Root Cause Analysis and Remediation for Quality and Value Improvement in Machine Learning Driven Information Models

Shelernaz Azimi and Claus Pahl
Free University of Bozen - Bolzano, Bolzano, Italy

Keywords: Data Quality, Information Quality, Information Value, Machine Learning, Data Quality Improvement, Data Analysis, Root Cause Analysis, Data Quality Remediation.

Abstract: Data quality is an important factor that determines the value of information in organisations. Information creates financial value, but depends largely on the quality of the underlying data. Today, data is more and more processed using machine-learning techniques applied to data in order to convert raw source data into valuable information. Furthermore, data and information are not directly accessed by their users, but are provided in the form of 'as-a-service' offerings. We introduce here a framework based on a number of quality factors for machine-learning generated information models. Our aim is to link back the quality of these machine-learned information models to the quality of the underlying source data. This would enable to (i) determine the cause of information quality deficiencies arising from machine-learned information models in the data space and (ii) allowing to rectify problems by proposing remedial actions at data level and increase the overall value. We will investigate this for data in the Internet-of-Things context.

1 INTRODUCTION

Large volumes of data are today continuously produced in many contexts. The Internet-of-Things (IoT) is a so-called big data context where high 'volumes' of a 'variety' of data types are produced with high 'velocity' (speed), often subject to 'veracity' (uncertainty) concerns (Saha and Srivastava, 2014). Another aspect of this V-model for big data is 'value' that needs to be created from data (Nguyen, 2018).

Raw data originating from various sources needs to be structured and organised to provide information that is then ready for consumption, i.e., providing value for the consumer. In recent times, machine learning (ML) is more and more often used to derive particularly non-obvious information from raw data, thus enhancing the value of that information for the consumer. Machine learning creates valuable information when manual processing and creation of functions on data is not possible due to time and space needs. Value is created if this information aids in monetising data in products or services that are provided. Information can also support organisations in improving operational and strategic decision making. Furthermore, self-adaptive systems can be controlled by this information, even dynamically.

The impact of data volume, variety, velocity and veracity on the quality and value can be a challenge, particularly if the information is derived through a machine learning approach. In order to better frame the problem, we need to define a quality framework that links data and the ML function level. We aim here to close the loop, i.e., mapping ML functional quality problems back to their data origins by identifying the symptoms of low quality precisely and map these to the root causes of these deficiencies. Furthermore, remedial actions to solve the data quality problem shall ultimately be proposed by our framework.

Our contribution in this paper consists of two parts: firstly, a layered data and information architecture for data and ML function layers with associated quality aspects; secondly, a symptom and root cause analysis, closing the loop to link observed quality concerns at ML model level to data quality at the source data level that might have caused the observed problems. This extends work presented in (Azimi and Pahl, 2020a), in particular in the second aspect, but also using a different core quality model here. The novelty of our approach lies in, firstly, the layering of data and ML model quality based on dedicated ML function types and, secondly, when data quality might not be directly observable, we provide a new way of inferring quality problem causes when needed. It also complements work in (Ehrlinger et al., 2019)

Azimi, S. and Pahl, C.
Root Cause Analysis and Remediation for Quality and Value Improvement in Machine Learning Driven Information Models.
DOI: 10.5220/0009783106560665
In *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020) - Volume 1*, pages 656-665
ISBN: 978-989-758-423-7

where ML quality analysis is proposed for an Industry 4.0 use case, but without providing a comprehensive quality framework. We report on case studies that we are conducting with a regional IT solution and service provider around Internet-of-Things applications (IoT). IoT is a typical domain that satisfies the V-model of big data. Therefore, we use IoT here as the application context in order to make qualities and impacting factors more concrete.

2 BACKGROUND TECHNOLOGIES

With our investigation, we target here the quality of information, specifically information that is created from data by using machine learning techniques. We will briefly introduce these aspects and also explain the role of IoT as the chosen application domain here. Data is a valuable asset in the IoT technology domain as a source for creating information and knowledge.

Data Quality: refers to how well data meets the requirements of its users. Each data user or consumer expects the respective data to meet given criteria that are essential for a task or objective. These criteria (also referred to as aspects or attributes of data quality) are for example Accuracy, Timeliness, Precision, Completeness, or Reliability.

