How Teams Communicate about the Quality of ML Models: A Case Study at an International Technology Company

JUMANA ALMAHMOUD, MIT CSAIL, USA

ROBERT DELINE, Microsoft Research, USA

STEVEN M. DRUCKER, Microsoft Research, USA

Machine learning (ML) has become a crucial component in software products, either as part of the user experience or used internally by software teams. Prior studies have explored how ML is affecting development team roles beyond data scientists, including user experience designers, program managers, developers and operations engineers. However, there has been little investigation of how team members in different roles on the team communicate about ML, in particular about the quality of models. We use the general term *quality* to look beyond technical issues of model evaluation, such as accuracy and overfitting, to any issue affecting whether a model is suitable for use, including ethical, engineering, operations, and legal considerations. What challenges do teams face in discussing the quality of ML models? What work practices mitigate those challenges? To address these questions, we conducted a mixed-methods study at a large software company, first interviewing 15 employees in a variety of roles, then surveying 168 employees to broaden our understanding. We found several challenges, including a mismatch between user-focused and model-focused notions of performance, misunderstandings about the capabilities and limitations of evolving ML technology, and difficulties in understanding concerns beyond one's own role. We found several mitigation strategies, including the use of demos during discussions to keep the team customer-focused.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: quality of ML models quality, communication, teams, communicating ML models, presentation of work, machine learning

ACM Reference Format:

Jumana Almahmoud, Robert DeLine, and Steven M. Drucker. 2021. How Teams Communicate about the Quality of ML Models: A Case Study at an International Technology Company. *Proc. ACM Hum.-Comput. Interact.* 5, GROUP, Article 222 (July 2021), 24 pages. <https://doi.org/10.1145/3463934>

1 INTRODUCTION

In the last decade, machine learning (ML) has become a central component of software development. For some software, ML enables key aspects of the user experience (UX), from recommendation systems to image manipulation, to voice recognition, and natural language dialogues. For other software, the ML works behind the scenes, for example, detecting fraud or predicting faults in the data center. Software teams also build ML models to support decision making on issues such as prioritizing work, understanding customers, or choosing price models. Hence, the success of software projects increasingly depends on creating high-quality ML models.

Authors' addresses: Jumana Almahmoud, jumanam@mit.edu, MIT CSAIL, Cambridge, MA, USA; Robert DeLine, Microsoft Research, Redmond, WA, USA, rdeline@microsoft.com; Steven M. Drucker, Microsoft Research, Redmond, WA, USA, sdrucker@microsoft.com.

This work is licensed under a Creative Commons Attribution-NonCommercial International 4.0 License.

© 2021 Copyright held by the owner/author(s).
2573-0142/2021/4-ART222. <https://doi.org/10.1145/3463934>

This emergence of ML also means that software teams increasingly need to communicate about ML models and their quality. Here, we use the general term *quality* to encompass not only technical issues about ML model evaluation such as accuracy, but also any issue affecting whether a model is suitable for use, including ethical, engineering, operations, and legal considerations. Because ML is involved in many aspects of software development—from UX, to engineering and operations, to management—this communication spans many roles on the team [7, 37]. Prior research explored how data scientists ensure high confidence in their analysis results [24, 36, 37, 57]. However, concerns about ML quality are not limited to issues related to the role of data scientists. Other roles on the team, including UX designers, program managers (PMs), developers, operations engineers, and product managers, also need to communicate about ML models and their quality, to support coordination and decision making. What are the challenges software teams face when communicating about ML models and their quality? What practices and tools can be introduced to mitigate these challenges?

To understand this issue, we conducted a case study in a large software company that deploys ML in a wide variety of its products. To follow a holistic approach in studying the communication process, we utilized Lasswell’s model of communication to design our interviews and survey. The model deconstructs an act of communication into component parts: the speaker, the content, the channel/medium, the audience, and the effect [25]. In our mixed-methods approach, we first conducted a series of 15 one-hour semi-structured interviews with employees in a variety of roles. We then surveyed 168 employees from a variety of roles to better understand current practices and pain points. Throughout the paper, we use the term “stakeholders” to refer to internal employees who are part of the same software organization working on the same ML-based product but are not the model developers such as UX, legal, sales, etc.

