Data Readiness Report

Shazia Afzal, Rajmohan C, Manish Kesarwani, Sameep Mehta, Hima Patel
{shaafzal, rajmohanc1, manishkesarwani, sameepmehta, himapatel}@in.ibm.com
IBM Research, India

ABSTRACT

Data exploration and quality analysis is an important yet tedious process in the AI pipeline. Current practices of data cleaning and data readiness assessment for machine learning tasks are mostly conducted in an arbitrary manner which limits their reuse and results in loss of productivity. We introduce the concept of a Data Readiness Report as an accompanying documentation to a dataset that allows data consumers to get detailed insights into the quality of input data. Data characteristics and challenges on various quality dimensions are identified and documented keeping in mind the principles of transparency and explainability. The Data Readiness Report also serves as a record of all data assessment operations including applied transformations. This provides a detailed lineage for the purpose of data governance and management. In effect, the report captures and documents the actions taken by various personas in a data readiness and assessment workflow. Overtime this becomes a repository of best practices and can potentially drive a recommendation system for building automated data readiness workflows on the lines of AutoML [8]. We anticipate that together with the Datasheets [9], Dataset Nutrition Label [11], FactSheets [1] and Model Cards [15], the Data Readiness Report completes the AI documentation pipeline.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Software and its engineering** → **Documentation**; • **Human-centered computing** → *Walkthrough evaluations*.

KEYWORDS

machine learning datasets, data quality, data pre-processing, documentation, data governance

1 INTRODUCTION

In recent times, there has been an increased emphasis on the importance of standardized documentation practices in the machine learning community. Of particular relevance are the foundational papers Datasheets, ModelCards and FactSheets ([1, 9, 15]) that highlight these standardization efforts for various AI assets in a systematic manner. Their primary objective is to enable transparent communication and transfer of information between creators and consumers of various AI assets. These efforts are also aimed at bench-marking the key characteristics and evaluations of AI assets to empower auditing and enable informed reuse.

In this paper, we propose the Data Readiness Report as a documentation of operations conducted during the data preparation, cleaning and quality analysis phases in a typical AI life cycle. At a conceptual level it describes the characteristics of a dataset and calibrates it with respect to various quality dimensions relevant to machine learning for different stakeholders. This apprises data

practitioners about the challenges in the data upfront and can potentially reduce iterative data exploration. Here, we begin with highlighting the importance of data profiling and quality analysis as a key step before data enters a ML pipeline. We discuss the shortcomings and inefficiencies in current practices related to data pre-processing and quality assessment. Finally we illustrate the concept of Data Readiness Report as a shareable asset that accompanies the data. The report consists of baseline quality indicators as well as operations performed as part of remediation or transformations. We foresee the key benefits of the Data Readiness Report to be as follows:

- Serve as a shareable data asset giving a complete overview of data readiness for machine learning tasks
- Document the data quality issues identified as part of data quality and readiness assessment process
- Bring in standardization by expressing properties of data in terminology that is meaningful to AI
- Serve as a repository of data operations and sequence in which they are applied
- Reduce repetitive data exploration and analysis thereby shortening model development time
- Enable auditing of user interactions with data, and therefore,
- Preserve *Who* has modified, transformed, or enriched data in *What* manner at *What* stage

We suggest that the Data Readiness Report serves as the missing link between the Datasheets and Model Cards/FactSheets thereby completing the transparency and standardization efforts in the AI documentation pipeline. Throughout the paper the terms data preparation, data cleaning, data quality analysis, etc. are used interchangeably under an umbrella term of data readiness and quality assessment to indicate data operations ordinarily undertaken before data enters a machine learning pipeline.

