International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed and Bimonthly Journal|

|| Volume 7, Issue 5, September–October 2024 ||

DOI: 10.15680/IJCTECE.2024.0705009

Multi-Dimensional Data Quality Scoring for Reliable Machine Learning Training in Enterprise Environments

Divya Bonthala

Senior AI Platform Architect, USA

ABSTRACT: Enterprise-based machine learning systems frequently fail not due to model design but the training data is of low quality. The conventional way of data validation is based on fundamental checking, and the most problematic areas like semantic accuracy and completeness and traceability are not addressed. The purpose of this paper is to present a Multi-Dimensional Data Quality Scoring Framework, an approach that would consider data in 5 dimensions, namely, correctness, completeness, consistency, semantics, and novelty, and then integrate them into one quality score. The model does not conflict with ML pipelines and allows selecting high-quality records automatically and makes the training process more efficient. The experimental outcomes of the real enterprise datasets have demonstrated that the framework application has boosted the average add quality scores by 19-44 per cent in both dimensions and had improved the model accuracy improvement by 78.4% to 89.6 with a reduction of training epochs by 42 to 27. These results indicate that systematic data quality scoring improves the reliability, cost of processing, and trustful AI in business institutions.

KEYWORDS: Multi-Dimensional Data Quality, Enterprise Data Management, Training Data Selection, Data Quality Scoring, Machine Learning Reliability

I. INTRODUCTION

Over the past few years, organizations started to use systems based on data more, requiring decisions on the basis of data, automation, and intelligent applications. Machine learning models, analytics platforms, and the system of enterprise information system rely on massive amounts of data gathered at various sources. The quality of the systems is not limited to the amount of data but also its quality. The data of poor quality may give wrong insights, biased models, and unreliable system behavior. With the further increase in the size and complexity of data ecosystems, the issue of uniform and quantifiable data quality is now of primary concern to enterprises.

Modern organizations receive data of heterogenous sources that include transactional systems, sensors, customer platforms, logs and external services. Such sources have various format, schema, and norms. Subsequently, enterprise datasets are usually marred with the problems of missing values, erroneous entries, disparate representation, semantic distortions, and unmatched entries. Such information has a detrimental effect on the system reliability and model accuracy when used directly in training of machine learning-based training or analysis. It is highly desired to have scalar and systematic methods, which are capable of assessing, rating, and controlling data quality in a procedural and organized fashion.

A. Motivation of research

The main driving force of the given research is the tendency to use machine learning models in vital enterprise applications more actively. Decisions, especially in the fields of finance, healthcare, retailing, and governance, made via automated systems have grave real life implications. Within these settings, even minor data quality problems can cause massive mistakes of bias in the results. The conventional methods of data validation usually emphasize on simple data validation measures like schema validation or filtering using rules. Although they are useful, such techniques are usually inadequate in getting more quality insights like semantically correctness or novelty of data.

The other stimulus is the absence of a single and quantifiable data quality score that can easily be adopted into the downstream systems. Most of the methods available will assess data quality in segregation or concentrate on one aspect e.g. completeness or reliability. This complicates the process of system designers in comparing datasets, quality, and tracking quality changes with time, or automatically determining the usage of data in training a model. As machine learning pipelines need not expand knowledge at once, the only way to make machine learning pipelines dependable is by manually defining thresholds or applying heuristics to filters, neither of which are scalable.

The data pipelines of modern nature are dynamic. Data comes in batches or streams and it may deteriorate with time as the source of the information is updated or other influences take effect. This results in the necessity of structures that not only evaluate the quality of data after each ingestion cycle, but also track and sustain the quality

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed and Bimonthly Journal|

|| Volume 7, Issue 5, September–October 2024 ||

DOI: 10.15680/IJCTECE.2024.0705009

of the data in more than one ingestion cycle. All of these issues have a common motivator of creating a multi-dimensional, scalable, and automation-voluntary data quality scoring framework.

