Anomaly Detection and Impute Missing Data in Supply Chain Management Using AI Techniques

Anandan Karuppana

PG Scholar, Department of Computer Science & Engineering, Sri Shanmugha College of Engineering and Technology, Sankari, Tamilnadu, India. Email: anandan.karuppana@gmail.com

Abstract: The performance of Supply Chain is becoming critically dependent on quality and consistency of data. The performance of Supply Chain is solely dependent on various master data objects like persistent definition of customers, items, products and locations. Inconsistent and inadequate management of data makes Supply Chain less competitive as surplus amount of money and time is spent in information management between systems and trading partners and little is available for innovation. Proper management of data helps to build an efficient Supply Chain, enabling resources to spend on innovation rather than muddling with problems. Supply chain Management requires Data Quality (DQ) and Master Data Management (MDM) to get rid of bad data and succeed in digitization and to ensure the data is harmonized, consistent, standardized across the lifecycle of the supply chain. The most common problems in Data Cleaning and Exploratory Analysis is to handling the missing values. Firstly, understand that there is no standard method to deal with missing data and Impute the missing entries of the incomplete data sets. Missing data leads to critical business problem because nearly all frequency distribution and statistical methods gather complete information of all the variables included in the analysis. Any absent observations on missing variables can dramatically reduce the sample size. As a result, the precision of confidence intervals gets affected, statistical power declines and the parameter estimates could be biased. Dealing with missing can be challenging as it requires a careful examination of the data to identify the type and pattern of data missing, empty, null values and clear understanding of different imputation methods are applied. In this research attempting to address SCM missing master data and Impute the missing entries of the incomplete data sets by applying the binary classification algorithm to find a structural solution.

I. Introduction

The performance of Supply Chain is becoming critically dependent on quality and consistency of data. Organizations need to pay more attention on Master Data Management (MDM). The Supply Chain managers may require changing their mindset on utilization of data, master data management can be defined as a technology-enabled business discipline that manages an organization's data to ensure consistency, accuracy, uniformity, and accountability of an enterprise's official, shared master data. The directive is to ensure that there are no multiple sources of potentially inaccurate data present in the organization. The performance of Supply Chain is solely dependent on various master data objects like persistent definition of customers, items, products, and locations. Inconsistent and inadequate management of data makes Supply Chain less competitive as surplus amount of money and time is spent in information management between systems and trading partners and little is available for innovation. Proper management of data helps to build an efficient Supply Chain, enabling resources to spend on innovation rather than muddling with problems. Master data has always been a part of Supply Chain, however the significance of its consistency in Supply Chain is growing. The main reasons behind it is increasing pressure on its performance, growing emphasis among organizations on identifying customer needs and collaboration of companies with trading partners to improve data business outcome. Advancement towards better management of master data and information governance programs still find great resistance. The reason behind such reluctance relates to the users who misunderstand the meaning of "information governance". Information governance mainly focuses in identifying data useful to the business and its desired business outcomes. Consider information governance only if they want to administer data with the sole purpose of realizing business value and if prepared to alter its business process to create, enhance or use data to extract that value. As a result, lack quality data and interoperability of data sets across different trading partners. Most supply chain master data available are difficult to correlate hence it is siloed in several systems, disparate tools and standards that causes of generating bad data. Fragmented and inconsistent data delays time-to-market, increases forecast errors, produces wrong production and inventory planning, affects cross-sell/upsell opportunities, and creates inefficiencies. Timely access to consistent, clean, and accurate data is essential not only for better communication between suppliers, partners, brokers, and shippers, and also for product commercialization and innovation. Though it is the IT environment where the MDM system publishes the master data SOR (System of Record) for use, it is not necessarily the system where master is created and maintained. The system accountable for maintaining the master data is called system of entry. Most organizations have multiple system of entry maintaining the master data. Master Data Management is beneficial to organizations in several ways. It helps in improving business agility by providing a single trust vision of people, processes and applications. It helps in making strategic decision making and enhances customer relationship. It assists organizations in reducing operation cost and boost compliance with regulatory requirements. The key issues in which MDM helps organizations are in handling of data redundancy and inconsistency, business inefficiency and support business change. Missing data in SCM is critical problem that leads to huge financial and effort loss and nearly all standard statistical methods gather complete information for all the variables included in the analysis. The most common problems in Data Cleaning and Exploratory Analysis is to handle the missing values. Firstly, there is no standard method to deal with missing and bad data. And the solutions for data imputation based on the kind of problem — Time series Analysis, ML, Regression etc. which is difficult to generalize the solution. In this research attempting to summarize the most used methods and trying to find a structural solution. Data Quality Management on Master Data Management are the comprehensive solution to define and manage all critical data which provides a single trusted view of data across the enterprise, data governance, agile self-service access, analytical graph-based exploration and data quality dashboard to continuously monitor the quality. Supply chain data management requires DQ and MDM to get rid of missing and bad data to ensure the data is standardized, consistent throughout the supply chain lifecycle and to succeed in digitation and decision making. This paper deals to address the missing data that are challenging for SCM, it requires a careful examination of the data quality to identify the type and pattern of data missing, empty, null values and apply the right imputation methods to populate the missing data to build a Single version of Truth and 360-degree view of the SCM Data.