Quality frameworks for data and information have already been presented (O'Brien et al., 2013). There is also a commonly accepted classification of (big) data aspects that can help in organising and managing the quality concerns, often called the 4V model (Saha and Srivastava, 2014; Nguyen, 2018): volume (scale, size), velocity (change rate/streaming/real-time), variety (form/format) and veracity (uncertainty, accuracy, applicability). Our chosen IoT domain exhibits all of those characteristics.

Machine Learning: (ML) techniques build a formal model based on given data (the training data) aiming to make predictions or decisions without having been programmed to do this. Machine learning techniques are typically classified into supervised learning, unsupervised learning and reinforcement learning. In *supervised learning*, the machine learning algorithm builds a formal model from a set of data that contains both the inputs and the desired outputs. *Classification and regression algorithms* are types of supervised learning. Classification is used when the output is a discrete number and regression is used when the output is a continuous one. In *unsupervised learning*, applying ML builds a model from a set of data that contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. *Reinforcement learning* algorithms are given feedback in the form of positive or negative reinforcement in a dynamic environment.

In the **Internet-of-Things** (IoT), so-called things (such as sensors and actuators) produce and consume data in order to provide services (Pahl et al., 2018; Azimi and Pahl, 2020b). In case the underlying data is inaccurate, then any extracted information and knowledge and also derived actions based on it are likely to be unsound. Furthermore, the environment in which the data harvesting occurs is often rapidly changing and volatile. As a result, many characteristics such uncertain, erroneous, noisy, distributed and voluminous apply (Pahl et al., 2019).

IoT is the application context here. In order to focus our investigation, we make the following assumptions: (i) all data is numerical in nature (i.e., text or multimedia data and corresponding quality concerns regarding formatting and syntax are not considered here) and (ii) data can be stateful or stateless. Thus, IoT is here a representative application domain for our investigation characterised as a V-model-compliant big data context with a specific set of applicable data types, making our results transferable to similar technical environments.

3 RELATED WORK

The related work shall now be discussed in terms of data level, machine learning process perspective and machine learning model layer aspects separately.

Data level quality was investigated in (O'Brien et al., 2013),(Casado-Vara et al., 2018),(Sicari et al., 2016). In the first paper, data quality problems where classified into 2 groups of context-independent and context-dependant from the data and user perspective and in the second one, a new architecture based on Blockchain technology was proposed to improve data quality and false data detection. In the third paper, a lightweight and cross-domain prototype of a distributed architecture for IoT was also presented, supporting the assessment of data quality. We adapt here (O'Brien et al., 2013) to out IoT application context.

The ML process perspective was discussed in (Amershi et al.,). A machine learning workflow with nine stages was presented in which the early stages are data oriented. Usually the workflows connected to machine learning are highly non-linear and often contain several feedback loops to previous stages. If the system contain multiple machine learning components, which interact together in complex and unexpected ways, this workflow can become more complex. We investigate here a broader loop from the later final ML function stages to the initial data and ML training configuration stages, which has not been comprehensively attempted yet.

The machine learning model layer has been studied in multiple papers (Plewczynski et al., 2006), (Caruana and Niculescu-Mizil, 2006), (Kleiman and Page, 2019), (Sridhar et al., 2018), (Ehrlinger et al., 2019). Different supervised learning approaches were used. They observed that different methods have different applications and analysed in this context the effect of calibrating the models via Platt scaling and isotonic regression on their performance as a quality concern.

In some of the above papers, specific quality metrics applied to machine learning techniques have been presented. (Kleiman and Page, 2019) for example discusses the area under the receiver operating characteristic curve (AUC) as an instance of quality for classification models. In (Sridhar et al., 2018), the authors propose a solution for model governance in production machine learning. In their approach, one can meaningfully track and understand the who, where, what, when and how a machine learning prediction came to be. Also the quality of data in machine learning has been investigated. An application use case was presented with no systematic coverage of quality aspects. We aim here to condense the different individual quality concerns in a joint ML-level model.