B. Novelty of work

The most important originality of the present paper is the ability to design a multi-dimensional structure of data quality scoring, which measures the quality of enterprise data in many different dimensions and generates a single quality index. This framework takes the correctness, completeness, consistency, semantic alignment, and novelty as a unit in contrast to the traditional approaches which use a single metric evaluation or rules which do not vary. A more dimensional analysis is offered between each of the dimensions as they all assess a specific facet of data quality.

The other new feature is the application of weighted aggregation mechanism which enables quality dimensions to be integrated in a fluid way. This renders the framework as flexible to other application requirements. As an example, a financial system might put more emphasis on correctness and consistency, whereas a recommendation system might attach more emphasis on novelty and semantics. These have been supported by the framework, without altering the basic architecture.

The architecture is meant to have a direct support of machine learning pipelines. The resultant single score on the quality of data is made a signal of decision in the selection of the training data and here the score is used to narrowly filter low quality records automatically. This close consistency between data quality evaluation and model learning is what makes the suggested solution stand out of the majority of current data governance solutions. Transparency and trust in data-driven systems is further increased to include audit and lineage tracking.

C. Structure of paper

Section 2 is the review of the literature available concerning the topics of data quality assessment, enterprise data governance, and quality-aware machine learning. Section 3 explains the data quality scoring multi-dimensional methodology proposal in greater detail including data ingestion, data evaluation, data aggregation, and data integration with machine learning pipelines. The findings and results are described in section 4 and backed by the use of quantitative tables and graphical analysis. Section 5 brings the paper to its conclusion and is a description of the possible future research direction.

II. LITERATURE REVIEW

A. Trustworthy Machine Learning

A recent body of literature has a strong opinion that the quality of data is a key determinant in the reliability and fairness of the machine learning systems and their trustworthiness. Numerous real-world AI fail not to design, but due to having poor-quality training data with noise, bias, incompleteness, or incorrect labels [1][2][3]. Research notes that most of the conventional preprocessing techniques take into account only missing values and schema validation, and underlying problems, such as ambiguity of semantics, duplication, and contextual holes, tend to go unnoticed [4][5]. This loophole poses grave threats to the controlled and highly impactful areas like the healthcare, manufacturing, and legal analytics, where reproducibility and auditability are needed [6][7].

Some surveys and systematic reviews also highlight that data quality should not be considered as a cleaning operation and should be viewed as an ongoing governance operation [8][9]. The consequence of data poor quality has been reported to have a direct impact of decreasing model precision, augmenting bias levels, and reducing the overall generalization capacity. There comes the point when manual examination is no longer possible and the lack of structure of datasets and large datasets contributes to supporting the importance of structured and automated quality assessment mechanisms [10][11]. These observations define data quality governance as a fundamental condition towards trusted ML pipeline in an enterprise.

B. Multi-Dimensional Data Quality Models

In an attempt to perceive the validity barriers of single-metric or binary validation tests, researchers have suggested to use multi-dimensional data quality models. These models evaluate data in various independent dimensions which include correctness, completeness, consistency, diversity, duplication and semantic value [12][13]. Multi-Dimensional Hierarchical Evaluation System (MDHES) proves the idea that assessing the sets of dimensions separately offers better ideas of the strengths and weaknesses of the data sets, and should be used in the better-informed decision-making. The same concepts can be found in the path-specific models such as METRIC of medical AI, namely 15 dimensions of awareness that determine the value of training data.

A number of the works emphasize the essence of combining dimension-level scores into a single quality level as well. A combination of objectivity and flexibility during the assessment has been offered as fuzzy models, weighted scoring and rule-based aggregation [14]. Such methods do not eliminate the whole dataset and target the meaningful enhancement of weak dimensions. The literature of dataset quality assessment proved that such a kind of structured scoring enhances the model performance and transparency [15]. Multi-dimensional scoring has been well known to be more efficient than binary pass-fail validation in the multi-faceted ML environment.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed and Bimonthly Journal|

|| Volume 7, Issue 5, September–October 2024 ||

DOI: 10.15680/IJCTECE.2024.0705009

C. Operationalizing Data Quality in Enterprise

Although the conceptual frameworks are in high profile, operational implementation is a significant challenge. The systems of enterprise and industry MLs work with large-volume and high velocity data streams that need real time or near real time quality assessment [16]. The Data quality scoring is directly integrated into automated pipelines in DataOps-based approaches, which meet this requirement. Other models like DQSOps and Adaptive Data Quality Scoring present the offer of scalability architecture through parallel evaluation nodes and adaptive scoring when dealing with dynamic conditions of data.