2 MOTIVATION

As machine learning continues to be applied across various domains, the challenges related to data acquisition, its management and control, as well as ensuring its conformance to quality criteria and readiness are also getting multiplied. Data is being ingested in various formats and from diverse sources at remarkable velocity. It is also being consumed and transformed by various technical and non-technical stakeholders at different stages. In a typical AI pipeline, these multiple human actors influence and interact with data in various ways. Figure 1 illustrates the role of various personas at different stages of the pipeline. From making the data available for a target use case, to setting access and privacy constraints, checking legal compliance, cataloguing it and getting annotation and labelling along with domain specific insights, are some of the processes that precede the data being acquired by data scientists or machine leaning engineers for use. Post cleaning and
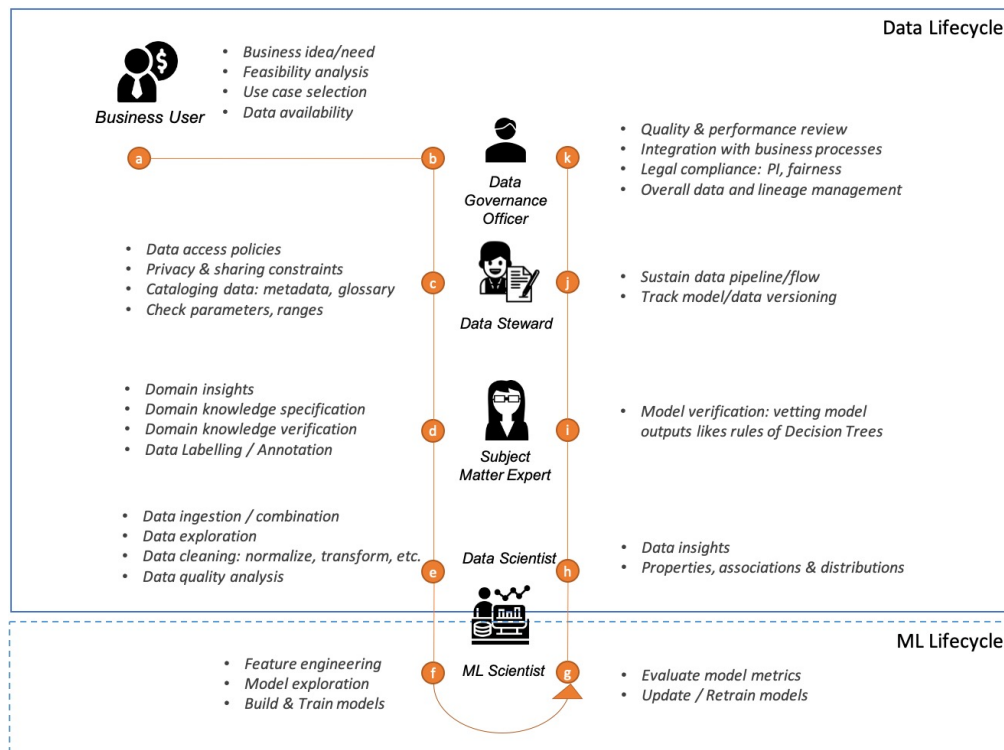
Data Lifecycle

Business User
- Business idea/need
- Feasibility analysis
- Use case selection
- Data availability

a — b — k

Data Governance Officer
- Quality & performance review
- Integration with business processes
- Legal compliance: PI, fairness
- Overall data and lineage management

- Data access policies
- Privacy & sharing constraints
- Cataloging data: metadata, glossary
- Check parameters, ranges

c — j

Data Steward
- Sustain data pipeline/flow
- Track model/data versioning

- Domain insights
- Domain knowledge specification
- Domain knowledge verification
- Data Labelling / Annotation

d — i

Subject Matter Expert
- Model verification: vetting model outputs likes rules of Decision Trees

- Data ingestion / combination
- Data exploration
- Data cleaning: normalize, transform, etc.
- Data quality analysis

e — h

Data Scientist
- Data insights
- Properties, associations & distributions

ML Lifecycle

- Feature engineering
- Model exploration
- Build & Train models

f — g

ML Scientist
- Evaluate model metrics
- Update / Retrain models

Figure 1: Various personas in an AI pipeline

pre-processing, models are built and their performance is reviewed. This sets the stage for post production downstream tasks generally centered along review, verification, versioning and integration. The primary objective of this figure is to show that the data ecosystem spans across an organization and human interaction at various levels is an essential component in this workflow. This emphasises the value and need for a transparent and efficient communication of role dependent information between personas. This would not only familiarize them with the profiling or analysis already performed on data but would also enable informed decision-making and infuse efficiency in the process. The Datasheets and Model Cards address such sharing and transparency concerns at data acquisition and model development stages, respectively. Here we focus on the data quality analysis process wherein data is subjected to various cleaning and validation checks before being considered ready for AI tasks.

The journey of data from its collection to serving as training data for an AI system mostly takes place in an adhoc manner and is barely documented. Anomalies and in-consistencies may creep into the data at any of the ingestion, aggregation or annotation stages. This impacts the quality or readiness of a dataset for building reliable and efficient machine learning models. It is estimated that data scientists spend about ~80% of their time in iterative pre-processing of data before it is considered to be fit for downstream machine learning tasks. This involves slightly less glamorous operations related to cleansing, validation and transformation of underlying data. Machine learning contexts also introduce additional quality requirements like ensuring adequate representation of target classes, collinearity checks, presence of discriminatory features, occurrence of bias, identification of mislabelled samples and various other data validation checks. These are time consuming and tedious operations but can have a cascading impact on the model design and performance. Currently there is no single integrated tooling/solution that systematically addresses these data issues in a streamlined manner. Moreover insights into the characteristics of data and its distribution, that are key to feature engineering, are also conducted in an exploratory and speculative manner. Thus, most effort in data exploration and quality assessment in the data preparation pipeline is dispersed and empirical, thereby not easily transferable across contexts. The knowledge and practice acquires a more experiential and subjective nature which can impact productivity and reuse. Data practitioners may develop affinity to frequently used methods and techniques without conscious regard to maximising value or adopting optimal workflows [17]. This together with a lack of standardized documentation makes the current process inefficient and less productive. The opacity of these operations keep the consumers of the dataset, like machine learning engineers, unaware of the pre-processing or actions already applied on the data. With no supplementary record they have no insight into the quality of incoming data and its readiness for a machine learning task and may spend numerous precious cycles exploring such properties.