4 INFORMATION AND DATA QUALITY: ANALYSIS AND REMEDIATION

Information is created from data by organising and structuring raw data that originates from data-producing sources (e.g., sensors in IoT environments), thus adding meaning and consequently value to data. We can illustrate the value aspect in different IoT applications (we choose weather and mobility here): (a) paid weather forecasting service, i.e., direct monetisation of the data and information takes place [weather]; (b) long-term strategic decisions, e.g., city planning, can be based on road mobility patterns [mobility]; (c) short-term operational planning, e.g., event management in city or region can be based on common and extraordinary mobility behaviour [mobility]; (d) immediate operation, e.g., in self-adaptive traffic management systems such as situation-dependent traffic lights [mobility].

In the remainder of this section, we introduce the context of data and ML with the respective quality models in Subsections 4.1, 4.2 and 4.3, then present in 4.3 the architecture of the feedback loop, address the se analysis for observed quality problems in 4.5 and 4.6, and finally look into remediation and the automation of the process in 4.7 and 4.8.

4.1 Data and Machine Learning

Our central hypothesis is that information, as opposed to just data, is increasingly provided through functions and models created using an machine learning (ML) approach. In many domains, such as IoT, there is historical information available that allows functions to be derived as machine learning models.

The ML functions fall into different categories. We distinguish here the following ML function types:

- *predictor*: predicts a future event in a state-based context were historical data is available.
- *estimator* (or calculator): is a function that aims to calculate a value for a given question, which is an estimation rather than a calculation if accuracy cannot be guaranteed.
- *adaptor*: is a function that calculates setting or configuration values in a state-based context where a system is present that can be reconfigured to produce different data.

4.2 Data and Information Quality

At the core of our framework is a layered data architecture, see Figure 1, that captures qualities of the data and the ML information model layer. Machine learning connects the two layers.

The *base layer* is the *raw data layer* consisting of unstructured and unorganised data, which would come from IoT sources in our case. Following (O'Brien et al., 2013), we can distinguish context-dependent and context-independent data quality aspects. We adjust the framework proposed in (O'Brien et al., 2013) to numeric data (i.e., we exclude text-based or image data for example):

- Context-independent data quality: missing/incomplete, duplicate, incorrect/inaccurate value, incorrect format, outdated, inconsistent/violation of generic constraint
- Context-dependent data quality: violation of domain-specific constraint

These form the lower data quality layer in Figure 1.

The *upper layer* of the model, at the top of Figure 1, is an ML-enhanced *information model*.

- To define a quality framework for the information function, we considered as input for function quality the following structural model quality:

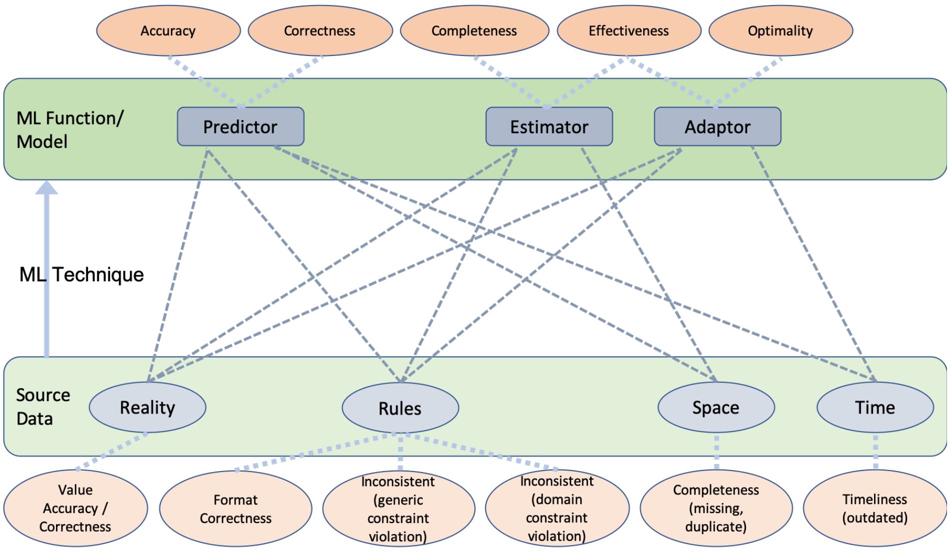

Figure 1: Layered Data and ML Information Model Architecture.

completeness, correctness, consistency, accuracy and optimality

- Based on these we associate a primary function quality aspect for each of the function types[1]:
  - predictor: correctness, accuracy.
  - estimator: effective, complete,
  - adaptor: effective, optimal.