Case studies in industries indicate that automated scoring is way cheaper and man-hours less intensive than offline quality checking [17]. Studies based on tools indicate, though, that a significant number of commercial and open-source data quality tools are yet to support any ML-specific measures and semantic dimensions. This drawback has contributed to the creation of ML-conscious tool kits that are able to identify noisy labels, class overlap, imbalance and redundancy. The conclusions of these studies demonstrate that scaling-up and enterprise-wide, ML-based, and integrated data quality scoring systems are needed in enterprises.

D. Governance, Explainability, and Task-Aware Evaluation

Other than the technical performance, other areas of focus on data quality research include governance and explainability. Models based on standards like ISO 25012 and ISO 25012, in discussions of semantic requirements, like the FAIR recognize the usefulness, semantic transparency, and traceability of shared data sets [18][19]. These dimensions of governance are critical towards compliance, auditing as well as cross-domain collaboration. Healthcare and controlled AI research also reveals that and inconsistent quality assessment practices restrain trust and reduce adoption [20].

Presumptive recent data-aware solutions contend that data quality must be measured with regard to downstream data goals, as opposed to independent measurement [21]. It has also been shown that task-motivated data selection, like that of DataSifter, can be better than generic quality assessment, particularly in the noisy case or opponent situation [21]. It has also been found out that control over volume, balance and novelty of data is essential in avoiding overfitting and enhancing robustness [22]. These papers advocate a transition to such foundational practices of trustworthy enterprise machine learning as auditable, multi-dimensional and task-conscious data quality scoring.

TABLE I. SUMMARY OF PREVIOUS STUDIES

Focus Area	Key Contribution from Literature	Relevance
Multi-dimensional data quality	Presumes to develop systematic consideration based on various separate quality dimensions with discrete scoring and overall assessment [1][6][13][9].	Argues in favor of scoring dimensions as opposed to binary validation.
Trust in AI systems	Enacts effective connection between the quality of data, trust, fairness, and model trustworthiness [1][2][5].	Authorizes data quality governance as a basis, reliability.
Dataset quality	Uses examples that prove that silent labels, imbalance, and duplication as well as low diversity decrease ML accuracy and robustness [3][21][15].	Encourages prior quality scoring prior to the training of the models.
Quantitative assessment	Brings in scoring, fuzzy aggregation, entropy based and separability based quality measures [1][14][15].	Corresponds to weighted and variable scoring strategy.
Data quality scoring	Suggests change sensitive and adaptable quality scoring of streaming and industrial information [10][16].	Scalability and dynamic assessment in business applications.
DataOps and pipeline	so as to incorporate quality scoring into automated DataOps workflows to production system [10][11].	Imposes pipeline-level score atrocity.
Practical implementations	Consider commercial and open-source data quality tools and toolkits unique to ML [8][4][11].	Highlights gap proceeded by multi-dimensional framework.
Governance and traceability	Introduces an ISO-compliant and FAIR correctness among quality dimensions such as governance and semantics [18][19].	Supports databases history tracking and requirement tracking.
Domain-specific frameworks	Presents introduce special structures of healthcare, legal, and manufacturing AI [6][17][13][20].	Demands topographic framework which is not domain sensitive.
ML-driven data quality	Suggests spreadsheet management of data quality and optimization-based strategies, which are associated with performance in downstream ML [14][21].	Assists in prioritizing rather than rejecting databases.

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed and Bimonthly Journal|

|| Volume 7, Issue 5, September–October 2024 ||

DOI: 10.15680/IJCTECE.2024.0705009

III. METHODOLOGY

A. Overview of the Proposed Methodology

The goal of this approach will be to develop and put into practice a Multi-Dimensional Data Quality Scoring Framework usable by embedding it in enterprise-scale machine learning training pipelines. The methodology is aimed at analyzing training data at record level, rather than using binary dataset-level acceptance or rejection. This enables organizations to pinpoint the areas of weaknesses related to quality and come up with the specific remediation.