Although Datasheets address the documentation of the data acquisition phase, there is an obvious gap underlining the data

Table 1: A comparison of Data Readiness Report with related work

Feature	Datasheets	DataMaid	Codebook	Dataset Nutrition Label	Mithra Label	Data Readiness Report
Basic metadata about dataset	yes	yes	yes	yes	yes	yes
Data characteristics	yes	yes	yes	yes	yes	yes
Data quality metrics	no	yes	yes	yes	yes	yes
Data ethics	yes	no	no	no	yes	yes
Task specific	no	no	no	no	yes	no
Remediations to quality issues	no	no	no	no	no	yes
Collaborative environment setting	no	no	no	no	no	yes
Lineage of operations	no	no	no	no	no	yes
Nature of output	static questionnaire	overview document	codebook document	interactive report	visual widgets	report document
Creation method	manual	automated	automated	automated	automated	automated

preparation and pre-processing phase. We assume a collaborative data quality framework that systematically addresses these issues by providing an integrated platform for data exploration and quality assessment and allows tracking of all user operations and decisions taken by various stakeholders in the pipeline, e.g [5, 17]. This not only streamlines the process and optimises productivity but also increases trust and reliability. It can enable creation and cataloguing of operational knowledge in the form of shareable artifacts that describe the data and operations conducted on it. Tasks that require collaboration between personas like a data steward and a subject matter expert can also be conducted in a streamlined manner with specific changes appropriately documented. We propose a Data Readiness Report as an artifact that serves as a certification of the baseline quality of ingested data as well as a record of operations and remediations performed on the data. Unlike Datasheets that focuses on data provenance, the aim here is to account for and record detailed statistical characteristics of data along with the operations conducted on it as part of data quality and readiness for AI modelling. The readiness report becomes a one stop lookup for understanding the quality of data and the transformations it has undergone. The following sections describe the components and use cases of such a report in the context of existing tools and practices.

3 RELATED WORK

Focus on standardized documentation to accompany AI assets is a fairly recent phenomenon. In this section we review prior art related to this theme to show how the proposed readiness report builds on and provides a holistic account of all the operations by data workers in a data quality and readiness pipeline.

One of the earliest significant works that highlighted this practice was the concept of Datasheets[9]. This was proposed as an

accompanying record for datasets to document the motivation, composition, collection process, recommended uses, dataset distribution and maintenance etc. thus encouraging transparency and accountability for data used by machine learning community. Data Statements[3] is similar to Datasheets but concentrated on linguistic datasets focusing specifically on NLP related categories.

Model Cards[15] were developed on the same principle as short documents that could accompany trained machine learning models and provide information about the models like general details, intended use cases, evaluation data and metrics, model performance measures etc. Again the motivation was to increase transparency between various stakeholders of machine learning models and systems. Similar to model cards but oriented towards AI services are FactSheets[1]. A FactSheet for AI Services contains various attributes of an AI service like intended use, performance, safety, and security etc. It tries to bridge the expertise gap between the producer and consumer of an AI service by communicating the various attributes of the AI services in a standardized way.

The Dataset Nutrition Label[11] proposes a nutrition label concept for datasets to assess the quality and robustness of a dataset before it is used for AI model development. It provides a comprehensive overview of the core components of a dataset in a standardized format through interrogation of the dataset via various flexible qualitative and quantitative modules. The prototype modules discussed are about metadata, provenance, statistics, variables, pair plots, probabilistic model etc.

MithraLabel[18] generates nutritional labels for datasets in the form of a set of task-specific flexible visual widgets. It helps determine the fitness of the dataset for specific data science tasks like classification, clustering and ranking. The widgets provide information about representation of minorities, bias and correctness

in addition to some general information about the dataset, value distributions, summary sketch, missing values, outliers, etc.

Datamaid[16] provides automated tools for data quality assessment in R. It auto-generates human-readable assessment overview documents containing data cleaning summaries. It does not provide any tools for fixing issues in the data. Codebook[2] is another R package that generates rich documentation of the dataset containing metadata, data summary statistics, distribution plots, missing values etc. It generates human and machine-readable metadata which is easy to share.

Lawrence[14] proposes the use of data readiness levels as a common language for assessing the quality and preparedness of a data set for better project management. It proposes an initial set of descriptors for data readiness in the form of three levels representing different stages of data readiness like accessibility of the dataset, faithfulness and representation of the data, appropriateness of the data set for the context. Castelijns et. al.[4] present a similar framework focused on machine learning.

The major focus of these works has been on highlighting the data characteristics and data quality issues in specific ways. They do not take into account the remediations to the identified quality issues and explanations for the same. They also do not capture the lineage of data assessment operations and the role of various personas in a collaborative data preparation environment. Table 1 encapsulates the positioning of data readiness report with respect to the previously mentioned prior art along various dimensions. The data readiness report not only provides a more comprehensive and holistic view of the data quality issues but it also encompasses other critical information from the data assessment lifecycle like remediations applied, lineage of operations and the actions of various personas during data assessment. The report generation process is largely automated except for the basic metadata module which relies on information outside the context of the data quality pipeline.

