International Journal of Engineering, Science and Mathematics
Vol. 8 Issue 12, December 2019,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell’s Directories of Publishing Opportunities, U.S.A

MAINTAINING FINANCIAL DATA QUALITY FOR BUSINESS INTELLIGENCE

NAVEEN KUNNATHUVALAPPIL HARIHARAN

Sr. Hyperion SME & Department of Information Technology, United States

ABSTRACT

Only when the input data is reliable can mathematical models and business intelligence systems for decision-making produce accurate and effective outputs. However, data taken from primary sources and gathered in a data mart may contain several anomalies that analysts must identify and correct. This research covers the activities involved in creating a high-quality dataset for business intelligence and data mining. Three techniques are addressed to achieve this goal: data validation, which detects and reduce anomalies and inconsistencies; data modification, which enhances the precision and robustness of learning algorithms; and data reduction, which produces a set of data with fewer characteristics and records but is just as insightful as the original dataset.

KEYWORDS:*Business Intelligence, Data modification, Data reduction, Data quality, Data validation*

1. Introduction

Making judgments based on poor data can have a significant influence on a company's strategy. The financial impact of data quality has been extensively researched and discussed. Poor data quality comes at a price, but good data quality comes with a reward.

These costs and benefits provide a sufficient return on investment for data quality initiatives to be justified (Işık, Jones, and Sidorova 2013; Watson and Wixom 2007). Data quality, which is integrated into projects by BI project managers, ensures the data warehouse's overall performance and the quality of data throughout the organization (Watson and Wixom 2007).

Although the success of a BI system is determined by the confidence of decision-makers in the data, data quality concerns the entire information architecture of the company, not just the front-end reporting tool or back-end data warehouse. The issue is that the BI system is frequently criticized for poor data quality since it raises the issue (Marshall and De la Harpe 2009). Data is viewed at the transaction level or within separate islands of information before it reaches the data warehouse. Because of the lack of a bigger view, the faults are obscured. The BI system's data aggregation magnifies data concerns, offering the company a bird's-eye view of the whole forest.

The economic impact of data quality has been studied extensively, and much has been written about it. Poor data quality comes at a price, whereas good data comes with a reward. These expenses and benefits give a sufficient return on investment to justify funding initiatives to improve data quality. Data quality should be incorporated into projects by BI project managers to assure not just the overall performance of the data warehouse and the quality of data throughout the company (Marshall and De la Harpe 2009).

Recognizing this chance to offer value, the BI project team must include data quality as part of their quality management plan throughout the planning phase. While the quality management plan is typically used to address the quality of project deliverables, the data quality plan should be considered a BI project deliverable in and of itself (Verbitskiy and Yeoh 2011). The project team uses this plan first, developing the organizational structures and processes that are required. It's subsequently handed over to the support organization as part of the go-live process. As a result, the company's data governance system is built on this foundation. Good data should be the most critical component because it doesn't matter how effective management sponsorship or business-driven incentive is if the organization does not have good data.

Any BI project will fail if the businesses lack the necessary data or the data is of insufficient quality (Eaton et al. 2008). Data profiling, which may characterize the data's content, consistency, and structure, is a good idea to undertake before implementation. It should be conducted as early as possible in the process. If the analysis reveals that the data is inadequate, the project should be placed on hold temporarily until the IT department figures out how to properly collect data (Jordan and Ellen 2009).

2. Data validation

Due to incompletion and noise, the quality of the input data may be inadequate.

2.1 Incomplete data

There are many reasons why certain records may have missing values corresponding to one or more characteristics. It's possible that certain data wasn't recorded in a systematic way at the source or that they weren't available when the transactions linked with a record occurred. In other cases, data may be missing due to recording equipment that is not working properly. It's also possible that some data was intentionally eliminated during earlier phases of the data collection process because it was deemed inaccurate. Failure to transmit data from operational databases to a data mart utilized for a specific business intelligence analysis can also result in incompleteness (Mircea, Ghilic-Micu, and Stoica 2011).

i. Methods for correcting Incomplete data

Several strategies are available for partially correcting incomplete data.

a)Deletion

All records for which one or more characteristics have no values can be discarded. If the value of the target attribute is absent, it is critical to discard a record in a supervised data mining study (Marsh 2005). When the proportion of missing values fluctuates irregularly across the different attributes, a policy based on systematic record deletion may be unproductive, as one risks losing a significant amount of data.

b)Inspection.

