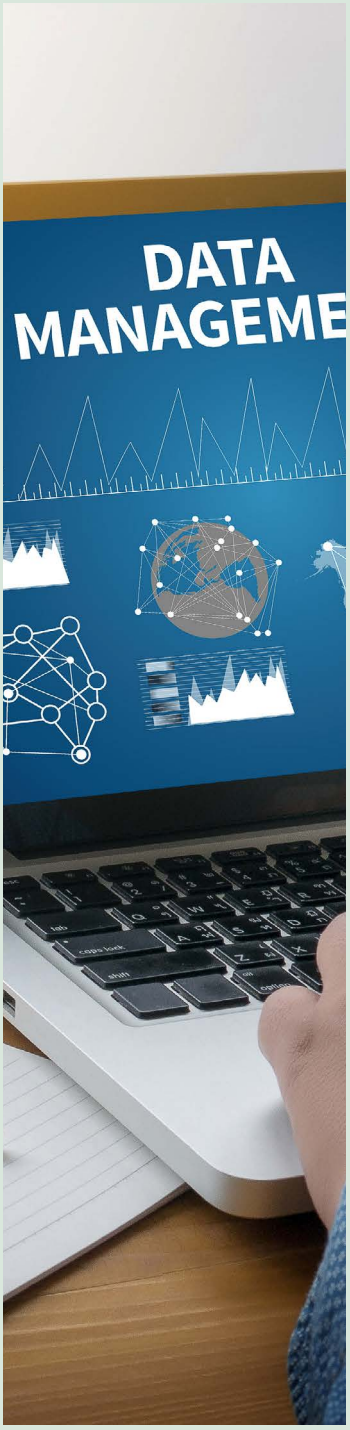
WHITE PAPER




# USING MACHINE LEARNING IN DATA QUALITY MANAGEMENT

Infosys®
Navigate your next

## Introduction

Recent improvements in computing power, decreasing costs of storage and the availability of suitable infrastructure has renewed the focus on artificial intelligence. We are on the cusp of an inflection point as far as artificial intelligence and its applications are concerned. A lot of focus has been on the ability of the new, emerging technologies to use the available data and create "use cases". But, for the technologies to fulfill their potential, the availability of rich and high quality data is essential.

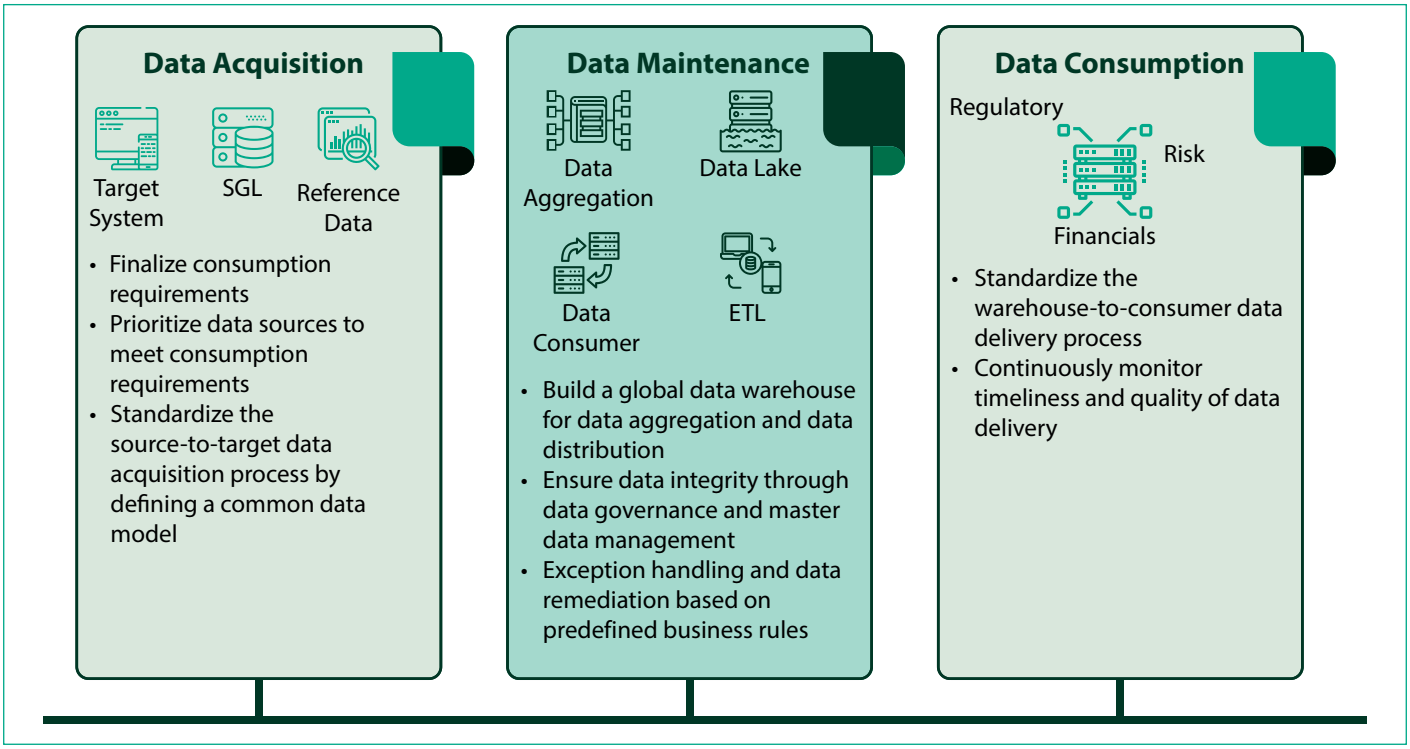
A recent Gartner research estimates that poor data quality is responsible for an average loss of $15million per year[1] in all organizations. As far as a financial institution is concerned, data quality does have a major impact on its daily activities. In big financial institutions, especially the sell-side firms, there has been an increasing focus on data quality to build mature data management capabilities[2]. With the changing nature of business, the advent of Fintech and stricter regulatory requirements, these existing approaches to data quality management may not be sufficient as they can be reactive in nature.