2 RELATED WORK

In this section, we look into previous work that explored aspects related to our contribution. We look into the communication around ML models from different perspectives, specifically considering how the “quality” of models is presented and perceived, and we explore the literature that tackles those aspects. We first look at communication and uncertainty; we then review collaboration in ML teams. Finally, we list documentation strategies used to facilitate the collaborative process around ML models and how algorithmic biases are captured.

2.1 Communicating Uncertainty and Error

Harold D. Lasswell [25] described the act of communication as the answers to the following questions: “*Who says What, in Which channel, to Whom, with What effect?*”. With that in mind, we wanted to explore how the quality of ML models is being communicated within teams in a large software company. Existing literature points out how certain roles in a software team and certain workflows are being affected by the introduction of ML [7, 36, 37]. While the previously referenced works mention communication as a challenge, to our knowledge, none of these existing studies investigate the challenges in communicating the quality of ML models.

According to Box, “*all models are wrong, but some are useful*” [14]. With that in mind, how do employees assess models to be of “good enough” quality? Models produce output that is uncertain by nature and has some level of error that needs to be reported in order for users to make informed decisions [21, 34]. Prior work looked into how to communicate the uncertainty in data through visualization and how it is perceived by users [6, 9, 32, 33, 50]. Other researchers surveyed how users deal with uncertainty in general [13, 31].

Another way to communicate the quality of ML models is to show errors in the output. [17] Researchers have built tools to explore errors in models results [8, 17, 49, 55]. Although these tools

were not built specifically with communication in mind, they enable data scientists to understand their models and problem areas. In addition, these tools can be used to generate visualizations that show localized error and communicate the quality of model in a more detailed way. In our work, we explore what methods are actually used in the organization to communicate the quality of ML models to a wide audience within the organization.

2.2 ML and Collaboration

Prior work contributed to our understanding of the collaboration process between data scientists and UX designers [19, 35, 39, 61], software engineers (SEs) [7, 37, 53], and people in business and other fields [38, 52, 62]. However, there has been little investigation of how members in different roles on the team communicate about ML, in particular about the quality of models. Zhang et al. [63] considered an inside view of collaboration focusing on data science workers. In this paper, we make a deeper examination by focusing on the communication process and the perceived notion of model “quality” within teams working on ML-based products. The teams are large and expand to sub teams that are beyond ML developers and include more set of roles and relations(SEs, UX, legal, business, marketing, etc.)

Existing model development tools mostly serve model developers by exposing the metrics they care about to assess the quality of their models [18, 41, 49, 55]. What about metrics that are important to other stakeholders who are likely to be different from the model builders, such as the person in charge of making the decision to release? Some prior work such as [43, 56, 58], looked into how tools could be used to allow a more effective collaboration process. However, that collaboration is only specific to two given roles, such as domain experts and data scientists, and is not generalized to collaboration among a diverse set of roles.

2.3 Documentation and Standardization

Studies have shown that considering only model performance and benchmarks is not enough to assess how well these models work [2, 4]. Proper documentation is essential [23], and many approaches for model documentation have been proposed recently, to be used as checklists to ensure the quality of models and increase users trust in AI. For example, Model Cards[45] and Factsheets [10, 20] were proposed as a general documentation approach for models and their data. . Other work argued that making the correct decision in choosing which is the right model relies on choosing the right dataset first. Therefore, Holand et al. proposed templates for documenting data. For example, Nutrition Labels [28] and Datasheets [22] serve as checklists to mitigate the risk of applying the wrong model on these datasets and assess the decision in which model best fits. However, one common limitation to these approaches and to documentation in general, is how to incentivize users to populate and use these documentations.

In addition to documenting models, model testing is essential in providing information to help teams assess the quality of models in the wild. For example, Breck et al. [15] suggested a “rubric” for scoring the readiness of an ML model based on its performance in the testing phase. Another example comes from the work done by Raji et al. [54], in which they proposed a framework for auditing even before deployment and testing to ensure the quality of the model and mitigate risks.

2.4 Algorithmic Bias

One important aspect of quality in ML systems is the issue of biases in data and algorithms [12, 40, 48]. As machine learning is becoming an integral part of our lives, researchers have been investigating the biases that could arise and how to mitigate them [5, 44]. Work on bias mitigation looked into the interpretability and transparency of these models [27, 59, 60], what industry practitioners need to improve the fairness in ML systems [29, 30], and the perceived fairness of

biased algorithms in current practices [26, 51]. In our work, we looked into how much bias and fairness in algorithms is communicated as an aspect of ML models' quality and who within teams and organizations is interested in this.