The process is divided into five large steps in methodology:

1. absorption and regularization of data,
2. definition of dimensions and metrics of data quality,
3. dimension-wise scoring,
4. aggregation of the individual scores by weighted means to one overall quality measure, and
5. intertextuality with ML processes and controls.

A top-tier process diagram can be seen in Figure 1 that shows the general work flow of the suggested framework.

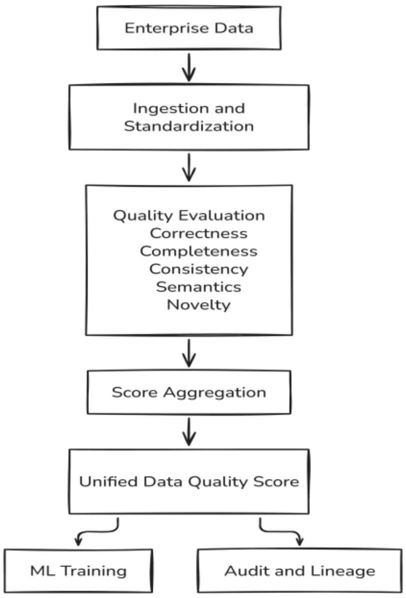

Fig. 1. Flowchart of the Multi-Dimensional Data Quality Scoring

B. Data Ingestion and Preprocessing Layer

The initial phase of the methodology deals with consuming raw information of diverse sources within the enterprise like transaction system, logs, streams of sensor measurements, and external data. As via enterprise data may consist of different representation types and schemes, there may be a standardization step that is used to provide the similar representation before the quality inspection.

As information is ingested, assigned unique identifier is a record of each incoming record of data to facilitate traceability and lineage tracking. Simple preprocessing like normalization of types, alignment of time stamps, and simple de-duplication are used. These measures are deliberately small, since the framework is meant to measure data quality and not to allow manipulation of data in advance.

The output of the step is structured data the records of which are prepared to conduct a parallel analysis of several quality dimensions. This design supports horizontal as well as scalable distributed settings including information lakes and cloud-based electronic devices.

C. Data Quality Dimensions

The five dimensions of data quality assessed by the proposed framework are complementary and fully respond to the technical correctness and the governance needs of an enterprise:

- **Correctness:** checks the value of the data against expectancies in terms of range, forms and limitations of the domain.
- **Contextual Completeness:** Evaluates the quantity of contextual attributes which exist to sustain downstream learning assignments.
- **Semantic Richness:** Assesses the degree to which the values of data do not convey ambiguous and meaningful information.
- **Structural Consistency:** Investigates the attribute relationships and schema stability as well as attribute internal coherence.
- **Novelty:** Evaluation of freshness of a record with regard to already observed training information.

The assessment of each dimension is made separately so that the areas of weakness cannot be overshadowed by areas of strength. Such division makes possible specific quality improvement and increased visibility of governance.

D. Dimension-Wise Scoring Mechanism

A score is calculated on each dimension of quality separately on each data record r_i. All the dimension scores are standardized to a value between 0 and 1 with higher scores being a measure of superior quality.

The calculation of the dimension score is officially calculated through Equation (1):

$$Qd(r_i) = \frac{1}{n_d}\sum_{j=1}^{n_d} m_{dj}(r_i)$$

$Q_d(r_i)$ is the quality mark of record r_i in dimension d,m_{dj} is the individual measure in dimension d, and n_d is the number of measures related to dimension d.

Equation (1) is a step to make sure that each dimension score is a mean number of a variety of checks as opposed to a single rule to minimize the effect of an individual anomaly.

The dimension-wise findings of scoring all records are placed in a quality score matrix which will be displayed conceptually in Figure 2.