Alternatively, an inspection of each missing value by experts in the application domain might be requested to receive suggestions for possible substitute values. This method is vulnerable to much arbitrariness and subjectivity and is cumbersome and time-consuming when dealing with massive datasets (Taniar and David 2008). On the other hand, if used correctly, it has proven to be one of the most accurate corrective methods.

c) Identification.

A third option is to utilize a standard value to encode and detect missing values, eliminating the need to delete complete entries from the collection. For example, any missing data can be assigned the number (- 1) for a continuous attribute that only accepts positive values. Similarly, for a categorical attribute, missing values could be replaced with a new value that differs from the attribute's default values (McDonald et al. 2015).

d)Substitution.

There are several criteria for replacing missing data automatically, albeit most of them appear to be random. Missing values of an attribute, for example, might be substituted with the attribute's mean determined for the remaining data. This method can only be used with numerical characteristics, and it will plainly fail if the values are distributed asymmetrically (Sugumaran, Sangaiah, and Thangavelu 2017). In a supervised analysis, missing values can be replaced by calculating the attribute mean solely for records that belong to the same target class. Finally, missing data can be replaced with the highest probability value, which can be calculated using regression models or Bayesian approaches (Thakkar 2018). Estimation processes, on the other hand, can become rather difficult and time-consuming when dealing with a large dataset having a large percentage of missing data.

b. Data affected by noise and mitigation techniques

A random disruption within the values of a numerical attribute is referred to as noise, and it frequently results in noticeable anomalies. Outliers are values that are erroneous or abnormal in the data. Other sources of noise to look for include malfunctioning data measurement, recording, and transmission devices (McKnight 2005). The presence of data stated in different measurement units, which must be converted, might lead to mistakes and abnormalities. Outliers in a dataset must first be detected so that they may be corrected and regularized, or entire records containing them can be removed.

A different approach is to use clustering methods and calculate the distance between observations. The observations that are not placed in any of the clusters are identified as outliers after the clusters have been identified, representing sets of records with a mutual distance that is fewer than the distance from the records included in other groups. Clustering techniques have the advantage of considering multiple attributes at the same time, whereas dispersion-based methods can only consider each attribute separately (Duan et al. 2007).

The techniques described above can be combined with expert opinion to identify actual outliers in regular observations, even if they fall outside the intervals where regular records are expected to lie (Ranjan 2009). Before taking corrective measures in the case of anomalous observations, it is best to consult with experts, especially in marketing applications.

The goal of data validation techniques is to check and implement corrective actions in the case of incomplete and inconsistent data, as well as data that has been contaminated by noise.

3. Data transformation

It is appropriate to apply several transformations to the data set in most data mining analyses so as to improve the accuracy of the subsequently developed learning models. External correction techniques are indeed examples of the original data transformations which facilitate later phases of learning (Thakkar 2018).

Techniques for data transformation

a)Standardization

Most models are based on preventive data standardization, also known as standardization. The most common standards include ii) the decimal scaling method, ii) the min-max procedure, and the ii) z-index method (Evren Seker, Cankir, and Lutfi Arslan 2014).

b)Feature extraction

The goal of standardization approaches is to replace an attribute's values with values acquired through a suitable transformation. More complicated transformations are employed in some cases to generate new characteristics that reflect a set of additional columns in the matrix M that represents the dataset D. Feature extraction is the term used to describe such transformations (Salinca 2015). Consider the case when a set of attributes represents each customer's spending over a period of time. The data can then be used to create new variables that capture trends in the data using differences or ratios between expenditure amounts over time. In other cases, more complicated transformations, such as Fourier transforms, wavelets, and kernel functions may be used. Attribute extraction can also entail the construction of new variables that condense the important information contained in a subset of the original attributes into a single variable (Zareapoor and Seeja 2015).

4. Data reduction

The transformations outlined above are usually sufficient to prepare inputs for a data mining study when dealing with a small dataset. When confronted with a huge dataset, however, it is also appropriate to shrink it in order to make learning algorithms more effective without jeopardizing the quality of the findings obtained.