3 METHOD

Using a mixed-methods approach, we first conducted semi-structured interviews with employees in a variety of roles, drawn from product teams for whom ML is an important aspect of the product. The flexibility of the interview format allowed us both to elicit an initial understanding of their communication practices and challenges, as well to learn the vocabulary the teams use to talk about these issues. To generalize our understanding, we deployed a survey across the company's product teams. Both the interviews and survey were inspired by Lasswell's communication model, which describes acts of communication around five key questions: *Who is the communicator? What is the content? Which channel/medium is used? Who is the receiver? What is the effect?* While in both the interviews and the survey we asked general questions about how the team uses ML in its product, we focused on communication about the quality of the models. We intentionally used a broad term like *quality*, which encompasses any aspect of the model that determines its suitability for use, to avoid biasing our participants toward any particular consideration. While the academic ML community typically focuses on agreed-upon measures (e.g., accuracy) and methods (e.g., train/test splits) for evaluating models, deploying models in a product often involves other ethical, engineering, impact, operational, and legal considerations. We designed our interviews and survey broadly to solicit data on all these issues.

3.1 Interview

We conducted one-hour, semi-structured interviews with 15 employees: 6 data scientists, 3 ML researchers, 2 PMs, and 4 UX designers/researchers. We recruited our participants through our network in the organization then used snowball sampling based on participants' recommendations. We asked participants background questions to understand where ML fits into their work practice and then focused on their current work practices around releasing models and communicating about their quality. We asked about the type of models they deal with, their use in the product, and how they ensure the quality of their models. For those participants who were decision makers, we asked what they look for in a model to approve its deployment; for other participants, we asked how they share and present their model if they were looking for approval. Finally, we asked them to tell us a success story for a model that was deployed. Our goal was to observe their journey to release the model and to identify interesting points in communication.

3.2 Survey

Informed by our interviews, we designed a survey to investigate the workflow in which models are communicated to an audience. From our interviews, we identified trends in how model developers share their work, and we used those trends as a starting point for our survey questions. While the interviews were open-ended in nature, in the survey we have provided a list of options from which users could choose their answers, or they could add their own. For example, for questions about with whom developers share their models, we provided the main roles we had collected from our survey as a starting points to choose from and indicated they could add other roles in the open text field. The survey split based on whether the participant presented models or whether models were presented to them. We asked participants to think of their most recent work related to ML and answer the questions accordingly. We made sure that our survey participants' team, product, or service include some aspect of ML by explicitly asking about that. After asking about their experience and education related to ML and their current role, we asked participants a filtering

question to direct them to the right set of questions on the survey. We asked how ML is involved in their role and whether they work directly with the ML model or indirectly, through working on some system feature that uses the ML model. Based on this question we presented a set of questions customized to the type of participants (model presenters or audience). In the context of a recent incident or meeting in which a model was discussed, we asked them about the following areas: their role; the model they discussed; in what form they had presented or shared their model; who was their audience; and the goal of discussing that model. We asked the same questions to participants who had the model presented to them. Our goal was to find gaps and mismatches in what was presented about a model and what was expected. We recruited our participants through internal mailing lists and chat groups that discuss or work on ML projects. We also snowballed to reach a diverse population including UX and program management roles.

3.3 Analysis

We began the analysis process by segmenting the interview transcripts based on the 5 questions we asked to better understand the communication process (Who is the communicator? What is the content? Which channel/medium is used? Who is the receiver? What is the effect?). For each question, we analyzed the transcripts to identify patterns and themes. We used the main themes we identified to guide the design of the survey. The themes included roles commonly involved in the discussion process around ML models, artifacts that were communicated from ML models, mediums in which those models were being communicated, and goals that our participants tried to achieve when communicating their ML models. In addition to the quantitative data collected from the survey, we analyzed the results from the open-ended questions in the survey to identify the challenges and best practices in communicating ML models. Through an iterative qualitative coding process and affinity diagramming [42]] we analyzed the open-ended questions until an agreement on primary themes was reached by all authors. Observations and findings from our interviews and surveys are presented in the results section.