It is essential to assess the quality provided by the ML models in order to provide value, which emerges in the different types such as predictors, estimators or adaptors. In Figure 1, we grouped source data into reality and rules aspects (this is sometimes called the intrinsic data quality category) and space and time aspects (called the contextual data quality category). We aligned the six individual qualities with these. At the ML model layer, the three functions predictor, estimator and adaptor are shown, which each of them having their primary quality concern attached.

In some situations, we need to refine the quality classification. For the adaptor function, effectiveness and optimality are criteria that often involve multiple goals, e.g., for the primary goal 'effective' for one aspect (which could be a performance threshold in a technical system), we could have as secondary goal 'optimality' for another aspect (such as energy or amount of resources sent to maintain performance).

[1]Other, so-called ethical model or function qualities such as fairness, sustainability or privacy-preservation have been introduced (Rajkomar et al., 2018). However, as there is uncertainty about their definition, we will exclude these.

4.3 ML Models Quality

The function qualities are defined in Table 1. In practical terms, the complexity of the quality calculation is of importance, since in an implementation the ML function assessment would need to be automated: The complexity of the quality assessment is a principle concern. Furthermore, often we need to wait for an actual observable result event (adaptor) as an example. We will return to this automation aspect later.

Table 1: Information Quality Definitions.

Quality	Quality Definition
correctness	Correctness is a boolean value that indicates whether a prediction was successful
accuracy	Accuracy is the degree to which a prediction was successful
effective	Effectiveness is a boolean value that incdicates the correctness of a calculation
complete	Completeness is the degree to which a estimator covers the whole input space
optimal	Optimility is a boolean value indicating whether the optimimal solution has been reached

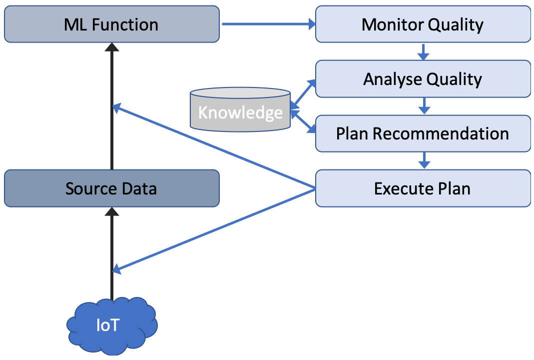

Figure 2: Closed Feedback Loop.

4.4 Quality Analysis: Architecture and MAPE-K Feedback Loop

Our objective is to analyse the reasons behind possible poor quality and performance of ML models and to identify insufficient data qualities either in the training data selection or the collected raw data as the root causes of the observed ML quality deficiency.

Our proposed *quality processing architecture* – shown in Figure 2 – implements the so-called MAPE-K control loop for self-adaptive systems. As inputs we have raw/source data from sources such as sensors in the IoT case. the MAPE-K feedback loop works as follows: *Monitor*: continuously monitor the performance of the ML models; *Analyse*: analyse the causes of possible quality problems; *Plan*: identify root causes and recommend remedial strategies; *Execute*: implement the recommended remedies and improvements. K represents the *Knowledge* component with the monitoring data, analysis mechanisms and catalog of proposable remedies. Output is an enhanced ML information model after improvement that remedies the quality problems. This is a feedback loop to control data and information quality.

4.5 Root Cause Analysis: Quality Mapping from ML Model to Data

Within the quality processing architecture, the mapping of ML function quality to data quality is the core of the MAPE-K Analysis stage. In order to illustrate some principles, we select a few cases of mappings of observed ML model problems to possible underlying data quality concerns (root causes), see Table 2. Some cases depend on whether the application context is stateful or stateless as in the 'outdated' case.

Across the layers of the data and information architecture, we have shown cross-layer dependencies. In Figure 4, we can see a mesh of dependencies,

Table 2: Information Quality (upper layer) to Data Quality (lower layer) Mapping.