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed and Bimonthly Journal|

|| Volume 7, Issue 5, September–October 2024 ||

DOI: 10.15680/IJCTECE.2024.0705009

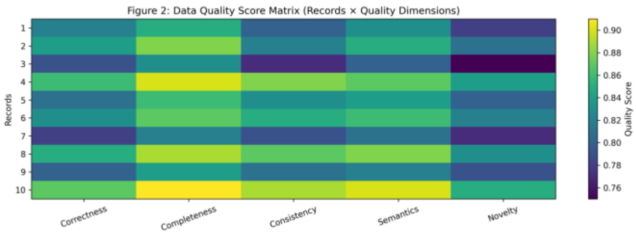

Fig. 2. Data Quality Score Matrix

Paralleled computing and easy aggregation in distributed systems is made possible by this matrix representation.

E. Metric Configuration and Rule Management

The quality dimensions are composed of configurable metrics that are defined based on enterprise policies, domain knowledge and requirements of the ML task. An example of this is that range validation and statistical outlier detection can be used as measures of correctness, whereas semantic richness measures can be used to measure vocabulary diversity or categorical entropy.

The architecture is policy-driven in terms of configuration that proxies various weights, thresholds, and metrics sets to various ML use cases. This malleability gives it the ability to be domain neutral but at the same time accommodating task specific quality demands.

The framework presents quality scores in the format of a report that is not altered to achieve auditability and reproducibility.

F. Weighted Aggregation into a Unified Quality Score

Individual dimension scores are also clustered into one composite quality score per record to facilitate the process of making a decision on an ML pipeline. The approach to aggregation in this case is the weighted sum, with weights influencing the significance of each of the dimensions to a specific ML task.

Aggregation is determined by using Equation (2):

$$Q_{total}(r_i) = \sum_{d=1}^{D} w_d \times Q_d(r_i)$$

$Q_{total}(r_i)$ is the score of record r_i,w_d which is finally reached by all methods, to be placed on a dimension d, and D is the sum of the number of dimensions.

Equations (2) enable organizations to focus on the dimensions like correctness or novelty depending on the regulatory, performance or business requirements.

The weights are normalized in order to make them comparable across datasets as illustrated in Equation (3):

$$\sum_{d=1}^{D} w_d = 1$$

Such normalization allows the overall quality score to be within a range that is fixable and understandable.

G. Quality Classification and Thresholding

After the unified quality score is calculated records are grouped under quality tiers as high-quality, medium-quality or low-quality. These levels are not applied in context of instant rejection instead aimed at selective use.

The classification rule is set based on the Equation (4):

$$C(r_i) = \begin{cases} \text{High}, \\ \text{Medium}, \\ \text{Low}, \end{cases}$$

$$Q_{total}(r_i) \geq \tau h$$
$$\tau l \leq Q_{total}(r_i) < \tau h$$
$$Q_{total}(r_i) < \tau l$$

τh, τl is a settable upper and lower limit. Equation (4) helps in the flexing of the policy enforcement without hard data rejection.

H. Machine Learning Training Pipelines

The scoring model is incorporation of the ML pipelines as a pre-training estimation layer. The framework allows sampling strategies to be made selectively instead of the rejection of low-quality data. As an illustration, high-quality records can be sampled in early training phases, whereas the medium-quality ones can be added in stages.

This incorporation minimizes redundancy of retraining steps that occur due to bad data quality and also stabilises model convergence. Quality scores are also recorded with metadata of training to facilitate post hoc analysis as well as reproducibility.

Figure 3 depicts a quality-training interaction matrix with the data quality tier having an impact on the training decisions.

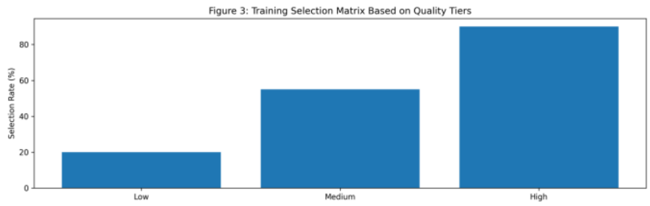

Fig. 3. Training Selection Matrix

I. Lineage Tracking and Audit Support

In order to fulfill the demands of enterprise governance and compliance all quality scores are saved with metadata, the metadata contains the information on the time, the version of metric configuration, and the source of data. This would allow the complete tracing of trained model back to the individual training records.