This point-of-view explains how machine learning can help in improving data quality management that can build operationally more efficient financial institutions.

## Data Quality – The cog in the wheel

### Data quality management – The current state in organizations

Typically, in any organization, the lifecycle of data starts at the source, evolves, before reaching the data consumers.




Over the past decade, organizations have taken additional steps to exercise a greater degree of control over data, in part due to regulatory requirements, and in part due to business needs. Organizations identify data requirements based on the needs of the consumers of treasury, risk, regulatory and tax data (The list is indicative and not exhaustive). The data management function in an organization is responsible for making good, clean data that is driven by data quality dimensions such as accuracy, consistency, completeness, timeliness, integrity, conformity, and veracity available to consumers. Given below is an indicative list of activities performed by an organization to manage data:

- Simple technological efforts such as data validation, data enrichment and data transformation have been put in place to prevent faulty data from reaching the data consumers.
- Data quality and business rules have been established that are mapped to the business policies.
- Time and effort have been spent on identifying the critical data elements (CDEs). The critical data elements are the building blocks of an organization. For example, the seemingly harmless

loan disposition date of a corporate loan or the final settlement date of an OTC derivative are CDEs that indirectly impact both the risk models as well as regulatory reporting.

- To ensure timely delivery of data, organizations have service level agreements (SLAs), both to receive incoming data and send outgoing data
- Finally, to monitor the quality of incoming and outgoing data, some organizations have built stand-alone data quality platforms that produce data quality scorecards, which help in the remediation process as they help the management to get a view at a data source, country and even at a contract level, to quantify the quality of data.

Organizations with a reasonably mature data management function use these metrics to create an impact taxonomy by tracing the data quality issues to the undesired financial impact that they may have on the business.

## Drivers of change

In organizations, there are opportunities to use resources more efficiently, to predict data quality issues and prescribe solutions, and to ultimately improve data management. For the purpose of this discussion, some of the key drivers of change in data quality managements have been provided below:

### Changing business models

Ironically, the world's most valuable banks which have existed for close to two centuries, and which have so much data, can be compared in market value with the tech companies. Data has been the key for the rapid growth of the tech companies, and banks do realize the importance of using data to increase their value. Increasingly, data is used for customer analytics, taking business decisions that can have an impact on the topline, and providing better user experience to address the changing customer behaviour. Data can help in identifying new segments of business by creating targeted products.

The increase in the volume of data available for analysis is fundamental to achieving microsegments which might lead to realizing the industry aspiration of the 'segment of one'. Unfortunately, the ability to manage data is not proportional to the increase in the volume of data as its management aims to address the physiological needs of data.