4 RESULTS

4.1 Interview

Our interviews served as a starting point for us to understand how ML is being communicated in a large software company, what roles we should be reaching out to, and which teams were involved. Through these semi-structured interviews, we observed that the aspects of model quality that employees looked at varied based on their role and the relationship that employee had with the model. As a result, we designed the survey to ask specific questions about the employee's involvement with the model and how that model was communicated by that employee to different stakeholders. After an iterative analysis of the open responses from the interviews, the authors grouped the responses into categories and identified the following trends: who, with whom, what, form and goals of communicating ML models. This step in the analysis influenced our questions in the survey, which are described in the survey results section.

4.1.1 Who and with whom. Out of the 15 employees we interviewed, 9 were experienced in model building (data scientists, researchers, SEs) and 6 were stakeholders of ML models working on AI based projects (UX designers and researchers, PMs, design managers). Our interviewees reported fluidity in their roles; for example, some of them were in management roles as PMs and researchers, yet they worked directly on models and interacted with data scientists on a daily basis. Others reported that while they were SEs, they worked on aspects directly related to the model such as data acquisition and training. We noticed the backgrounds of the audience who were involved in discussions around ML models varied depending on the stage of the model and the goal of

the meeting. For example, model developers that were working as researchers mostly dealt with technical audience as they were involved in the initial phases of developing models and standards for teams of data scientists. However, a model builder who was an SE or a PM dealt with a wider set of audience such as legal teams, UX, team leads, etc.

(a) Reported roles from the survey.

(b) Reported answers to the question, how is ML involved in your role?

Fig. 1. Results from survey before filtering out “Neither of the above”

4.1.2 *What.* From our conversations with participants we have gathered 6 aspects of quality that they shared when talking about a model:

- **Errors:** Sharing when the model failed was one of the aspects of model quality that was most emphasized as needed but rarely presented and discussed. An ethics and AI researcher mentioned that one important area for future work that they were then focused on was standardizing an “Error Analysis Report” with each model. Other ML researchers mentioned

that localized error was also important, and that they were building and advocating for tools that show local and global errors in the data from the model. Data scientists also mentioned that when they discuss updates in AI based products, it is important to discuss “what got better and what got worse in the updates of the model.”

- **Visualization:** A PM indicated that from his years of experience with discussing models, he had learned that “stakeholders want to see colors and ranges of an aggregate measure like accuracy. Just red, yellow or green.” He was sharing his frustration about his teams’ effort to share more metrics about their models, while what their stakeholders actually wanted to hear about was different. Their stakeholders, which included PMs, UX designers, and legal teams, want to see visual representations that are easy to understand right away. ML researchers all agreed that visualization and looking at the model in the context of the training set, testing set, and unlabeled data was very helpful for their audience, which was mainly data scientists. Data scientists on the other hand used visualization mostly for exploration. However, one group of data scientists mentioned that they made biweekly and monthly reports in the form of a Business Intelligence (BI) dashboard that they shared in their mailing list.
- **Examples and Demos:** Models’ expected behavior was a recurring topic in our interviews. When a model was presented, stakeholders (collaborators and the audience of a presentation) wanted to know how this model or feature was going to provide a positive impact. One designer mentioned that it was always helpful for him to see a subset of the data in which the model performs well or poorly in a demo to make sure he understood how to handle those edge cases. Another UX researcher confirmed that “context of use is where the meat is. What stress cases affect how the software performs.” A PM mentioned that when data scientists on his team presented a model to him, he really cared about the experience that was wrapped around the model. According to him, “program managers care about how is this going in the wild.” Data scientists felt that they needed to personalize their discussions around the model depending on the audience. For example one data-scientist mentioned: “I present precision, but I have to explain it through an example, e.g., out of every 10 times we get 8 right and so on.” Another mentioned: “Model efficiency, performance and other model metrics that I have to explain and provide examples of how it relates to our goals (e.g., number of bugs fixed in the bug detection NN).” Data scientists presenting to other data scientists and engineers favored exposing more technical details in demos. Some data scientists took the approach of sharing different runs of the model after tuning and choosing the best run.
- **Model Metrics:** Presenting and discussing statistical metrics was mostly what data scientists were comfortable sharing. Given the technical domain of the software company, some PMs had worked as data scientists, some PMs were also comfortable with model metrics. PMs mentioned that they care about precision and coverage as they are important for customer acceptance. Data scientists mentioned ordinary differential equation (ODE) coefficients, min, max, median, precision, recall, etc. Others mentioned “impact metrics” such as an increase in precision and how that impacts profit, or efficiency vs. accuracy and so on. When presenting these metrics, some data scientists were creative in their ways, for example, “I provide ranges for aha moments!”
- **Organizational Goals:** When discussing models, some of our interviewees mentioned that, depending on the audience, it was very important to relate the discussion back to organizational and team goals. One ethics and AI researcher mentioned that when he was presenting the success of his model he “measure[d] success around the total application, not in terms of precision/recall/accuracy of the model.” Another fairness researcher mentioned that the stakeholders they dealt with, in general, cared about how the business was improving, how much money they were losing or making, and what the model looked at to make predictions.

