ENSURING DATA QUALITY FOR DATA MINING – A CASE STUDY

(Practice-Oriented Paper)

Varun Madhok
Global Business Intelligence Solutions
IBM Canada Ltd.
3600 Steeles Avenue East
Markham, ON L3S 9Z7 Canada
Internet: vmadhok@ca.ibm.com

ABSTRACT

This paper describes a data quality assurance exercise undertaken for a bank as part of a larger project in relationship marketing. The experience reported here is framed in a general context that is replicable across diverse data mining projects. This paper should be of specific interest to quality assurance practitioners for projects that harvest warehouse data for decision support to the management.

The assessment comprised ten checks in three broad categories, to ensure the quality of data collected over 807 attributes. The assessment discovered two critical gaps in the data that had to be corrected before the data could be transitioned to the analysis phase. The check and correct cycles lasted fifteen days.

Key words: Relationship marketing, data mining, data warehouse, quality assessment.

INTRODUCTION

This quality assurance exercise was part of a project undertaken for a bank to predict the propensities of its customers to buy its profitable offerings. At the core of the project was a charter that crystallized the objectives and limited the scope to specific categorizations of service offerings. Based on the charter, a detailed Data Requirements Document was generated - to list the data crucial to the analysis, to establish naming conventions and to identify the data's respective table(s) of origin in the warehouse. The sample data for analysis were extracted as per the Data Requirements Document as a flatfile. Since these data were used for generating predictive models, the validity and the accuracy of the analysis hinged on the integrity of the data with respect to the Data Requirements Document. According to the framework proposed by Wang et al [1], the work presented in this paper would be classified as a 'Production' data quality assessment.

The quality assurance was restricted to assessments of the data along the following two criteria.

1. Consistency: The extracted data had to be consistent with the minimal data requirements for the project - as stipulated by the Project Charter and as listed in the Data Requirements Document.
2. Accuracy: The extracted data had to be verified against the respective origins in the warehouse. The data in the warehouse were not assessed for accuracy[1].

Issues such as outlier-handling and missing-value processing were not in scope of this phase of the project.

THE DATA

The engagement objective was to score the bank's customers on their respective

[1] Aberrations in the data discovered in the course of the assessment were documented and submitted to the warehouse administrators. However, an evaluation of the warehouse data was beyond the immediate scope.

propensities to purchase each of twelve services offered by the bank. The solution proposed by the author required data on the customers' relationships with the bank categorized on the service and on the associated channels; the pertinent relationship attributes were frequency of transactions, recency of transactions, duration of relationships, and the monetary worth. Demographic data on the customers were also included.

Roll-ups and joins were the critical issues in the data preparation. The analysis had to be performed at the customer level while their associations with the bank were typically recorded at the account level. The bank offered close to a hundred products over the analysis period – these products had to be categorized into twelve service categories.

The Project Charter categorized the sixty-plus bank offerings into the services relevant to the project. The Data Requirements Document clarified the operations required to roll-up data from the account level to the customer level. The latter document listed 807 attributes on which data were required, and identified their source tables in the warehouse. The large volume of data was a result of the collation performed per month for twelve months, for each of the twelve product categories, for several variables. A unique customer number identified each datum.

The above details have been highlighted as being crucial to understanding the structure for the data quality assessment explained in the next section.

THE QUALITY CHECKS

Upon examination of the Data Requirements Document, and the associated extraction process, the focal points for the extraction process were identified as the following.

- Mappings: The data extraction required linking data from the business definitions, as identified by the Project Charter, to their encoding for the warehouse. Quality assurance required an examination of these mappings.
- Parallel extraction: The extraction process for certain data was identical across the twelve product categories. Data quality could be assessed through examination of such data for a single product category for a single month.
- Peculiar extraction: Certain data were peculiar to specific product categories. These data had to be examined individually for assurance of quality.

Quality assessment comprised comparison of the extracted data with the parent data in the warehouse, and checks on the code used in the extraction.