II. RELATED WORK

Categorical data are important in many fields of research, examples are surveys with multiple-choice questions in the social sciences (Chen and Shao, 2000), single nucleotide polymorphisms (SNPs) in genetic association studies (Schwender, 2012) and tumor or cancer studies (Eisemann et al., 2011). It is most likely that some respondents/patients do not provide the complete information on the queries, which is the most common reason for missing values. Sometimes, also the information may not be recorded or included into the database. Whatever the reason, missing data occur in all areas of applied research. Since for many statistical analyses a complete data set is required, the imputation of missing values is a useful tool. For categorical data, although prone to contain missing values, imputation tools are scarce. It is well known that using the information from complete cases or available cases may lead to invalid statistical inference (Little and Rubin, 2014). A common approach is to use an appropriate imputation model, which accounts for the scale level of the measurements. When the data are categorical the log-linear model is an appropriate choice (Schafer, 1997). The simulation studies of Ezzati-Rice et al. (1995) and Schafer (1997) showed that it provides an attractive solution for missing categorical data problems. However, its use is restricted to cases with a small number of attributes (Erosheva et al., 2002) since model selection and fitting becomes very challenging for larger dimensions. A non-parametric method called hot-deck imputation has been proposed as an alternative (Rubin, 1987). This technique searches for the complete cases having the same values on the observed variables as the case with missing values. The imputed values are drawn from the empirical distribution defined by the former. The method is well suited even for data sets with a large number of attributes (Cranmer and Gill, 2013). A variant, called approximate Baysian bootstrap, works well in situations where the standard hot-deck fails to provide proper imputation (Rubin and Schenker, 1986). But the hot-deck imputation may yield biased results irrespective of the missing data mechanism (Schafer and Graham, 2002), and it may become less likely to find matches if the number of variables is large (Andridge and Little, 2010). Another popular non-parametric approach to impute missing values is the nearest neighbor’s method (Troyanskaya et al., 2001).

The relationship among attributes is taken into account when computing the degree of nearness or distance. The method may easily be implemented for high-dimensional data. However, the k-nearest neighbors (kNN) method, originally developed for continuous data, cannot be employed without modifications to non-metric data such as nominal or ordinal categorical data (Schwender, 2012). As the accuracy of the kNN method is mainly determined by the distance measure used to calculate the degree of nearness of the observations, one needs different distance formula when data are categorical. Some existing methods to impute attributes are based on the mode or weighted mode of nearest neighbors (Liao et al., 2014). Schwender (2012) suggested a weighted kNN method to impute categorical variables only, that uses the Cohen or Manhattan distance for finding the nearest neighbors. The imputed value is calculated by using weights that correspond to the inverse of the distance. One limitation of this approach is that it can handle only variables that have the same number of categories. Also, the value of k, which strongly affects the imputation estimates, is needed. There are some methods for imputing mixed data that can also be used for categorical data, for example, see Liao et al. (2014) and Stekhoven and Buhlmann (2012). The ¨ latter transform the categorical data to dichotomous data and use the classical k-nearest neighbors method to the standardized data with mean 0 and variance 1.