4 DATA READINESS REPORT

Figure 2 depicts how various data workers like data stewards, subject matter experts and data scientists are involved at different stages in the data ingestion and processing workflow. Their interaction with data is determined by their role in the overall pipeline of model development and each may have different function with respect to the kind of input they provide. A data steward may be concerned with bias present in the data, adherence to business constraints, acceptable parameters and ranges for attributes and also monitoring of lineage. An SMEs role could be to validate the annotation quality, identify label noise, review various recommendations and also offer remedial actions to remove the anomalies. Similarly a data scientist may explore the statistical properties of data and look at other dimensions of quality like outliers, class imbalance, missing values etc. However this involvement of various stakeholders in current practice occurs in an adhoc manner often involving trial and error based iterative workflows. This reduces the scientific rigor of this exercise and also results in loss of productivity as expert human knowledge remains tacit and non-transferable.

In order to increase efficiency and learn best practices over time for driving automation, we introduce the concept of a Data Readiness Report that enables sharing and communication of operations

that are done on data to bring it in accordance with the requirements of AI models. A landscape of 'data readiness' is generated by describing features and properties of data along with details of operations applied on it, using terminology that is meaningful to AI. The report serves to quantify the data insights and issues in a format that allows informed use of data. We conceptualise the term 'data readiness' to describe its fitness for machine learning and AI projects. As such data readiness is defined in terms of dimensions or attributes that measure or describe the characteristics of data that directly or indirectly impact model development. This includes the data operations undertaken to get data insights, explore features and their associations, measure data quality through validation checks, detect deficiencies and anomalies, and get an estimation of various other criteria that are relevant to data or ML practitioners. Moreover, the remediations and transformations done to clean and enrich the data and make it ready for AI also contribute to the data readiness framework.

The Data Readiness Report in itself serves as a documentation of all the quality and readiness analysis including the lineage of transformations applied. It calibrates the quality of a dataset against various dimensions relevant to machine learning. By certifying the dataset in terms of various quality indicators or metrics, it is aimed to be a reference point for a user who wants to understand the data readiness of a given dataset in a transparent manner. The quality analysis should be accompanied with a description of each operation and an explanation of the outputs. The report also records the recommended transformations or suggested remediations for rectifying the issues in data and improving its quality. Since it records the lineage of transformations or operations conducted as part of the remediation, a repository of best practices is learned over time that can be used for improvising recommendations and optimizing data workflows. This can further help in building scalable and principled data assessment solutions by finding the right balance between automation and human input with the latter used strategically for reviewing, verification or approval of automatic decisions.

Figure 2 also shows some examples in which the report can benefit various personas discussed earlier. For instance, a data scientist A can quickly understand the challenges and quality issues in the data from the report without having to repeat similar analysis conducted by say data scientist B. A subject matter expert can review the remediations done on data and learn how data issues affect quality scores. Similarly, data stewards may be interested in tracking user interactions with data and enabling reliable data versioning based on lineage records. Non technical users like data governance officers or business users may be more interested in overview information like data profile and quality indicators amongst others. The figure shows an illustration of the utility of information provided by the Data Readiness Report for different stakeholders. In practice, users may require and consume information from the report depending on their specific task and requirements.

In the remainder of this section, we describe the composition of the Data Readiness Report in the context of a data quality assessment framework. These components represent information that we believe can be helpful to data practitioners. However this is certainly not exhaustive as various personas in a federated enterprise may require a different view or level of detail.
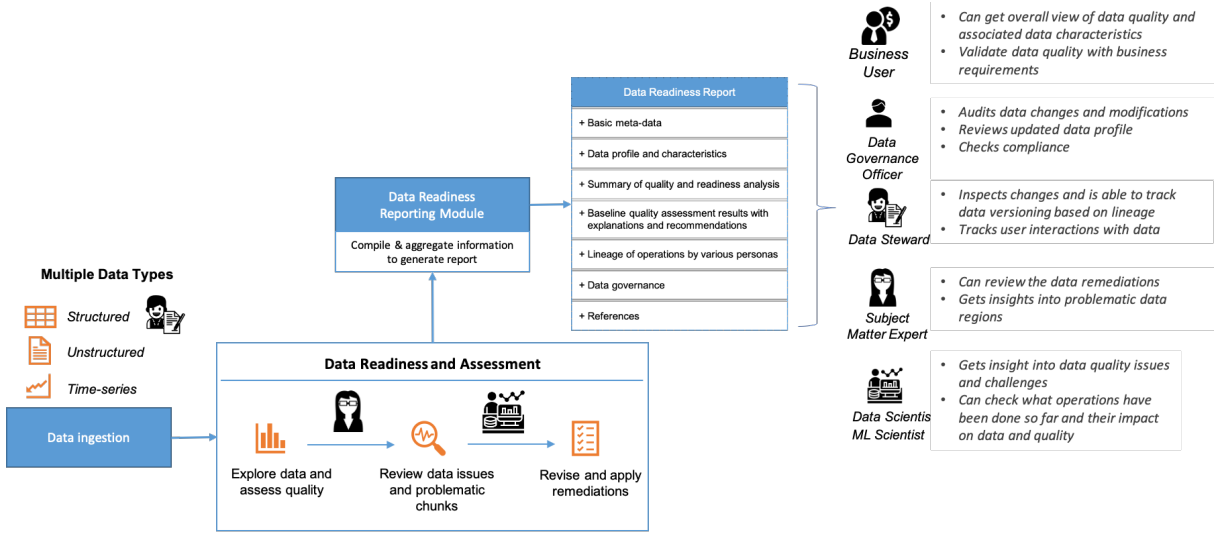
Figure 2: Generation of Data Readiness Report and its Utility Across Personas