Information Quality (Observed Symptom)	**Data Quality (Possible Root Cause)**
Predictor accuracy	*Possible Causes*: data incomplete, data incorrect, data duplication, outdated. *Example*: count/average services per areas (hospitals) could suffer from outdated or duplicate data
Predictor correctness	same as above
Estimator effectiveness	*Possible Causes*: outdated data. *Example*: applies sometimes as in heating systems in building, when measurements are not reflect up-to-date
Adaptor ineffective	*Possible Causes*: could be caused by incorrect data format. *Example*: Celsius vs Fahrenheit in temperature measurements

with only adaptors not strictly requiring space qualities (i.e., allow systems to work in the case of incompleteness by not taking an action) and estimators not essentially based on a state/time notion.

4.6 Possible Root Causes: Data in IoT

The dependency mesh in Figure 4 shows possible causes of problems at the data layer. This can be advanced one step by also looking at the causes of data quality problems, which would in our case arise from the underlying IoT infrastructure that provides the raw data (Samir and Pahl, 2019). We analysed possible causes and categorise them as follows:

- *Deployment Scale.* IoT is often deployed on a global scale. Data comes from a variety of devices and sensors. A large number of devices increases the chance of errors and resulting low data quality.
- *Resources Constraints.* Things in IoT suffer often from a lack of resources (e.g. power, storage, etc.). Their computational and storage capabilities do not allow complex operations support. Considering the scarce resources, data collection policies, where trade-offs are generally made, are adopted, which affect the quality of data.
- *Network.* Intermittent loss of connection in IoT is frequent. IoT can be seen as a constrained IP network with a higher ratio of packet losses. Things

are often only capable of transmitting small-sized messages due to constrained resources.

- *Sensors.* Embedded sensors may lack precision or suffer from loss of calibration or even low accuracy especially when they are of low cost. Faulty sensors may also result in inconsistencies in data sensing. The casing or the measurement devices could be damaged due to extreme conditions like extreme heating or freezing which can also cause mechanical failures. The conversion operation between measured quantities is often imprecise.
- *Environment.* The sensor devices are not only deployed in safe environments. In order to monitor some phenomena (e.g., weather), sensors are deployed in environments with extreme conditions. The maintenance of such sensors is rarely ensured considering the inaccessibility of terrains. In those conditions, sensors may become non-functional or unstable due to a variety of events (e.g., snow accumulation, dirt accumulation).
- *Vandalism.* Things are generally defenseless from outside physical threats. In addition, their deployment in the open nature makes them susceptible to vandalism. Such acts often result in rendering sensors non-functional, which definitely affects the quality of produced data.
- *Fail-dirty.* Here a sensor node fails, but keeps up reporting readings which are erroneous. It is generally an important source of outlier readings.
- *Privacy Preservation Processing.* Data quality could be intentionally reduced in the context of privacy preservation processing.
- *Security Vulnerability.* Devices are vulnerable to security attacks. Their lack of resources makes them harder to protect from security threats (e.g., no support for cryptographic operations because of their high consumption of resources). It is possible for a malicious entity to alter data in sensor nodes causing data integrity problems.
- *Data Stream Processing.* Data gathered by things are sent in the form of streams to back-end applications which process them further. These data streams could be processed for a variety of purposes (e.g., extracting knowledge, decreasing the data stream volume to save up on the scarce resources). Here, data stream processing operators (e.g., selection) could, under certain conditions, affect the quality of the underlying data.

4.7 Remediation: Problem Causes and Remedial Actions

The association of root causes allows us to use analysis results for remediation and improvement. Recommendations for remedies and improvement actions can be given. Two principle recommendation targets exist, indicated in Figure 2. *Data collection*: the suggestion could be to collect other raw/source data (for instance more, different or less data), guided by the above problem causes in the IoT infrastructure domain. *ML training*: the proposition to configure other ML training/testing data to be selected in the problem can be attributed to the ML training process rather than the data quality itself.

4.8 Automation of Analysis and Remediation

Another concern is how to automate the problem cause identification. We propose here the use of statistical and probabilistic models, e.g., Hidden Markov Models (HMM) allow us to map observable ML function quality to hidden data quality via reason-based probability assignment, which could address the above assignment of root causes to symptoms. The proposal would be a probability assignment of the cause likelihood. This kind of implementation, however, remains at this stage future work.