The imputed data are re-transformed to obtain the estimates. However, it has been confirmed by several studies that rounding may lead to serious bias, particularly in regression analysis (Allison (2005), Horton et al. (2003)). For categorical data one has to use specific distances or similarity measures, which are typically based on contingency tables. Commonly used distance measures include the simple matching coefficient, Cohen’s kappa κ c (Cohen, 1960), and the Manhattan or L1 distance. The Euclidean or variants of the Minkowski distance give an equal importance to all the variables in the data matrix when computing the distance. But for a larger number of variables, the equal weighting ignores the complex structure of correlation/association among these variables. As will be demonstrated, better distance measures are obtained by utilizing the association between variables. More specific, we propose a weighted distance that explicitly takes the association among covariates into account. Strongly associated covariates are given higher weights forcing them to contribute more strongly to the computation of the distances than weakly associated covariates.

III. GENERALIZED MODEL

A. supply chain management lifecycle

Supply chain management is the management of everything from R&D, production to product development, flow of goods, services and all processes that transforms raw materials into final marketable products. SCM Process involves the streamlining of a business's supply-side activities to maximize customer value and expand a competitive advantage in the marketplace. SCM suppliers to develop and implement supply chains model that are efficient and economical. information systems plays vital role in Supply chains cover everything from idea blue print, R&D, production to product development to capture data as an asset.

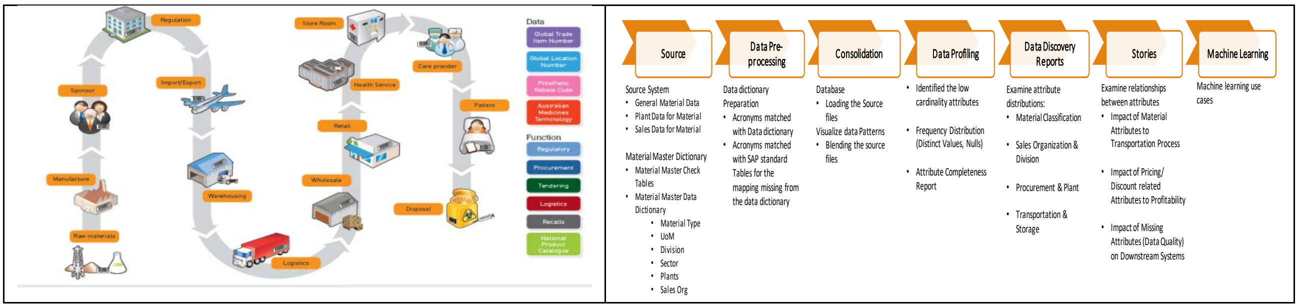
B. Supply Chain Data Management

Data management is an administrative process that includes acquire, harmonize, validate, store, secure, and processing required data to ensure the timeliness, accessible and reliable of the data for its users. Organizations are making use of Data more than ever before to make business decisions and gain deep insights of customer behavior, trends, and opportunities to create extraordinary customer experiences. Supply Chain Management produces various data over the lifecycle are as follows

- Transactional Data
- Master Data
- Reference Data
- Unstructured Data
- Golden Data

The Objective is to create Golden data which is a cleansed, de-duplicated, consolidated, validated version of the original master data. It is called as "Single Version of Truth" or "360° Customer View"