There are three basic criteria for deciding whether or not to utilize a data reduction technique: The models generated were created with efficiency, accuracy, and simplicity in mind (Chapman 2005).

Efficiency.

When learning methods are used on a dataset that is smaller than the original, the computation time is frequently reduced. If the algorithm's complexity is a super-linear function, as it is for most well-known algorithms, the efficiency gains from reducing the dataset size can be significant (Lee et al. 2006). It is common practice to run several different learning algorithms during the data mining process in order to find the best accurate model. As a result of the shorter processing times, the analyses can be completed more quickly.

Accuracy.

The accuracy of the models developed is a vital success factor in most applications, and it is thus the primary criterion used to choose one class of learning procedures over another. As a result, data reduction approaches should not have a substantial impact on the model's accuracy (Al-Najjar and Alsyouf 2003). It's also possible that some data reduction strategies based on attribute selection will result in models with greater generalization power on future records, as illustrated below.

Simplicity.

Some data mining tools, which are more concerned with interpretation than a prediction, require that the models developed be easily converted into simple rules that specialists in the application domain can understand. Decision-makers sometimes are ready to accept a minor drop in accuracy in exchange for simplified guidelines. When it comes to creating more simply interpretable models, data reduction is frequently used (Al Shalabi, Shaaban, and Kasasbeh 2006).

Because developing a data reducing technique that represents the best solution for all of the criteria given is difficult, the analyst will seek an appropriate trade-off among all of the requirements.

A reduction in the number of observations can be achieved through sampling, a reduction in the number of characteristics can be achieved through choice and estimation, and a reduction in the number of values can be achieved through discretization and aggregation.

a) Sampling

By extracting a statistically significant sample of observations from the original dataset, the amount of the original dataset can be reduced even further. Classical inferential reasoning is used in this form of reduction. As a result, the size of the sample must be determined in order to ensure the level of precision required by the subsequent learning algorithms, as well as an appropriate sampling technique. Depending on whether one wants to keep the percentages of the original dataset in the sample for a categorical attribute that is regarded critical, sampling might be simple or stratified.

Most learning models can be trained with a sample of a few thousand observations in most cases. Setting up numerous independent samples, each with a prespecified size, to which learning algorithms should be applied is also beneficial. In this method, computation durations rise linearly with the number of samples determined, and the multiple models developed may be compared to assess the robustness of each model and the quality of the information derived from data in the face of random fluctuations in the sample. When the models and rules created generally remain consistent when the sample set employed for training varies, the conclusions reached can be considered robust.

b) Feature selection

The goal of feature selection, also known as feature reduction, is to remove a subset of variables from the dataset that aren't considered relevant for the data mining activities. One of the crucial components of the learning process is determining which combination of predictive factors is most suited to adequately explain the event being studied. Feature reduction has a number of possible benefits. Learning algorithms can run faster on the smaller datasets than on the original one because there are fewer columns. Furthermore, models created after non-influential attributes are removed from the dataset are generally more accurate and understandable.

Filter methods, wrapper methods, and embedding methods are the three basic types of feature selection methods.

b.1. Filtering techniques

Filter methods choose the relevant properties before moving on to the next learning phase and are thus unaffected by the algorithm being utilized. The most important traits are chosen for learning, while the others are left out. Several alternative statistical metrics for evaluating the predictive potential and relevance of a set of features have been offered. In general, these are monotone metrics, meaning their value grows or decreases as the number of qualities analyzed increases or decreases. The most basic supervised learning filter includes evaluating each individual attribute depending on its amount of correlation with the target. As a result, the traits that appear to be most connected with the aim were chosen (Beniwal and Arora 2012).

b.2. Wrapper techniques

If the goal of the data mining research is classification or regression, and therefore performance is primarily measured in terms of accuracy, the predictive variables chosen should be based not only on the amount of significance of each individual feature but also on the learning technique used. Wrapper methods can meet this requirement since they evaluate a set of variables using the same classification or regression approach that was used to predict the value of the target variable (H. Liu et al. 2010). Each time, the algorithm learns using a different set of attributes, identified by a search engine that searches the whole set of all possible combinations of variables and chooses the set of attributes that guarantees the best accuracy. Wrapper approaches are typically computationally intensive since evaluating every conceivable combination found by the search engine necessitates dealing with the whole learning algorithm's training phase.

b.3. Embedded techniques

For embedded techniques, the attribute selection process is included in the learning algorithm, allowing for direct selection of the best set of attributes during the model creation phase. Embedded approaches include things like classification trees. They utilize an evaluation function at each node of the tree to calculate the predictive value of a single characteristic or a linear combination of factors. The required properties are automatically identified, and the rule for separating the records in the related node is determined as a result (Sharma and Saroha 2015).