4.1 Basic Metadata

The introductory section of the report should describe details of the dataset to set the context for the remaining sections. It should also provide accessible links to the original dataset for reference along with standard usage information. It may include the following:

- basic details of the dataset like name, version, generation date, etc.
- plain text description of what it contains
- type of the data: structured, unstructured, time-series, etc
- intended usage
- source of the dataset and any specific restrictions on usage

Note that this section may have some overlap with a datasheet if available. However, in absence of a detailed data description as required in a datasheet, this first section can serve as an effective summary of key information. Since the information contained in this section is likely to be external to the data pipeline, filling this may require manual effort. The purpose of having this section is to give some background of input data for which a report is generated. This is not anticipated to be an extensive exercise but more of a concise lookup of key facts about the dataset.

4.2 Summary of Quality and Remediation Assessment

As the name suggests, the primary goal of this section is to give an overview of the quality and readiness assessment performed on the data. As the number of operations and transformations performed on the data become quite large over time, this summary view provides a quick snapshot of the data profile and quality profile of latest data in comparison to the baseline data. Using a visual info-graphic for illustrating this is a good way to make the review process easier and attractive for consumers of the report.

4.3 Baseline Data Profile

This section in the report captures basic characteristics and statistical properties of input data. To simplify the presentation, here we assume that the data is of structured type but in general, the data could be structured, unstructured or time-series and the properties and characteristics would be defined relative to the dataset type. For example, in structured data, a data profile section would include aspects like number of rows and columns, description of each column including datatype, minimum value, maximum value, missing data percentage, number of unique values and other statistical measures. Further for each column, the report can also include details of identified patterns, value distributions and data constraints. The purpose of this section is to provide a detailed profile of the original data by exhibiting all the general traits of the data. It also serves as a benchmark to compare against any updates made on the data during the process of quality assessment and associated transformations.

4.4 Baseline Data Quality and Readiness Analysis

This is the core section of the report that provides details of the data quality and readiness analysis. This would include the various techniques used to assess the data against dimensions or attributes relevant to a machine learning task like noise, missing values, outliers, high dimensionality, imbalanced class, inconsistency, redundancy, etc. [5]. For example, it may include details of how a dataset fares across certain pre-defined quality metrics known to influence model building efforts [12]. The specific format of visualizing this information may vary depending on output requirements and constraints. But key information like the methods used, parameters or configurations applied, details of results including appropriate explanations and suggested actions if any, should be documented. We strongly recommend the use of proper explanations in describing the results of various data operations to enable diverse consumers

Data Readiness Report

- **Basic Metadata:** Basic information about data
  - Data Owner
  - Version
  - Generation Date
  - Type: structured, unstructured or semi-structured
  - Description
  - Tags
  - Intended Usage
  - Contact Person
- **Summary of Quality and Readiness Assessment:** Overview of original against final data quality
  - Original and updated data profile
  - Original and updated quality profile
- **Baseline Data Profile:** Characteristics of original data
  - Number of rows and columns
  - Basic properties of each column including datatype, description, min value, max value, median, standard deviation, missing data percentage, number of unique values etc
  - Column data constraints
  - Value distributions
  - Pairwise column correlations
- **Baseline Quality and Readiness Assessment:** Evaluation of data on various quality dimensions along with explanations and recommended remedial actions. e.g. Label Noise, Class Imbalance, Correlation, Data Homogeneity, Outliers, Missing Values, Inconsistent values, Data Bias, amongst others.
- **Updated Data Profile:** Characteristics of data after applying remediations suggested during the quality assessment and readiness analysis phase.
- **Updated Quality and Readiness Assessment:** Quality checks on updated data after application of remediations.
- **Lineage of Operations:** Record of data operations by various personas
  - Methods used with input and output parameters
  - Detailed Results
  - Explanation of results
  - Recommendations or suggested actions
  - SME inputs and remediations applied
  - Changes to dataset
- **Data Governance:** Set of rules and policies for this data
  - Source of Data
  - Usage Restrictions
  - Policy Restrictions
  - License
- **References**
  - Details of metrics
  - Overview of remediations
  - Cite implementation details

Figure 3: Data Readiness Report Template

of the report to understand the issues or challenges in input data. This would improve transparency and ensure that readers of the report make meaningful inferences about the quality checks performed on data. Having appropriate recommendations to improve quality scores wherever possible is strongly advised so that users are not just left with a static analysis but are also made aware of possible ways to fix the identified quality issues.

4.5 Updated Data Profile

The raw data profile may undergo changes once remediations or transformations are done to improve the data quality or remove anomalies. For example, to rectify class imbalance sampling strategies may be used to make the data more balanced which may result in an increase or decrease in the number of rows depending on the resampling method used [13]. Similarly modifications done to resolve column heterogeneity may change the value distributions of target columns [6, 10]. All such changes ought to be documented and this section serves the purpose of providing the relevant updates to the data profile.