C. Solution Approach


Regulation
Store Room
Sponsor
Import/Export
Care provider
Health Service
Patient
Retail
Manufacture
Warehousing
Wholesale
Disposal
Raw materials
Logistics
Data
Global Trade Item Number
Global Location Number
Prosthetic Rebate Code
Australian Medicines Terminology
Function
Regulatory
Procurement
Tendering
Logistics
Recalls
National Product Catalogue
Source
Data Pre-processing
Consolidation
Data Profiling
Data Discovery Reports
Stories
Machine Learning
Source System
General Material Data
Plant Data for Material
Sales Data for Material
Material Master Dictionary
Material Master Check Tables
Material Master Data Dictionary
Material Type
UoM
Division
Sector
Plants
Sales Org
Data dictionary Preparation
Acronyms matched with Data dictionary
Acronyms matched with SAP standard Tables for the mapping missing from the data dictionary
Database
Loading the Source files
Visualize data Patterns
Blending the source files
Identified the low cardinality attributes
Frequency Distribution (Distinct Values, Nulls)
Attribute Completeness Report
Examine attribute distributions:
Material Classification
Sales Organization & Division
Procurement & Plant
Transportation & Storage
Examine relationships between attributes
Impact of Material Attributes to Transportation Process
Impact of Pricing/ Discount related Attributes to Profitability
Impact of Missing Attributes (Data Quality) on Downstream Systems
Machine learning use cases

- Source data are captured from the applications through real time or batch files. Material data are the master data considered for this presentation.
- Data preprocessing is to classify and segment the input data with reference to the industry standard data dictionary and maintain the consistent definition across the SCM lifecycle.
- Input data from various feeds and systems are consolidated to analyze the patterns, missing data, frequency analysis and visualize the analysis frequency patterns.
- Train the data through Machine learning algorithms such as Binary Classification to identify the Missing values and Anomalies, Predict the right Data and Impute
- Validate the correctness of the prediction and imputed values.

D. Algorithm Selection Process

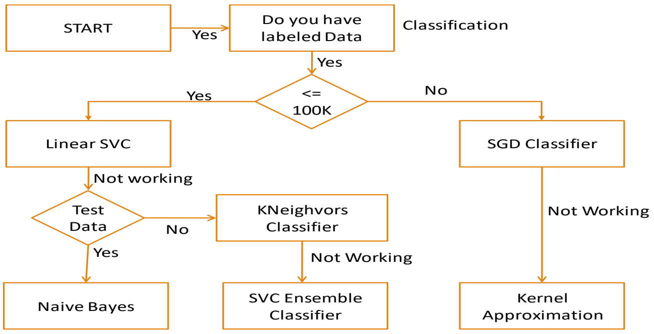


Fig.1 Algorith Selection Flow

E. Imputation

Understand the reason for the data goes missing.

- Missing at Random (MAR): Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
- Missing Completely at Random (MCAR): The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
- Missing not at Random (MNAR): Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

In the first two cases, it is safe to remove the data with missing values depending upon their occurrences, while in the third case removing observations with missing values can produce a bias in the model. So we have to be really careful before removing observations. Note that imputation does not necessarily give better results.

F. Multiple Imputation

- Imputation: Impute the missing entries of the incomplete data sets m times (m=3 in the figure). Note that imputed values are drawn from a distribution. Simulating random draws doesn't include uncertainty in model parameters. Better approach is to use Markov Chain Monte Carlo (MCMC) simulation.
- Analysis: Analyze each of the m completed data sets.
- Pooling: Integrate the m analysis results into a final result

G. Imputation of Categorical Variables

- Mode imputation is one method but it will definitely introduce bias
- Missing values can be treated as a separate category by itself. Create another category for the missing values and use them as a different level.
- Prediction models: Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable (training) and another one with missing values (test). We can use methods like logistic regression and ANOVA for prediction
- Multiple Imputation - KNN (K Nearest Neighbors)

In this method, k neighbors are chosen based on some distance measure and their average is used as an imputation estimate. The method requires the selection of the number of nearest neighbors, and a distance metric. KNN can predict both discrete attributes (the most frequent value among the k nearest neighbors) and continuous attributes (the mean among the k nearest neighbors). The distance metric varies according to the type of data:

1. Continuous Data: The commonly used distance metrics for continuous data are Euclidean, Manhattan and Cosine
2. Categorical Data: Hamming distance is generally used in this case. It takes all the categorical attributes and for each, count one if the value is not the same between two points. The Hamming distance is then equal to the number of attributes for which the value was different.