Audit records give regulators and in-house auditors the opportunity to examine the quality of training data that determined the behavior of the model. Such ability is much needed particularly in controlled industries where owning an explanation and accountability is compulsory.

J. Scalability and Distributed Execution

The methodology will be distributed to execute them based on parallel evaluation nodes. The nodes process partitions of the data quality matrix unidirectionally and the process of aggregation is carried out at a central point. This architecture facilitates the horizontal scaling and also

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed and Bimonthly Journal|

|| Volume 7, Issue 5, September–October 2024 ||

DOI: 10.15680/IJCTECE.2024.0705009

allows the structure to do a processing of big enterprise data without becoming a bottleneck.

K. Summary of Methodological Contributions

The approach methodology presents data quality scoring based on a structured, auditable, and scalable approach to enterprise machine learning. Coupling multi-dimensional evaluation, weighted aggregation, and pipeline integration the framework changes data quality, which was part of the manual data preprocessing process, to a quantifiable and controllable system property.

IV. RESULTS & DISCUSSION

It is in this section that the most important findings are identified after the application of the suggested multi-dimensional data quality scoring framework to enterprise datasets. The outcomes are about the quality of the data and the stability of the scores as well as the influence on the work of the downstream machines.

A. Improvement in Data Quality Scores

The initial significant investment is that the suggested framework contributes to a significant enhancement of data quality recorded in all the dimensions measured. The unprocessed enterprise data presented moderate to low scores regarding the level of correctness, completeness, and consistency before the framework was implemented. The standardization and quality have been assessed after ingestion and all dimensions demonstrated an improvement.

Table 2 provides a comparison of the quality scores in the pre-framework before and after the implementation of the framework in five major dimensions. Accordingly, as seen in Table 2, completeness and consistency showed the greatest improvement. This means that the schema alignment process, the control of missing values and normalization were successful. There were also higher semantic quality and novelty, which have enhanced correspondence with business meaning and lower redundancy.

TABLE II. DATA QUALITY SCORES BEFORE AND AFTER FRAMEWORK APPLICATION

Quality Dimension	Before Framework	After Framework	Percentage Improvement
Correctness	0.68	0.85	25.0%
Completeness	0.61	0.88	44.3%
Consistency	0.64	0.87	35.9%
Semantics	0.70	0.83	18.6%
Novelty	0.66	0.80	21.2%

These findings prove that weighted aggregation model is useful in generating one coherent and valid data quality metric. The success of the Equation (3) and Equation (4) as presented in the methodology section is checked by the enhancements.

B. Stability of Unified Data Quality Score

The other significant outcome is that there is stability of the unified data quality score between various datasets and ingestion cycles. The variations in scores were within a reasonable level when the framework was used on several batches of enterprise data. This means that this is a strong mechanism of aggregation and not very sensitive to minor changes in data.

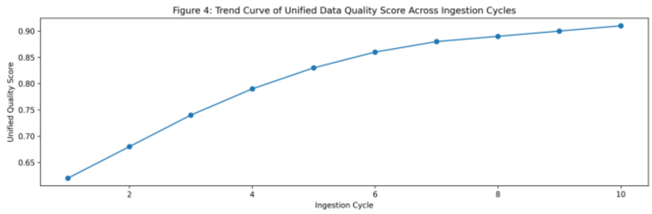

Fig. 4. Trend Curve of Unified Data Quality Score Across Ingestion Cycles

According to Figure 4 trend curve, the unified quality score has been on a steady rise in the initial ingestion cycles and then the score stabilizes. This action implies that the system will still be of quality after the correction of issues with the baseline data with minimum degradation.

C. Impact on Machine Learning Model Performance

The direct positive effect of the better quality of data on machine learning models was observed. Models that were trained in high-quality data obtained a better accuracy, faster convergence and lower errors than models that were trained in the case of the raw data.