4.6 Updated Quality and Readiness Analysis

In a typical quality analysis scenario, after checking the data quality on raw data, many transformations or remediations are applied to improve the data readiness. This is typically validated by re-running the quality analysis on the updated dataset and ensuring that modifications done to data have in fact been effective. This should therefore be reflected in appropriate documentation to easily compare the impact and value of performing such operations. The expectation is that the quality and readiness metrics on the updated data would be superior compared to the baseline assessments.

4.7 Lineage of Operations

In a data readiness assessment pipeline, various personas interact with data in different ways for a variety of reasons. This section will provide a detailed documentation of all the data transformations and operations performed by them in chronological order along with timestamps and other relevant details. This digital trail provides a comprehensive record of how the data has evolved due to interactions with humans in the loop. Having a complete record of all the data quality assessments and applied remediations across roles equips data governance process greatly. With quality control across versions, a data scientist can pick any version of the dataset that best suits them based on the quality scores and the task at hand. It also helps in building a repository of best practices for reuse. For example, comparison of reports can serve as a good review of how the data operations affect data quality and drive automation efforts to recommend optimal sequence of transformations and operations.

4.8 Data Governance

Every dataset may be accompanied by a unique set of policy and legal constraints which might restrict its accessibility in different operating environments. This section in the readiness report will highlight any such data usage guidelines. For instance if the data is generated in EU and EEA areas, then GDPR guidelines must be followed. In some cases, data owner might outline custom usage policies which must be adhered to.

4.9 References

This final section catalogs the quality metrics or dimensions used in the report along with appropriate references to give details of methods and algorithms used, parameters and configurations applied. The idea behind this is to improve the scientific rigour so that methods and techniques documented in the report can be effectively re-implemented.

5 EXAMPLES

Figures 4 and 5 exhibit sample Data Readiness Reports on the Breast Cancer dataset and Adult dataset respectively from the UCI data repository [7]. These mainly contain the key sections covered in the data readiness report template described in Figure 3. We made use of some of the machine learning related quality metrics mentioned in [12] for illustration. Due to space constraints, we include only limited features just to exemplify how key information from the quality analysis process can be represented in the report.

6 SUMMARY

Data that enters an ML pipeline is subjected to pre-processing by various stakeholders in their own distinctive manner using tools and methods acquired overtime through experience, habit or preference. This adhoc and iterative nature of work limits reuse and results in loss of productivity. Data practitioners also spend a significant percentage of their time in exploring and tackling various data quality issues. This is because they have very limited knowledge about the challenges present in incoming data and whether any modifications or changes have been done to it, and if so, by whom. We introduce the concept of a data readiness report as an accompanying documentation to a data quality and readiness assessment framework that allows data consumers to get detailed data insights into the quality of input data across various standardized dimensions. It serves as a comprehensive documentation of all data properties and quality issues including data operations by various personas to give a detailed record of how data has evolved. Together with the Datasheets and Model Cards this completes the AI pipeline with respect to key factors of transparency, documentation and governance.

REFERENCES

[1] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, A Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[2] Ruben C Arslan. 2019. How to automatically document data with the codebook package to facilitate data reuse. *Advances in Methods and Practices in Psychological Science* 2, 2 (2019), 169–187.

[3] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[4] Laurens A Castelijns, Yuri Maas, and Joaquin Vanschoren. 2019. The ABC of Data: A Classifying Framework for Data Readiness. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 3–16.

[5] David Camilo Corrales, Agapito Ledezma, and Juan Carlos Corrales. 2018. From theory to practice: A data quality framework for classification tasks. *Symmetry* 10, 7 (2018), 248.

[6] Bing Tian Dai, Nick Koudas, Beng Chin Ooi, Divesh Srivastava, and Suresh Venkatasubramanian. 2006. Rapid identification of column heterogeneity. In *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, 159–170.

[7] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[8] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and robust automated machine learning. In *Advances in neural information processing systems*. 2962–2970.

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[10] Sumit Gulwani. 2011. Automating string processing in spreadsheets using input-output examples. *ACM Sigplan Notices* 46, 1 (2011), 317–330.

[11] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677* (2018).

[12] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. 2020. Overview and Importance of Data Quality for Machine Learning Tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3561–3562.

[13] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30, 1 (2006), 25–36.

[14] Neil D Lawrence. 2017. Data readiness levels. *arXiv preprint arXiv:1705.02245* (2017).

[15] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[16] Anne H Petersen and Claus Thorn Ekstrøm. [n.d.]. dataMaid: your assistant for data cleaning in R. *Journal of Statistical Software* ([n. d.]), 35.

[17] El Kindi Rezig, Mourad Ouzzani, Ahmed K Elmagarmid, Walid G Aref, and Michael Stonebraker. 2019. Towards an End-to-End Human-Centric Data Cleaning Framework. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. 1–7.

[18] Chenkai Sun, Abolfazl Asudeh, HV Jagadish, Bill Howe, and Julia Stoyanovich. 2019. Mithralabel: Flexible dataset nutritional labels for responsible data science. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2893–2896.