### Impact of data quality on operational risk capital

Banks consider one or all of the revenue from business units, the loss from operations, scenario analysis or other control factors like Key Risk Indicators (KRIs) to calculate the operational risk (OpRisk) capital[3]. The quality of data has an impact on the OpRisk models used by firms and their business units to calculate this capital. This is especially true if the modeler uses scenario analysis when they are building models. Scenario analysis is performed during focused workshops where experts offer their opinions on scenarios and their possible outcomes. In a bank, the KRIs play a major role in shaping the opinion of these experts. There is a causal relationship between the business impact and the KRIs, and an OpRisk modeler takes advantage of this relationship. The data quality issues introduce model risk and make the OpRisk environment intractable which have a cascading effect on OpRisk measurement.

### Operational effectiveness

Data quality issues can be due to data entry errors, inconsistent format of incoming data, duplicate or missing data, incorrect mapping between sources and acquiring systems, complicated data transformation among other causes. These causes are mainly due to the inorganic growth of systems and applications in financial institutions. Adding to this complexity, due to the changing operational and business models, new sources of data are becoming less in-house and data itself is becoming more unstructured. Organizations may lose control over the quality of incoming

data, which is especially a nagging issue in tackling financial crime. The existing data remediation strategies are manual, reactive and consume a lot of time and effort on non-value add activities due to increased throughput time. They can be inadequate too, as seen in a sample data set of a bank, in which only 9% of the incorrect records were flagged off as non-reportable records[4]. This retrospective method of identifying root causes is used as an anecdotal evidence(s) to reactively resolve specific, future issues. Needless to say there is a lot of girth in this methodology, which provides an opportunity to make it leaner. Within these boundaries and given the non-prescriptive nature of such efforts, the timely delivery of accurate data to the downstream systems, especially when the volume is peaking, can also become questionable.

### Shrinking revenues and increasing regulatory fines

A profitable trade can easily result in a loss for the trading unit that results in the bank paying fines to the counterparties due to risks such as operational risk and settlement risk. Data integrity is one of the main reasons for such issues. Lack of appropriate controls on data could lead to extraordinary losses. In 2005, a typing error during a sell-order resulted in a loss of $225 mn for a brokerage firm[5]. Globally, since 2008, banks have paid $321 bn in fines, poor data quality being one of the causes of any regulatory fine[6]. Increasingly, regulators are asking financial organizations to explain their data lineage process, failing which, levy heavy fines on the organizations[7].

### Increase in the non-financial risk to an organization

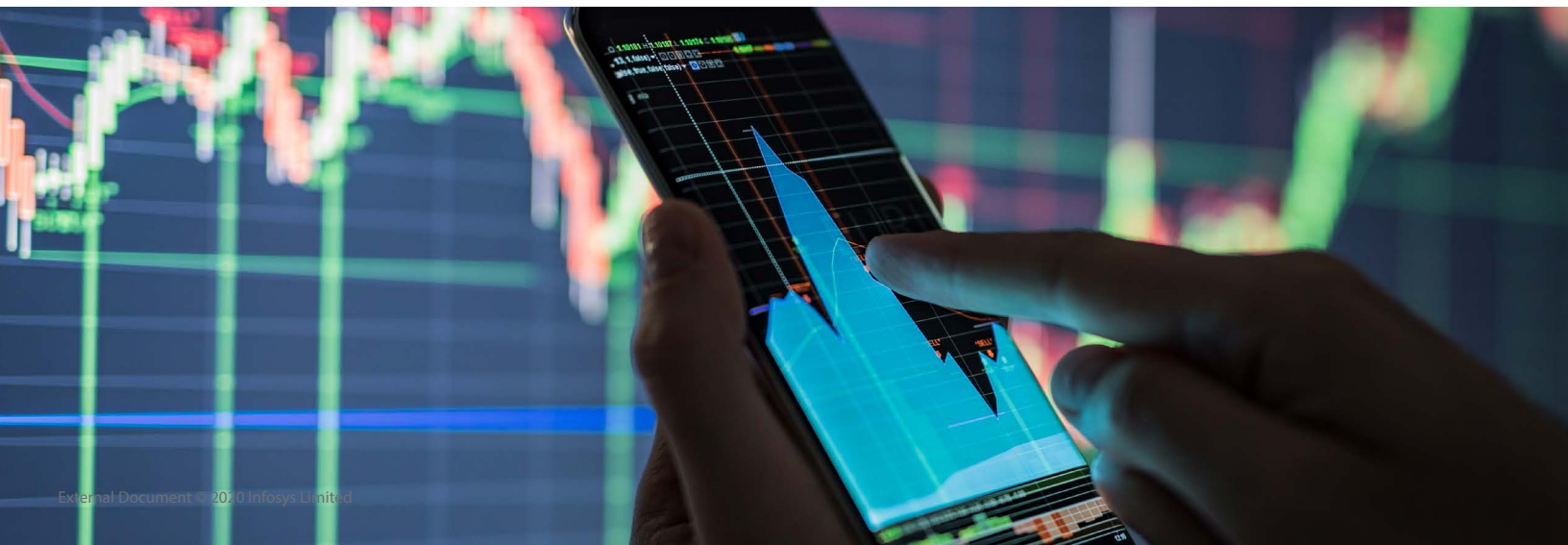
The impact of data quality issues is not restricted to losses suffered and fines paid by the organizations. The impact can be both internal and external. Within the organization, data quality issues can break the trust required from other units to use the data for forecasting and reporting. Additionally, regulatory reporting and reputation of the organization are affected due to poor data quality.