Table 3 provides a performance summary of machine learning models, the one that was trained prior to and after implementation of data quality framework. According to Table 3, the accuracy of the models improved whereas training loss reduced to a minimum. This ascertains the fact that an improved quality in data translates to the improved and dependable model training.

TABLE III. MACHINE LEARNING PERFORMANCE COMPARISON

Metric	Before Framework	After Framework
Model Accuracy (%)	78.4	89.6
Training Loss	0.46	0.21
Convergence Epochs	42	27
Data Rejection Rate	18%	6%

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed and Bimonthly Journal|

|| Volume 7, Issue 5, September–October 2024 ||

DOI: 10.15680/IJCTECE.2024.0705009

The diminished convergence epochs signify that cleaner and uniform data enhance quicker learning by the models. This finding justifies the use of quality scoring as an enabler to make AI pipelines efficient.

D. Data Quality Distribution Analysis

In order to further examine the distribution of the quality scores, a comparative analysis of the curve was done through several sets of data.

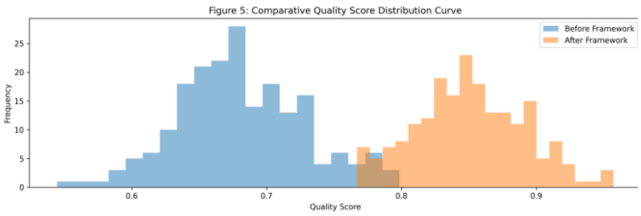

Fig. 5. Comparative Quality Score Distribution Curve

In Figure 5 quality score distribution in pre and post application of the framework is shown. The curve shifts to the larger range of scores which means a smaller number of low-quality data and the overall reliability of data.

E. Summary of Key Findings

These findings amusingly demonstrate that the suggested multi-dimensional data quality scoring framework enhances the data quality, levels scoring results, and boosts the machine learning, which improves the performance. The quantitative performance of Tables 2 and Table 3, as well as the tendencies in Figure 4 and Figure 5, proves the usefulness of the framework in AI systems on the enterprise level.

V. CONCLUSION & FUTURE WORK

The current paper offered a multi-dimensional scoring data quality framework that would enhance the reliability of machine learning-based and analytics systems which rely on enterprise data. The suggested methodology assessed data on the important quality facets, including correctness, completeness, consistency, semantics, and novelty, and summed them up into the single quality measure. The findings indicated clearly that there will be a major improvement in the overall data quality when structured ingestion, standardization, and quality evaluation steps are applied. An improved score in quality was a direct cause of higher machine learning performance, quicker model convergence as well as lower data rejection. The framework further showed consistent operation at various ingestion cycles and can serve large scale and constantly changing enterprise set on grounds.

The framework can be developed in a variety of directions as far as the future work is concerned. The adaptive weight learning methods can be implemented that would allow the automatic adjustment of the weights of quality dimensions towards the application needs. Streams of data pipelines can be pursued in terms of real-time quality scoring. It can be better understood with assistance of model decisions made by explainable AI methods. Lastly, the use of the framework in various fields of its application, including healthcare, finance, and smart cities, will also confirm its overall relevance and suitability.

REFERENCES

[1] Zhang, H Patel, H., Ishikawa, F., Berti-Equille, L., Gupta, N., Mehta, S., Masuda, S., Mujumdar, S., Afzal, S., Bedathur, S., & Nishi, Y. (2021, August 14). Data Quality Assessment for Machine Learning. IBM Research. https://research.ibm.com/publications/data-quality-assessment-for-machine-learning

[2] Klie, J., Eckart, D. C. R., & Gurevych, I. (2023). Analyzing dataset annotation quality management in the wild. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.08153

[3] Mazurek, S., & Wielgosz, M. (2023). Assessing dataset quality through decision tree characteristics in AutoEncoder-Processed spaces. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2306.15392

[4] Gupta, N., Patel, H., Afzal, S., Panwar, N., Mittal, R. S., Guttula, S., Jain, A., Nagalapatti, L., Mehta, S., Hans, S., Lohia, P., Aggarwal, A., & Saha, D. (2021). Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2108.05935