5 VALIDATION AND EVALUATION

ML functions provide information value for (i) monetisation through services/products and (ii) for decision support for strategic (long-term), operational (mid-term) and adaptive (short-term/immediate) needs. We have already used weather and traffic data for motivation. In order to illustrate better and validate our framework, we first discuss a more detailed use case in Section 5.1, before looking at some other evaluation criteria in Section 5.2.

5.1 Use Case Validation

We now detail the Mobility case further, which actually also involves weather data. This serves here as an illustration, but also validation of our concepts.

5.1.1 Collected Data

The raw data sets from the traffic and weather sensor sources are (1) road traffic data: number of vehicles (categorised), collected every hour and is accumulated, (2) meteorological data: temperature and precipitation, collected every 5 minutes. From this, a joined data set emerges that links traffic data with the meteorological data. Since we cannot assume the weather and traffic data collection points to be co-located, for each traffic data collection point, we associate the nearest weather collection point.

5.1.2 Information Models and Their Value

Machine learning can in this situation be utilised to derive different *types of information*: (1) the *predicted* number of vehicles for the next 5 days at a certain location; (2) the *predicted* level of traffic (in 4 categories light, moderate, high, very high) for the next 5 days at a certain location; (3) an *estimation* of average number of vehicles in a particular period (which needs to be abstracted from concrete weather-dependent numbers in the data); (4) an *estimation* of the correct type of the vehicle such as car or motorbike; (5) an *adaptation* through the determination of suitable speed limits, in order to control (reduce) accidents or emissions.

The ML model creation process can use different techniques, including decision trees, random forests, KNN, neural networks etc. This ultimately driven by a need for accuracy as a key quality concern. A model will be created for each traffic location. ML model creation (training) takes into account historical data, which in our case is a full year of meteorological and traffic data for all locations.

The purpose is to support the following objectives across several *value types*, with objective and ML function: *strategic*: for road construction based on prediction/estimation; *operational*: for holiday management based on prediction; *adaptive*: for speed limits based on adaptation.

5.1.3 Quality Analysis of ML Functions

We now select four functions, covering the three function types, that shall be described in more detail in terms of their functionality and quality:

- **Strategic [Estimator].**
  - *Function*: the long-term strategic aspect is based on traffic, but not weather. The estimated average number of vehicles over different periods is here relevant.
  - *Construction*: supervised learning – classify.
  - *Quality*: effective (allows useful interpretation, i.e., effective road planning), complete (available for all stations)
- **Operational [Predictor].**
  - *Function 1*: the operational aspect needs to predict based on past weather and past traffic, taking into account a future event (holiday period here). Concrete predictions are traffic level and traffic volume (number of cars)
  - *Construction*: supervised learning – classify.
  - *Quality*: correct (right traffic level is predicted), accurate (number of cars predicted is reasonable close to the later real value)
- **Operational [Predictor].**
  - *Function 2*: a second operational function could determine the type of car, e.g., if trucks or buses should be treated differently
  - *Construction*: unsupervised – cluster.
  - *Quality*: correct (right vehicle is determined), accurate (categories determined are correct for correct input data)
- **Adaptive [Adaptor].**
  - *Function*: a self-adaptive function that changes speed limit settings autonomously, guided by an objective (such as reducing accidents or lowering emissions).
  - *Construction*: unsupervised learning – reinforcement learning.
  - *Quality*: effective (speed reduction is effective), optimal (achieves goals with proposed action)

5.1.4 Quality Root Causes in Data

The quality of the raw data can be a problem in the following cases: *incomplete*: can arise as a consequence of problems with sensor connectivity and late arrival of data (causing incompleteness until the arrival), *duplicate*: sensors might be sending data twice (e.g., if there is no acknowledgement), *incorrect*: as a consequence of sensor faults, *incorrect format*: if temperature data is send in Fahrenheit instead of Celsius as expected, *outdated*: if either the observed object has changed since data collected (road capacity has changed) or data that has arrived late, *inconsistent*: where generic consistency constraints such as 'not null' in data records are violated.