Data Readiness Report

Basic Metadata

- **Name** : Breast Cancer Data Set
- **Owner**: Matjaz Zwitter & Milan Soklic
- **Description**: The data contains details of 286 patients with known diagnostic status 5 years after the operation
- **Intended Usage**: Prediction task for prognosis of Breast Cancer Recurrence
- **Source**: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Summary of Data Quality Analysis

(Bar chart: initial_score, updated_score)
- Feature Relevance
- Correlation Detection
- Data Completeness
- Outlier Detection
- Data Homogeneity
- Class Overlap
- Label Noise

Data Profile — Individual Column Profiles

Class | Age | Menopause | Tumor-Size

Baseline Quality and Readiness Assessment

Quality Metric	Description	Explanation	Recommendations
Label Noise (0.92)	Identifies noise ratio and noisy samples in the data. A score of 1 indicates no noise is present in the data.	Noisy labels detected in 23 / 286 rows giving a quality of 0.92.	Apply label noise remediation to the result.
Correlation Detection (1)	Identifies correlated numerical columns in the data. A score of 1 indicates no correlated columns found in the data.	No correlated column pairs found in the data.	No action required
Outlier Detection (0.99)	Identifies outlier samples in the data. A score of 1 indicates no outliers found in the data.	Outliers detected in 2 / 286 rows giving a quality score of 0.99.	Remove Outliers

Applied Recommendations and Updated Quality

- Label Noise — Remediation Applied — 0.98
- Correlation Detection — No action suggested — 1
- Outlier Detection — Remediation Applied — 1

Data Governance

Please refer to the Machine Learning Repository's citation policy:
https://archive.ics.uci.edu/ml/citation_policy.html

References

Method and implementation details of metrics used along with references

Note: Due to space constraints, this example report shows only some representative sections from the proposed template to illustrate the value and nature of key information affiliated to quality and readiness assessment.

Lineage Of Operations

User A
User B
User C

Data version 1 — Quality Score : 0.92 — 23 samples modified — Quality Score : 0.98

Data Profile
- Done By: User A
- Timestamp: 2020-10-06 14:39:17

Label Noise
- Done By: User C
- Timestamp: 2020-10-06 14:40:17

Label Noise Remediation
- Done By: User B
- Timestamp: 2020-10-06 14:41:40

Label Noise
- Done By: User C
- Timestamp: 2020-10-06 14:41:17

Data version 2 — Quality Score : 0.99 — 2 samples identified as outliers are dropped — Quality Score : 1.0

Outlier Detection
- Done By: User A
- Timestamp: 2020-10-06 14:42:27

Outlier Removal
- Done By: User B
- Timestamp: 2020-10-06 14:45:40

Outlier Detection
- Done By: User A
- Timestamp: 2020-10-06 14:46:17

... other operations

Figure 4: A Sample Data Readiness Report – Breast Cancer Dataset

Data Readiness Report

Basic Metadata

- **Name** : Adult Data Set
- **Owner**: Ronny Kohavi and Barry Becker
- **Description**: Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))
- **Intended Usage**: Prediction task is to determine whether a person makes over 50K a year
- **Source**: https://archive.ics.uci.edu/ml/datasets/Adult

Summary of Data Quality Analysis

Legend: initial_score, updated_score

- Feature Relevance
- Correlation Detection
- Data Completeness
- Outlier Detection
- Data Homogeneity
- Class Overlap
- Label Noise

(x-axis: 0.0, 0.2, 0.4, 0.6, 0.8, 1.0)

Data Profile   Individual Column Profiles

Age: (82,90), (75,82), (68,75), (60,68), (53,60), (46,53), (38,46), (31,38), (24,31), (17,24) — (0, 2,000, 4,000, 6,000)

Education: Prof-School, 7Th-8Th, 10Th, Assoc-Acdm, 11Th, Assoc-Voc, Masters, Bachelors, Some-College, Hs-Grad — (0, 2,000, 4,000, 6,000, 8,000, 10,000)

Relationship: Other-Relative, Wife, Unmarried, Own-Child, Not-In-Family, Husband — (0, 10,000)

Sex: Female, Male — (0, 5,000, 10,000, 15,000, 20,000)

Baseline Quality and Readiness Assessment

Quality Metric	Description	Explanation	Recommendations
Feature Relevance (15% / 85%, 0.85)	Identifies and ranks the feature based on Relevance. A score of 1 indicates that all features are relevant.	Less relevant features found in the dataset are 2 / 14 giving a quality score of 0.85	Drop less relevant features from the data.
Data Completeness (100%, 1)	Identifies missing values in the given data. A score of 1 indicates no missing values found in the data.	Missing values detected in 0 / 488415 entries giving a quality score of 1.0.	No action required
Outlier Detection (11% / 89%, 0.89)	Identifies outlier samples in the data. A score of 1 indicates no outliers found in the data.	Outliers detected in 3262 / 32561 rows giving a quality score of 0.89.	Remove Outliers

Applied Recommendations and Updated Quality

- Feature Relevance — Remediation Applied — 15% / 85%, 1
- Data Completeness — No action suggested — 100%, 1
- Outlier Detection — Remediation Applied — 4% / 1% , Raw Data 0.89, 0.94

Data Governance

Please refer to the Machine Learning Repository's citation policy:
https://archive.ics.uci.edu/ml/citation_policy.html

References

Method and implementation details of metrics used along with references

Note: Due to space constraints, this example report shows only some representative sections from the proposed template to illustrate the value and nature of key information affiliated to quality and readiness assessment.