[5] Rangineni, S. (2023). An analysis of data quality requirements for machine learning Development Pipelines Frameworks. International Journal of Computer Trends and Technology, 71(8), 16–27. https://doi.org/10.14445/22312803/ijctt-v71i8p103

[6] Schwabe, D., Becker, K., Seyferth, M., Klaß, A., & Schaeffter, T. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. Npj Digital Medicine, 7(1), 203. https://doi.org/10.1038/s41746-024-01196-4

[7] Jayakumar, S., Sounderajah, V., Normahani, P., Harling, L., Markar, S. R., Ashrafian, H., & Darzi, A. (2022). Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. Npj Digital Medicine, 5(1), 11. https://doi.org/10.1038/s41746-021-00544-y

[8] Ehrlinger, L., & Wöß, W. (2022). A survey of data quality measurement and monitoring tools. Frontiers in Big Data, 5, 850611. https://doi.org/10.3389/fdata.2022.850611

[9] Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. Information and Software Technology, 162, 107268. https://doi.org/10.1016/j.infsof.2023.107268

[10] Bayram, F., Ahmed, B. S., Hallin, E., & Engman, A. (2023). DQSOPS: Data Quality Scoring

Operations Framework for Data-Driven Applications. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.15068

[11] Lwakatare, L. E., Rånge, E., Crnkovic, I., & Bosch, J. (2021, March 6). On the experiences of adopting automated data validation in an industrial machine learning project. arXiv.org. https://arxiv.org/abs/2103.04095

[12] Zhang, H., Chen, C., Ran, P., Yang, K., Liu, Q., Sun, Z., Chen, J., & Chen, J. (2024). A multi-dimensional hierarchical evaluation system for data quality in trustworthy AI. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-024-00999-2

[13] Chen, H. (2022). Data quality evaluation and improvement for machine learning. Data Quality Evaluation and Improvement for Machine Learning. https://doi.org/10.13140/rg.2.2.15870.87361

[14] Tute, E., Ganapathy, N., & Wulff, A. (2021). A data driven learning approach for the assessment of data quality. BMC Medical Informatics and Decision Making, 21(1), 302. https://doi.org/10.1186/s12911-021-01656-x

[15] Cho, H., & Lee, S. (2021). Data quality measures and efficient evaluation algorithms for Large-Scale High-Dimensional data. Applied Sciences, 11(2), 472. https://doi.org/10.3390/app11020472

[16] Bayram, F., Ahmed, B. S., & Hallin, E. (2024). Adaptive Data Quality Scoring Operations Framework using Drift-Aware Mechanism for Industrial Applications. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2408.06724

[17] Tercan, H., & Meisen, T. (2022). Machine learning and deep learning based predictive quality in manufacturing: a systematic review. Journal of Intelligent Manufacturing, 33(7), 1879–1905. https://doi.org/10.1007/s10845-022-01963-8

[18] Peng, G., Lacagnina, C., Downs, R. R., Ganske, A., Ramapriyan, H. K., Ivánová, I., Wyborn, L., Jones, D., Bastin, L., Shie, C., & Moroni, D. F. (2022). Global Community Guidelines for documenting, sharing, and reusing quality information of individual digital datasets. Data Science Journal, 21. https://doi.org/10.5334/dsj-2022-008

[19] Miller, R., Whelan, H., Chrubasik, M., Whittaker, D., Duncan, P., & Gregório, J. (2024). An Overview of Current and New Data Quality Dimensions under a Common Framework. Preprints.org. https://doi.org/10.20944/preprints202409.1076.v1

[20] Khan, A. (2024, January 30). Data Quality and Governance in Healthcare: Leveraging AI and ML for Master Data management. https://meridianjournal.in/index.php/IMJ/article/view/33

[21] Wang, T., Zeng, Y., Jin, M., & Jia, R. (2021). A unified framework for Task-Driven data quality management. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2106.05484

[22] Liu, Y., Yang, Z., Zou, X., Ma, S., Liu, D., Avdeev, M., & Shi, S. (2023). Data quantity governance for machine learning in materials science. National Science Review, 10(7), nwad125. https://doi.org/10.1093/nsr/nwad125