When working with really large datasets with a significant number of attributes, filter methods are the best option. Wrapper approaches are ineffective in these situations due to the lengthy processing times. Furthermore, filter approaches are adaptable and, in theory, can be used with any learning algorithm. When the scale of the problem at hand is moderate, however, it is recommended to use wrapper or embedded methods, which provide higher accuracy levels than filter methods in most circumstances.

Wrapper approaches pick attributes using a search technique that inspects numerous subsets of attributes in order to apply the learning algorithm to each subset and assess the resulting correctness of the related model, as mentioned above. When a dataset has n properties, there are 2n potential subsets; therefore, even for low values of n, an exhaustive search technique would take too long to complete. As a result, selecting the characteristics for wrapper methods is normally a heuristic process, based in most cases on greedy logic that assesses for each attribute a sufficiently defined relevance indication and then selects the attributes based on their relevance level (Mikut and Reischl 2011)(Wang and Zhu 2018).

Forward, backward, and forward-backward search are three different search techniques that can be used.

Forward

As per the forward search scheme, also known as bottom-up search, the exploration begins with an empty set of attributes and then introduces the attributes one by one based on the relevance indicator's rating. When the relevance index of all the attributes that are still excluded falls below a predetermined level, the algorithm comes to a halt (Y. Liu and Schumann 2005).

Backward

The top-down search technique, also known as backward search, starts by picking all attributes and then eliminating them one by one depending on the desired relevance indicator. The algorithm comes to a halt when the relevance index of all remaining attributes in the model exceeds a predetermined threshold (Jović, Brkić, and Bogunović 2015).

Forward-backward.

The forward-backward technique is a trade-off between the aforementioned schemes in that it introduces the best attribute among those omitted at each phase while eliminating the worst trait among those included. The halting criterion is also determined by threshold values for the included and excluded properties in this situation (Kumar and Minz 2014).

Conclusion

The most fundamental expectation of business intelligence is that it will provide decision-makers with the information they need to make informed decisions. The implicit assumption in this statement is that the data used to produce this information is accurate. Unspoken assumptions, however, are usually translated into unaddressed requests. As a result, the BI project team must treat data quality with the same rigor and focus as other objectives like system uptime, system reaction time, and network performance. Data quality must be included right into the BI system's fabric.

Any BI project will fail if the businesses do not have enough data or the data is of poor quality. An analysis for quality of data should be done as early in the process as feasible, and if the analysis reveals that the data is inadequate, it is a good idea to put the project on hold until the IT department figures out how to properly collect and prepare quality data.

References

Al-Najjar, Basim, and Imad Alsyouf. 2003. "Selecting the Most Efficient Maintenance Approach Using Fuzzy Multiple Criteria Decision Making." *International Journal of Production Economics* 84 (1): 85–100.

Al Shalabi, L., Z. Shaaban, and B. Kasasbeh. 2006. "Data Mining: A Preprocessing Engine." *Journal of Computer Science*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.2072&rep=rep1&type=pdf.

Beniwal, Sunita, and Jitender Arora. 2012. "Classification and Feature Selection

Techniques in Data Mining." *International Journal of Engineering Research & Technology (ijert)* 1 (6): 1–6.

Chapman, Arthur D. 2005. *Principles of Data Quality*. GBIF.

Duan, Lian, Lida Xu, Feng Guo, Jun Lee, and Baopin Yan. 2007. "A Local-Density Based Spatial Clustering Algorithm with Noise." *Information Systems* 32 (7): 978–86.