## Potential use cases for improving data quality management using machine learning

Artificial intelligence can be defined as a branch science that aspires to imitate humans. Machine learning is a subset of artificial intelligence which uses historical data and algorithms to build predictive models that can learn by itself, identify patterns and predict outcomes. The data management function is ideal for machine learning algorithms to detect anomalies and prescribe remedies that can improve error-detection.

To better manage data quality using the emerging technologies in artificial intelligence, a few potential use cases have been identified below.

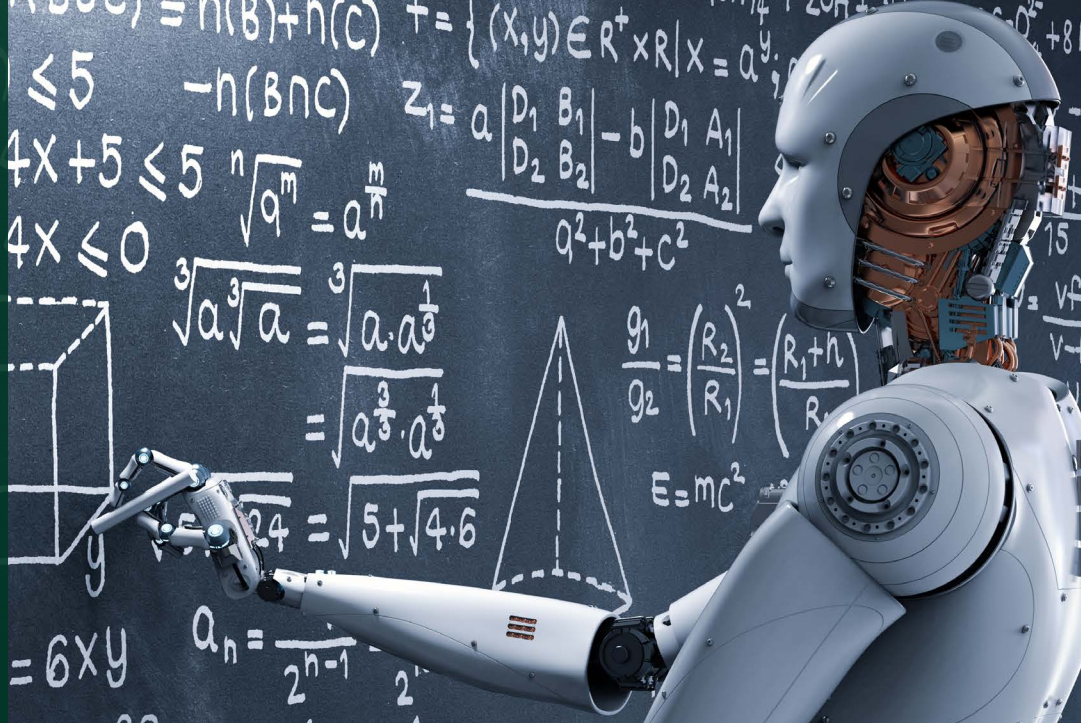
| Use case | Description |
|---|---|
| Automated data entry | In many organizations, considerable time is spent on manually entering the data to the different systems. With the increase in the amount of automated and semi-automated sources of data, this may not be sustainable in the future. To make the processes more efficient, technologies like OCR and speech-to-text can be used to scan documents of any type and convert recordings of human speech to text respectively. For example, unstructured, but relevant data, can be extracted from ISDA documents using machine learning algorithms, which can then be automatically updated to individual applications. The use of such technologies will help in transforming the business and reducing the time-to-market drastically. |
| Improve data quality at the source | Common errors during manual data entry include incorrect spellings, incomplete addresses and other fat-finger errors. Machine learning algorithms can help in resolving these common errors and also help in standardizing the data. For example, if the address of a particular customer is 123, ABC street and the data comes in as 132, ABC street, the algorithms can be used to rectify the mistake at the source. Supervised machine learning algorithms can continuously learn against the reference data to improve the accuracy of the predictions. If the reference data is not available, record linkage can be used to link the different data sources and identify the closest match to predict values. |
| Fill data gaps | The features of a record are usually correlated in one way or the other. For example, if product code is one of the attributes, they are usually related to the product identifier, country, and region among other features. When there is a relationship between the missing data and the other features of the record or when there is historical information on the probable data values, supervised machine learning algorithms can be used to fix the data issues by predicting the values to correct the missing values. When historical information is not available, expectation maximization algorithms can be used to find the maximum-likelihood estimates. Feedback can help the algorithms to learn and build accuracy over time. |
| Improve incorrect reporting | During regulatory reporting, incorrect records that shouldn't be reported may be inadvertently reported to the regulators. Machine learning algorithms can be used to remove these incongruous records from the report. The current method used to identify such records may be manual and may have the ability to identify only a small portion of these out of place records. Supervised learning algorithms such as Random Forest algorithms or Logit model[8] can be used to identify and flag off any such incorrect records to improve the quality of regulatory reporting. |