5.1.5 Use Case Discussion

With this use case, we can validate the suitability of our quality framework. The use case is sufficiently rich in features to allow meaningful statements about the framework: (i) all information *value types* (strategic, operational, adaptive) are covered, (ii) all ML *function types* (predictor, estimator, adaptor) are covered, and (iii) all ML *function qualities* are relevant and applicable. In Table 3, the root cause analysis of

Table 3: ML Model Quality Problem and Root Cause Analysis. Notes: 1 – for this supervised learning case, only true positives and false negatives apply; 2 – Example: no complete record, e.g., for a specific period and varying weather conditions; 3 – Example: no complete record, e.g., for a specific period none at all.

ML Function / Model	**ML Quality Problem**	**Reason – Data Quality**	**Technical Root Cause** (examples only)
Estimator	accuracy[1]	raw data: completeness, accuracy, consistency	completeness: sensor failure; loss of network connection
	effectiveness[2]	training data: incomplete	incomplete: similar situations not covered in training data
		raw data: accuracy, incompleteness	incomplete: sensor outages cause records to be missing
	completeness[3]	[as for effectiveness]	[examples as above]
Predictor 1	accuracy	raw data: completeness, accuracy	completeness: no data for similar situations available
		training data: incorrect labelling	similar relevant items are incorrectly labeled
	correctness	raw data: correctness	correctness: sensor failure
Predictor 2	accuracy	raw data: incorrect	incorrect sensor data
	completeness	raw data: completeness	incomplete sensor data
Adaptor	effective	raw data	sensor / environment failure
	optimal	training data: incomplete, incorrect labeling	incomplete: not all relevant categories are labeled in sufficient numbers in training data
		raw data: incorrect	caused by malfunctioning sensors

ML model quality problems for our use case is presented. The table here is not meant to be exhaustive, i.e., does not reflect a comprehensive analysis of the problem cases. The aim is to illustrate the possibility of attributing data deficiencies and, if possible, underlying root causes to the ML model problems. A note applies to the likelihood of these. The table reflects the possible problem causes. An assignment of probabilities would be possible if extensive experience with monitoring and analysing these systems existed.

5.2 Technical Evaluation

The evaluation aims at validating the proposed quality framework. Partly, the traffic use case we discussed above serves as a proof-of-concept application. However, we also cover other criteria more systematically and comprehensively.

The **General Evaluation Criteria** for our quality framework are the following: (i) *completeness* of the selected qualities at both data and information model levels, (ii) *necessity* of all selected quality attributes, i.e., that all are required for the chosen use case domains, (iii) *conformance* of the mapping between the layers, (iv) *feasibility* of automation and complexity of function quality calculation, and (v) *transferability* to other domains beyond IoT. The first, second and fifth criteria have already been demonstrated elsewhere (Azimi and Pahl, 2020a), where the basics of the layered quality model were introduced (here we add the close loop with the analysis and remediation part as novel elements.

The **Conformance** of the ML model with the underlying data sets is the key concern here in this investigation. This relates to a core property of ML models: accuracy, i.e., how well the model represents the underlying real truth. This is largely linked to the ML model construction through training. As said, it concerns a key property, but since it requires the consideration of concrete ML training details, this shall not be discussed here in full detail. Some general statements can, however, be made.

For a concrete application, the accuracy can be measured through precision and recall. Precision (positive predictive value) is the fraction of relevant instances among the retrieved instances. Recall (sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. They are based on true and false positives and negatives calculated by the model. The aim is perfect precision (no false positives) and perfect recall (no false negatives). This is application-specific, but the metrics for their calculation are generally agreed.

Automation applies here to the automation of the quality assessment, i.e., whether human intervention is necessary for ML model assessment and subsequent analysis and also the time aspect (whether assessment is immediately possible). This applies to (i) the initial ML model quality assessment (e.g., accuracy as described above), (ii) the mapping of model quality to data quality through probabilistic models as suggested, and (iii) root cause identification of data quality deficiencies, also using probabilistic models.

It needs to be noted that some aspects such as qualities of predictors and adaptors refer to future events (an external event will have happened for predictors or a future system adaptation will have become effective for adaptors). This still allows to make quality assessments, but just not immediately. A detailed coverage of this aspect is beyond the scope of this paper and shall be addressed at a later stage.