The checks on the 'mappings' were performed on the items tabulated below.

'Mappings' assessed for quality
Product identifiers - It was checked that the roll-ups from the granular product levels to the product categories were accurate.
Transaction identifiers - It was checked that the transactions used to measure the relationships among the bank and its customers were restricted to customer-initiated transactions.
Time identifiers - It was checked that the usage of the time identifiers to collate data from the fact tables was consistent with the encoding.

Once it was verified that the mappings, as identified above, had been accurately interpreted, the quality checks on the 'common extraction' items corresponded to verifying their extraction for a single month in any given product category. The quality checks for the 'parallel extraction' items were performed on the following.

'Parallel extracted' data assessed for quality
Monthly balances per product category - It was checked that the monthly balance for a certain customer in a certain product category was the sum of the balances for all the customer's accounts in that product category for the same month.
Monthly count of valid accounts per product category - It was checked that the number of accounts held by a given customer in a given product category for a given month was calculated correctly.

The quality checks for the 'peculiar extraction' items were performed on the following. The quality checks comprised the verification of the respective data as the cumulative over all the accounts held by the customer in the particular product category.

'Peculiarly extracted' data assessed for quality
Credit limits.
Days to maturity.
Overdraft limits.
Promotional pricing information.
Life/Disability insurance indicators.

QUALITY ISSUES

The quality checks on the data discovered the following critical gaps.

Critical gaps
In October 1999, the Bank phased out a credit offering. A new product was introduced at the same time and was assigned the code previously assigned to the expired offering.
Customer accounts for certain investment products were mapped to the customer identifiers using a linking scheme distinct from that of the other products. The quality issue was that the extraction process identified a 'new' set of investors with no associations to any of the other products offered by the bank.

The issues listed above were rectified after verification with the bank analysts and the warehouse administrators.

Other quality issues were unpopulated data fields and unary data. In each case, these gaps were communicated to the warehouse, but were considered non-critical and did not require immediate address.

SUFFICIENCY

The data mining analysis did not directly use all the variables listed in the Data Requirements Document. However, representative behavioral data on the customers' relationships with the bank were derived from these data, as were the appropriate indicator variables for use as the dependent variable in the analysis. The checks tabulated in the above lists were judged sufficient for ensuring success of the analysis.

DISCUSSION

The data-mining phase of the project was a prediction analysis. The models produced from this phase were each tested on a dataset that had been isolated from the design. The performance of the models was gauged based on their accuracy in identifying potential customers for the respective products/services. In the Project Charter, the client had established the minimal criteria for the prediction performance for each of the models. These measures were exceeded by each of the models in the validation. Project success indicated the robustness of the quality assurance phase.

The analyst who performed the quality assurance contributed to their design, but did not directly participate in the data extraction except to provide decision support to the data specialist. The Data Requirements Document served as a roadmap to the data

specialist, and was critical to the success of the project.

The main contribution of this work is the illustration of a structured method that condensed the task of verifying the project data to ten quality checks. The quality checks listed here can not be transferred to other prediction analyses without modification. However, their categorization as 'mappings', 'parallel extraction' and 'peculiar extraction' is a general, transferable framework. This proposition is elucidated in a general context below.

Typically, the granularity afforded by product and channel coding can not be translated to actionable business strategy. When extracting information from customer data, it is common practice to map product codes and service channels to coarse groupings. In projects where trend information is critical to the analysis, time-series data is needed and the associated extraction has a 'parallel' form. Experience suggests that the bulk of the data preparation is repetitive in nature and quality checks can be condensed to checks on 'mappings' and to spot-checks of representative attributes from the parallel-extracted data.

BIBLIOGRAPHY

[1] Wang, R., V. Storey, and C. Firth, "A framework for analysis of data quality research" *IEEE Tran. on Knowledge and Data Engg.* Vol. 7, No. 4, 1995.