One of the most attractive features of the KNN algorithm is that it is simple to understand and easy to implement. The non-parametric nature of KNN gives it an edge in certain settings where the data may be highly "unusual". One of the obvious drawbacks of the KNN algorithm is that it becomes time consuming when analyzing large datasets because it searches for similar instances through the entire dataset. Furthermore, the accuracy of KNN can be severely degraded with high-dimensional data because there is little difference between the nearest and farthest neighbor.

IV Conclusion

Missing data is a critical business problem in SCM because nearly all standard statistical methods presume complete information for all the variables included in the analysis which impact the decision-making process. And the Objective to Impute data for the missing values through the above mention architecture is to reduce the business impact.

Comparison Matrix

Accuracy: (TP + TN) / Total Population where TP True Positive and TN = True Negative

- Accuracy is a ratio of predicted observation to the total observations. Accuracy is the most important performance measure.
- True Positive: The number of correct predictions that the occurrence is positive
- True Negative: The number of correct predictions that the occurrence is negative

F1-Score: (2 x P x R) / (P + R) where P=Precision and R=Recall

- F1-Score is a weighted average of Recall and Precision. Therefore, score takes of both false positives and false negatives into account. F1-Score is more useful than accuracy, when the sample has uneven class distribution.
- Precision: is about being precise, When a positive value is predicted or predict 100% correct.
- Recall: is about capturing all cases and not predict 100% correct.

Algorithm Accuracy		
Algorithm	***Accuracy***	***F1-Score***
Logistic Regression	84.60%	0.6337
Navie Bayes	80.11%	0.6005
Stochastic Gradient Descent	82.20%	0.5780
K-Nearest Neighbours	83.56%	0.5924
Decision Tree	84.23%	0.6308
Random Forest	84.33%	0.6275
Support Vector Machine	84.09%	0.6145

Table 1: Algorithm Comparison Matrix

REFERENCES

[1] J. Östlin, E. Sundin, and M. Björkman, "Importance of closed-loop supply chain relationships for product remanufacturing," International Journal of Production Economics, vol. 115, pp. 336-348, 2008.

[2] H. Min and H.-J. Ko, "The dynamic design of a reverse logistics network from the perspective of third-party logistics service providers," International Journal of Production Economics, vol. 113, pp. 176-192, 2008.

[3] Hastie, T., Tibshirani, R., and Friedman, J. Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd ed. Springer-Verlag, New York, 2009.

[4] Gelman, A. and Hill, J.Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge, 2007.

[5] Allison, P. “Multiple Imputation For Missing Data: A Cautionary Tale.” Sociological Methods and Research, 28, pp.301-309,2000.

[6] Altmayer, L. “Hot-Deck Imputation: A Simple DATA Step Approach.” Proceedings of the 2002 Northe, 2002.

[7] Faraway, J.J. Linear Models with R, Chapman & Hall, New York.. 2005.

[8] Horton, N.J., Kleinman, K.P. “Much Ado About Nothing: A Comparison Of Missing Datamethods And Software To Fit Incomplete Data Regression Models." American Statistician, 61, 1, pp.79-90, 2007.

[9] Little, R.J.A. and Rubin, D.B. Statistical Analysis with Missing Data, 2nd ed. John Wiley, New York, NY, 2002.

[10] Yarandi, H.N. “Handling Missing Data with Multiple Imputation Using PROC MI in SAS.” Proceedings of the Southeast SAS User Group, Savannah, GA, 2002.

[11] UCLA: Academic Technology Services, Statistical Consulting Group. “Multiple Imputation in SAS, Part1.” http://www.ats.ucla.edu/stat/sas/seminars/missing_data/part1.htm (Accessed April 24, 2011)

[12] Wang, K. and Jiang, W. “High-Dimensional Process Monitoring and Fault Isolation via Variable Selection.” Journal of Quality Technology, 41, 3, pp.247-258, 2009.