6 CONCLUSIONS

Raw data is without additional processing of little value. More and more, machine learning can help with this processing to create meaningful information. We developed here a quality framework that combines quality aspects of the raw source data as well as the quality of the machine-learned information models derived from the data, We provided a fine-granular model covering a range of quality concerns organised around some common types of machine learning function types.

The central contribution here is the mapping of observable ML information model deficiencies to underlying, possible hidden data quality problems. The aim was a root cause analysis for observed symptoms. Furthermore, recommending remedial actions for identified problems and causes is another part of the framework.

Some open problems for future work emerge from our discussion. The assessment of the information model requires further exploration. We provide informal definitions for all concepts, but all aspects beyond accuracy need to be fully formalised. The automation of assessment and analyses is a further concern. In the paper here, we only covered the framework from a conceptual perspective. A further part of future work is to move the framework towards digital twins. Digital twins is a concept that refers to a digital replica of physical assets such as processes, locations, systems and devices. These are often based on IoT-generated data with enhances models and function provided through machine learning. We plan to investigate deeper the complexity of these digital twins and the respective quality concerns that would apply.

REFERENCES

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and Zimmermann, T. Software engineering for machine learning: A case study. In *International Conference on Software Engineering - Software Engineering in Practice*.

Azimi, S. and Pahl, C. (2020a). A layered quality framework in machine learning driven data and information models. In *22nd International Conference on Enterprise Information Systems - ICEIS 2020*. SciTePress.

Azimi, S. and Pahl, C. (2020b). Particle swarm optimization for performance management in multi-cluster iot edge architectures. In *International Conference on Cloud Computing and Services Science CLOSER*. SciTePress.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, page 161–168.

Casado-Vara, R., de la Prieta, F., Prieto, J., and Corchado, J. M. (2018). Blockchain framework for iot data quality via edge computing. In *Proceedings of the 1st Workshop on Blockchain-Enabled Networked Sensor Systems*, page 19–24.

Ehrlinger, L., Haunschmid, V., Palazzini, D., and Lettner, C. (2019). A daql to monitor data quality in machine learning applications. In Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A. M., and Khalil, I., editors, *Database and Expert Systems Applications*, pages 227–237.

Kleiman, R. and Page, D. (2019). Aucμ: A performance metric for multi-class machine learning models. In *International Conference on Machine Learning*, pages 3439–3447.

Nguyen, T. L. (2018). A framework for five big v's of big data and organizational culture in firms. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5411–5413. IEEE.

O'Brien, T., Helfert, M., and Sukumar, A. (2013). The value of good data- a quality perspective a framework and discussion. In *ICEIS 2013 - 15th International Conference on Enterprise Information Systems*.

Pahl, C., Fronza, I., El Ioini, N., and Barzegar, H. R. (2019). A review of architectural principles and patterns for distributed mobile information systems. In *Proceedings of the 15th International Conference on Web Information Systems and Technologies*.

Pahl, C., Ioini, N. E., Helmer, S., and Lee, B. (2018). An architecture pattern for trusted orchestration in iot edge clouds. In *Third International Conference on Fog and Mobile Edge Computing (FMEC)*, pages 63–70.

Plewczynski, D., Spieser, S. A. H., and Koch, U. (2006). Assessing different classification methods for virtual screening. *Journal of Chemical Information and Modeling*, 46(3):1098–1106.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., and Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872.

Saha, B. and Srivastava, D. (2014). Data quality: The other face of big data. In *2014 IEEE 30th International Conference on Data Engineering*, pages 1294–1297.

Samir, A. and Pahl, C. (2019). A controller architecture for anomaly detection, root cause analysis and self-adaptation for cluster architectures. In *Intl Conf Adaptive and Self-Adaptive Systems and Applications*.

Sicari, S., Rizzardi, A., Miorandi, D., Cappiello, C., and Coen-Porisini, A. (2016). A secure and quality-aware prototypical architecture for the internet of things. *Information Systems*, 58:43 – 55.

Sridhar, V., Subramanian, S., Arteaga, D., Sundararaman, S., Roselli, D. S., and Talagala, N. (2018). Model governance: Reducing the anarchy of production ml. In *USENIX Annual Technical Conference*.