Eaton, Scott, Michael Ostrander, Jennifer Santangelo, and Jyoti Kamal. 2008. "Managing Data Quality in an Existing Medical Data Warehouse Using Business Intelligence Technologies." *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, November, 1076.

Evren Seker, Sadi, Bilal Cankir, and Mehmet Lutfi Arslan. 2014. "Information and Communication Technology Reputation for XU030 Quote Companies." *arXiv E-Prints*, June, arXiv:1406.5073.

Işık, Öykü, Mary C. Jones, and Anna Sidorova. 2013. "Business Intelligence Success: The Roles of BI Capabilities and Decision Environments." *Information & Management* 50 (1): 13–23.

Jordan, John, and Clive Ellen. 2009. "Business Need, Data and Business Intelligence." *Journal of Digital Asset Management* 5 (1): 10–20.

Jović, A., K. Brkić, and N. Bogunović. 2015. "A Review of Feature Selection Methods with Applications." In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200–1205. ieeexplore.ieee.org.

Kumar, Vipin, and Sonajharia Minz. 2014. "Feature Selection: A Literature Review." *SmartCR* 4 (3): 211–29.

Lee, Yang W., Leo Pipino, James D. Funk, and Richard Y. Wang. 2006. *Journey to Data Quality*. MIT press Cambridge.

Liu, Huan, Hiroshi Motoda, Rudy Setiono, and Zheng Zhao. 2010. "Feature Selection: An Ever-Evolving Frontier in Data Mining." In *Proceedings of the Fourth International Workshop on Feature Selection in Data Mining*, edited by Huan Liu, Hiroshi Motoda, Rudy Setiono, and Zheng Zhao, 10:4–13. Proceedings of Machine Learning Research. Hyderabad, India: PMLR.

Liu, Y., and M. Schumann. 2005. "Data Mining Feature Selection for Credit Scoring Models." *The Journal of the Operational Research Society* 56 (9): 1099–1108.

Marshall, L., and R. De la Harpe. 2009. "Decision Making in the Context of Business Intelligence and Data Quality." *South African Journal of Information.*

https://journals.co.za/doi/abs/10.10520/EJC46314.

Marsh, Richard. 2005. “Drowning in Dirty Data? It’s Time to Sink or Swim: A Four-Stage Methodology for Total Data Quality Management.” *Journal of Database Marketing & Customer Strategy Management* 12 (2): 105–12.

McDonald, Kevin, Andreas Wilmsmeier, David C. Dixon, and W. H. Inmon. 2015. *Mastering the SAP Business Information Warehouse: Leveraging the Business Intelligence Capabilities of SAP NetWeaver*. John Wiley & Sons.

McKnight, W. 2005. “Text Data Mining in Business Intelligence.” *Information Management*. https://search.proquest.com/openview/301b972716e420fa70d5a72ebad15221/1?pq-origsite=gscholar&cbl=51938.

Mikut, Ralf, and Markus Reischl. 2011. “Data Mining Tools.” *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery* 1 (5): 431–43.

Mircea, M., B. Ghilic-Micu, and M. Stoica. 2011. “Combining Business Intelligence with Cloud Computing to Delivery Agility in Actual Economy.” *Journal of Economic Computation and*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.5151&rep=rep1&type=pdf.

Ranjan, Jayanthi. 2009. “Business Intelligence: Concepts, Components, Techniques and Benefits.” *Journal of Theoretical and Applied Information Technology* 9 (1): 60–70.

Salinca, Andreea. 2015. “Business Reviews Classification Using Sentiment Analysis.” In *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 247–50. ieeexplore.ieee.org.

Sharma, Nitika, and Kriti Saroha. 2015. “Study of Dimension Reduction Methodologies in Data Mining.” In *International Conference on Computing, Communication Automation*, 133–37. ieeexplore.ieee.org.

Sugumaran, Vijayan, Arun Kumar Sangaiah, and Arunkumar Thangavelu. 2017. *Computational Intelligence Applications in Business Intelligence and Big Data Analytics*. CRC Press.

Taniar, and David. 2008. *Data Mining and Knowledge Discovery Technologies*. Idea Group Inc (IGI).

Thakkar, Mohit. 2018. *Data Mining & Business Intelligence: QUESTIONS, ANSWERS, & EVERYTHING IN BETWEEN*