| Use case | Description |
|---|---|
| Monitor load and SLAs | Feeds from source systems can be daily, weekly or monthly and vary by regions, products. SLAs govern both the acquisition of source data and supply of aggregated data to the consumers. To better adhere to SLAs, supervised machine learning algorithms like regression and classification can be used on the historical data on volume and SLAs to predict the load in an incoming feed or identify potential breaches in SLAs in advance. |
| Create business rules | Supervised machine learning algorithms like decision tree algorithms can be used to learn from any existing business rules engine and use the data from the warehouse to both create new business rules and improve upon the existing business rules to assist in building a robust rules engine framework. For example, there might be a business rule which might define the loss threshold at $1 mn. The total operational loss of the business unit may be determined based on number of losses above $1 mn which also has an impact on the OpRisk capital because it ignores all the losses below $1 mn. Machine learning algorithms can help resolve such issues by categorizing loss events into different buckets of varying degrees (Bucket 1: $100,000 to $200,000, Bucket 2: $200,000 to $500,000, Bucket 3: $500,000 to $ 1mn) based on both frequency and size to determine the total loss that will help the risk teams to identify thresholds more accurately. |
| Reconciliation | Using historical data and user actions that may have resolved reconciliation issues in the past, machine learning algorithms can learn about exceptions management during reconciliation. Using machine learning algorithms like fuzzy logic, reconciliations issues can be resolved more effectively. |

## Conclusion

Authorities recognize the need to improve data quality and to take advantage of the more powerful IT systems that are available to analyze data[9]. Organizations are also increasingly looking at using data to improve their abilities in front office, middle office and back office, and not just as a tool to fulfill reporting obligations. The good news is that the technologies that require clean data have the ability to do exactly the same.The techniques suggested in this paper shall complement the existing efforts at data quality management and propel it to the next level of maturity.



## Reference


1 https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/

2 https://www.thetradenews.com/data-quality-among-top-issues-for-buy-side-fixed-income-desks/

3 According to Basel II, banks use one of the BIA, Standardized approach or AMA approach to calculate the operational risk capital.

4 https://www.bis.org/ifc/events/ifc_nbb_workshop/ifc_nbb_workshop_2d3.pdf

5 https://www.finextra.com/news/fullstory.aspx?newsitemid=14643

6 https://www.bloomberg.com/professional/blog/worlds-biggest-banks-fined-321-billion-since-financial-crisis/

7 https://finopsinfo.com/regulations/aml-exams-data-quality-takes-center-stage/

8 https://www.bis.org/ifc/events/ifc_nbb_workshop/ifc_nbb_workshop_2d3.pdf

9 https://www.fsb.org/wp-content/uploads/P290617-1.pdf


## About the Author


**Aadarsh Raghavan**

*Senior Consultant – Financial Services, Infosys Consulting*


Aadarsh is a business consultant with experience in executing advisory and delivery engagements, primarily in the Capital Markets and Risk Advisory domains, for financial institutions in the US, UK and India. Select experience involves working in engagements on areas such as Finance Transformation, Trading, Clearing and Settlement, Operational Risk Management, Data Management, Wealth and Asset Management.


For more information, contact askus@infosys.com


Infosys®
Navigate your next