return on investment in their resources used to develop Basel II compliant framework. They are increasing the accuracy of created models and are dramatically reducing the costs of marketing and sales campaigns towards their clients

REFERENCES

Tapiero, C., *Risk and Financial Management: Mathematical and Computational methods*, Wiley & Sons 2004

Monge, C. Elkan; *The field Matching Problem: Algorithms and Applications*, 1995

Nadinic: *The Challenge of Data Quality: Business and Regulatory Implications and how to overcome it*, Proceedings of the SAS Forum Adriatic Region, October 2006

W. Ericson: *Data Profiling- Minimizing risk in Data Management projects*, DM Review 2003

M. Hernandez, S. Solfo: *Real World Data is Dirty: Data Cleansing and The Merge/Purge problem*, Journal of Data Mining Knowledge Discovery 1998