Quality issues in MLSSs form

In this survey, we explore the opinions of researchers/practitioners about quality issues in ML Software Systems. You will be presented with a series of quality issues and will be asked to rate on a scale from 1 to 5 how common the issues are in your experience. **If you feel you do not have the experience to answer a question, you may easily leave the answer field empty**. The questionnaire is composed of 6 sections, each one corresponding to a quality aspect. You can expect 3-4 questions per section.

1. What is your job role?

   Mark only one oval.

   - ◯ Data Scientist
   - ◯ ML Engineer
   - ◯ Data Engineer
   - ◯ Project Manager
   - ◯ Manager (e.g. Director)
   - ◯ Other: _____

2. How many years of professional experience do you have?

   Mark only one oval.

   - ◯ 0-2
   - ◯ 3-5
   - ◯ 6-9
   - ◯ 10+

3. How many years of experience do you have **with ML**?

   Mark only one oval.

   - 0-2
   - 3-5
   - 6-9
   - 10+

4. If you are interested in compensation (i.e. the 25$ gift card), please provide your email. We will contact you only if you win the card.

Evaluability

Evaluability is the ability to evaluate the quality of ML models and datasets.

5. **Issue**: Evaluating the quality of a model offline (i.e. not in a production environment) is inaccurate even when the dataset used for evaluation is representative of the data distribution in production. A model is of quality if it answers the needs of the application.
   Potential reasons: (1) Evaluating a model on a test dataset using a metric such as accuracy or F1-score does not take into consideration the system in which the model is embedded, and, as a result, might evaluate models incorrectly. (2) ML performance metrics (e.g. accuracy, F1 score) do not necessarily reflect how effective the model is for the end application, because the goal that is targeted with an ML metric (e.g. accuracy) is different from what is important for the end application.

   In your experience, how often have you encountered that issue?

   Mark only one oval.

	Never	1	2	3	4	5	Frequently
	◯	◯	◯	◯	◯	◯	

6. **Issue**: Defining a good business metric for evaluating an MLSS is difficult. For MLSSs, a business metric refers to the degree to which the MLSS successfully achieves the goal it has been built for.
   Example: How do we define what is the best recommendation (for a recommendation engine)?

   In your experience, how often have you encountered that issue?

   Mark only one oval.

   never
   1
   2
   3
   4
   5
   frequently

7. **Issue**: Trying to simulate the environment/system in which the model will operate (to evaluate the model in this simulated environment) is difficult and error-prone.
Example: When A/B testing can not be used to evaluate a model, a person might want to test its model in a simulated environment. Reproducing the production environment is difficult.

   In your experience, how often have you encountered that issue?

   Mark only one oval.

   never
   1
   2
   3
   4
   5
   frequently

8. **Issue**: Evaluating the quality of a dataset (e.g. presence of incorrect labels, noisy/incorrect features, wrong format data, etc.) is difficult and time-consuming.

   In your experience, how often have you encountered that issue?

   Mark only one oval.

   never
   1
   2
   3
   4
   5
   frequently

9. Do you have any comments on evaluability?

Explainability

Explainability refers to any technique that tries to explain the decisions (e.g. predictions) of a model.

10. **Issue**: Explaining a model's predictions to people without ML knowledge (e.g. business stakeholders, users) using explainability techniques is challenging because these techniques require technical knowledge and interpretation.

    Potential consequence(s): ML Explainability techniques can not directly be used to explain the predictions of a model to users or stakeholders without ML knowledge.

    In your experience, how often have you encountered that issue?

    Mark only one oval.

    never

    1 ◯

    2 ◯

    3 ◯

    4 ◯

    5 ◯

    frequently

11. **Issue**: The explanation techniques sometimes provide explanations that do not make sense and can not be relied on.
Potential consequence(s): ML Explainability techniques can not be trusted.

In your experience, how often have you encountered that issue?

Mark only one oval.

never
1
2
3
4
5
frequently

12. Do you have any comments on explainability?

Debuggability

Debuggability refers to the degree to which a system can be easily debugged.

13. **Issue**: Reproducing bugs in an MLSS is difficult because of unstable data sources. A data source is unstable if it returns different values for the same queried record.
Example: A feature store could generate embeddings daily, and replace the old version with the new one. Accurately reproducing a day-old bug would be impossible, since the embedding that triggered a bug would not be available anymore.
Potential consequence(s): Some bugs are ignored.