One PM mentioned, “I usually share one number (aggregate measure like accuracy); they just want to see data related to the requirements.” Data scientists mentioned that when they were presenting a model they cared how that model satisfied the business requirements. For example “in search ads” when discussing “main line vs side bar,” what he presented about the model was “relevance, profit and ad weight; how much it gets us money.”

4.1.3 Form. The following are the forms observed in our interviews that employees working on ML used to share and discuss their models with their audience:

- **PowerPoint Presentations and Shared Documents:** All of our participants had used a presentation to share their work at some point. Two data scientists shared their presentations with us. One uses PowerPoint with hidden slides explaining every detail of his work; he shows those hidden slides depending on the audience. The other creates tables of all metrics in a shared document and shares it with his audience before his meetings.
- **Visualization/Dashboards:** PowerBI is one of the popular tools used by data scientists and PMs. It allows them to share an interactive dashboard or export visualizations of their model. One team shares PowerBI dashboards through email periodically within the team. One data scientist mentioned that “visualization is used to understand the data in the initial process.” Some use, for example, a decision tree visualization of data and features. Some use visualization tools (e.g., error analysis and debugging tools) to share specific visualizations of the model.
- **Emails:** Given the nature of close collaboration and continuous updates, teams use emails extensively to share information and discuss their models. As we mentioned in the visualization section, some teams use email as the form of communication to share dashboards and charts in between meetings. According to data scientists and managers, communicating through email is more convenient, and team members can ask questions and familiarize themselves with the models before meeting, which lowers the burden on the data scientists when presenting their work and gives team members time to learn what they need. Reports, BI dashboards, tables of metrics, and runs are some of the elements that are shared through email. One data scientist mentioned that he sends metrics and runs to himself as an email and actually uses that email to present his work during meetings.
- **Meetings:** Meetings, in general, are a recurring pattern and an inevitable task. However, PMs are the ones mostly elaborating on their workflows during meetings. One PM mentioned that they follow an informal approach in which they sit in the same room with the model builder following an iterative process. Another PM stated that “discussions about whether information is reasonable is easier with the modeler in the room.” What we have observed is that PMs prefer to pair different roles in those meetings (e.g., DS, UX, and developers), and they act as the bridge between those different roles during the discussion.
- **Demo or Prototype:** A demo is often used to describe the model and experience of a product when AI is introduced to it. A team lead suggested that a prototype to mimic the experience of a service is always more helpful than seeing numbers and metrics. Furthermore, one observation we made is that data scientists themselves resort to building UX around their model or integrate it with an existing service to explain the model. Two of the data scientists we interviewed mentioned that due to the shortage of UX designers on their team, they actually learned and built their prototype internally to help them discuss their model and its outputs.

4.1.4 Goals. During our interviews, participants shared stories about their experiences presenting ML models. We asked them about the goals they were trying to achieve by communicating their models. The following are the synthesized reported goals of sharing models:

- **Shipping:** As we have observed, the main goal of discussing models is to ship a product in which that model is involved. As one PM mentioned, the goal of his discussions is to “provide enough information to allow my audience to make a decision for shipping.” One data scientist mentioned that they often use a demo to convince decision makers to ship, not just to explain the model’s behavior.
- **Increase Confidence:** An ML researcher mentioned that they work on tools to increase confidence and trust in the decisions made. For example, “the goal of a tool like our Error Analysis Tool is to help people make choices.” One data scientist had an interesting workflow in which he builds a demo for his customer-facing models to increase his confidence in his work before sharing it with his stakeholders: “For the demo, it helps me to convince myself that I have the right model.” He then followed up with “a demo is very powerful to give confidence to the decision maker.”
- **Explain:** Another goal that was implicitly derived from our interviews is to explain the model. Those discussions usually happen between data scientists, software engineers, and PMs. The PM then takes the responsibility to fully understand the model to explain it to a wider range of stakeholders. One data scientist also mentioned that he sends out emails about his model and documentation to explain it in an email thread before the meeting or to prepare more explanatory material.