Lineage Of Operations

User A
User B
User C

Data Profile
— Done By: User A
— Timestamp: 2020-10-03 12:39:17

Data version 1

Feature Relevance
— Done By: User A
— Timestamp: 2020-10-04 17:23:17
Quality Score : 0.85

2 columns identified as less relevant features are dropped

Feature Relevance Remediation
— Done By: User B
— Timestamp: 2020-10-04 17:36:23

Feature Relevance
— Done By: User A
— Timestamp: 2020-10-04 17:37:17
Quality Score : 1.0

Data version 2

Outlier Detection
— Done By: User C
— Timestamp: 2020-10-05 18:40:17
Quality Score : 0.89

3262 outlier samples dropped

Outlier Removal
— Done By: User B
— Timestamp: 2020-10-05 18:45:40

Outlier Detection
— Done By: User C
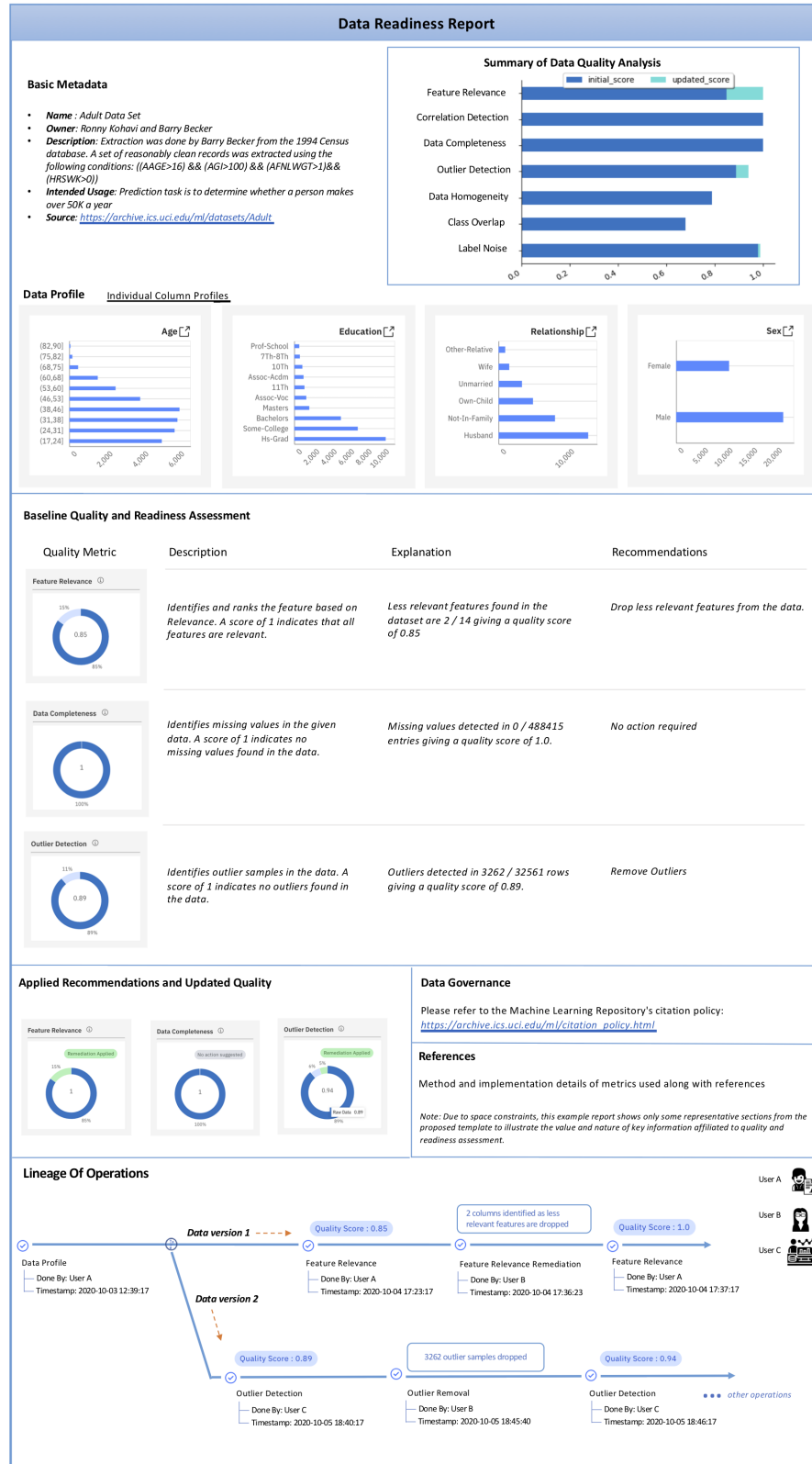— Timestamp: 2020-10-05 18:46:17
Quality Score : 0.94

••• other operations

Figure 5: A Sample Data Readiness Report – Adult Dataset