In your experience, how often have you encountered that issue?

Mark only one oval.

	never					frequently
	1	2	3	4	5	

14. **Issue**: Debugging data streaming systems (e.g. Hadoop) is difficult because it is challenging to picture what data should look like at each step of the data pipeline.

In your experience, how often have you encountered that issue?

Mark only one oval.

	never	1	2	3	4	5	frequently
		◯	◯	◯	◯	◯	

15. **Issue**: Debugging an MLSS is time-consuming when its data sources are managed by other teams, because it may require inspecting these data sources.
Example: An MLSS that consumes data from 5 different data sources to make a prediction. Any issue with one of the 5 data sources might hinder a model's performance. Thus, debugging might require inspecting the integrity of all the data sources (in the worst case). Since the data sources may be managed by other teams, this can become a time-consuming process.

In your experience, how often have you encountered that issue?

Mark only one oval.

	never				frequently
	1	2	3	4	5

16. Do you have any comments on debuggability?

Efficiency

Efficiency refers to the performance relative to the amount of resources used.

17. **Issue**: Training models consume a lot of resources (e.g. time, computing power, etc.) and it is an issue.
Potential consequence(s): Slow development time and hefty spending.

In your experience, how often have you encountered that issue?

Mark only one oval.

	never				frequently
	1	2	3	4	5

18. **Issue**: The queries sent to an MLSS are not answered timely (i.e. latency/delay issues).

In your experience, how often have you encountered that issue?

Mark only one oval.

never	
1	◯
2	◯
3	◯
4	◯
5	◯
frequently	

19. **Issue**: At inference time, ML models consume too much memory.
Example: A model that must run on a device with limited resources (e.g. a cell phone).

In your experience, how often have you encountered that issue?

Mark only one oval.

	never				frequently
	1	2	3	4	5

20. Do you have any comments on efficiency?

Maintainability

Maintainability refers to the degree of effectiveness and efficiency with which a product or system can be modified for the maintenance purposes (updates, fixes, …).

21. **Issue**: Maintaining an MLSS is difficult because there is not enough information describing the data the MLSS consumes.
Example: A dataset that does not explain what a feature represents (i.e. obscure column name with no description), what are the units of a feature, or what is the meaning of special tokens.

In your experience, how often have you encountered that issue?

Mark only one oval.

never	
1	◯
2	◯
3	◯
4	◯
5	◯
frequently	

22. **Issue**: Maintaining a model is difficult, because there is not enough information describing how the model was generated.

In your experience, how often have you encountered that issue?

Mark only one oval.

	never					frequently
	1	2	3	4	5	

23. **Issue**: Managing the dependencies (i.e. software libraries) of an MLSS is challenging and error-prone.
Example: The production environment of an MLSS is broken because a package manager automatically updates library versions, which broke the environment.
Potential consequence(s): Breaking the production environment of an MLSS.

In your experience, how often have you encountered that issue?

Mark only one oval.

	never
1	◯
2	◯
3	◯
4	◯
5	◯
	frequently

24. Do you have any comments on maintainability?

Reliability

Reliability refers to the degree to which a system, product, or component performs specific functions failure-free under specified conditions for a specified period of time.

This is the last section.

25. **Issue**: Having a reliable model is difficult because of concept or data drift.
Example: When the Covid epidemic happened, some models became a lot less accurate and even useless.

In your experience, how often have you encountered that issue?

Mark only one oval.

never	
1	◯
2	◯
3	◯
4	◯
5	◯
frequently	

26. **Issue**: Having a reliable model is difficult because of external data providers. An external data provider is any external entity (i.e. not the team that manages the MLSS) that provides data for an MLSS. The external data providers are unstable if they (1) are unavailable at times or (2) serve inconsistent information (i.e. the same record fetched at different times has different values).
Example: A weather service that corrects past data because of a faulty sensor. As a result, a model that consumes that data delivers poor predictions because the data distribution has changed.

In your experience, how often have you encountered that issue?

Mark only one oval.

never	
1	◯
2	◯
3	◯
4	◯
5	◯
frequently	

27. **Issue**: Having a reliable MLSS is difficult because of the data pipelines which are brittle and have technical debt.

In your experience, how often have you encountered that issue?

Mark only one oval.

	never	1	2	3	4	5	frequently
	◯	◯	◯	◯	◯	◯	

28. Do you have any comments on reliability?

Final comments

29. Do you have any other comments?