4.2 Survey

Informed by the results we collected during our interviews, we designed the survey to include some of our interview findings to validate our results on a larger scale. We also allowed users to enter free text in case they had something else to say other than what we provide in the options. We received 168 responses and, after cleaning the data and removing incomplete responses, we had 125 responses that contributed to the results of this work. A huge number of the survey participants were data scientists and engineers because we conducted this survey in a software company (see figure 1).

We synthesized the reported challenges to the following: 1) no problems, meaning that the people in the discussion had sufficient knowledge about ML, and 2) problems, meaning that lack of technical knowledge required additional explanations or more effort in the discussion or made the discussion less effective than it could be. As for participants’ relationships to the model, we identified two groups. The first were model builders working directly on the model. The second group worked indirectly by using an output or a feature from the model. For our initial analysis, we filtered out those who worked neither directly nor indirectly on models, which resulted in 95 participants (76% of the cleaned survey results). Those who were filtered out received open-ended questions about their relationship to their team, because they are on teams working on AI-based products.

Table 1 shows that most of the participants are more on the technical side and work directly on models. The results also show that more employees are facing problems in communication. The table shows the top goals that the participants reported. In the following sections, we share more of the reported goals categorized by the roles of the participants or the audience.

4.2.1 Who and with whom. As shown in Table 1, 67% of the participants work directly on some aspect of model building (e.g., producing data for the model, evaluating the accuracy of the model, deploying the model, etc.). The other 33% work on some aspect of the product or service that

Table 1. Overall results from the survey after filtering out the data. Percentage shown is derived from the number of employees in each of the 4 groups

		Directly working on model n=64		Model Stakeholder n=31	
		Problems n=44	No Problems n=20	Problems n=22	No Problems n=9
Role	Data Scientist	26 (59%)	10 (59%)	-	-
	Team Lead	4 (9%)	2 (10%)	1 (4.5%)	1 (11%)
	Software Eng.	8 (18%)	6 (30%)	12 (54.5%)	5 (56%)
	Program Mgr.	5 (11%)	-	-	-
	Researcher	1 (2%)	2 (10%)	-	-
	UX	-	-	3 (14%)	1 (11%)
Avg. Exp.	Job	12.3	13.8	11.84	16.7
	ML	5.9	9.13	3.25	6.33
Education	Formal	29 (66%)	12 (60%)	9 (41%)	4 (44%)
	On Job Training	13 (30%)	6 (30%)	10 (45.5%)	3 (33%)
	No Training	1 (2%)	2 (10%)	3 (14%)	2 (22%)
Position	Individual	37 (84%)	14 (70%)	21 (95%)	7 (78%)
	Manager	7 (16%)	3 (15%)	1 (4.5%)	2 (22%)
Model	Customer Facing	30 (68%)	14 (70%)	14 (64%)	6 (67%)
	Internal	14 (32%)	5(25%)	8 (36%)	3 (33%)
Goals	To Ship	32 (73%)	15 (75%)	13 (59%)	6 (67%)
	Explain Model	28 (64%)	13 (65%)	12 (54.5%)	4 (44%)
	Integrate	20 (45%)	7 (35%)	11 (50%)	2 (22%)

interacts with a model, but they do not work on the models directly. Almost 70% of the participants indicated problems in communicating models, whether they were communicating the model or the model was communicated to them. Almost 56% of those working directly on models are data scientists. The rest are software engineers 22%, team leads 9%, ML researchers 5% and 8% PMs. As for those indirectly working on ML models, they were mostly software engineers given the nature of the software organization (55%). The rest are PMs 26%, UX 13%, team leads 6%. As for whom these models are communicated, the results are shown in Figure 2. Others include business-related positions such as marketers, sales force, and attorneys.

4.2.2 What. In addition to the results shared in Table 2, survey participants who work with legal and sales teams mentioned legal implications for training on data usage and how exactly the model arrives at answers. In addition, some employees working with software engineering teams mentioned that they provide slides and examples comparing AI with traditional methods such as rule-based workflows.

4.2.3 Form. Table 3 lists the main forms in which participants communicate their models. Face-to-face meetings, emails, and PowerPoint presentations are the forms most used to communicate models. Emails could include dashboards and snapshots of the model results. Emails are also used as guiding agendas in meetings, as they are sent before the meetings, then projected during the meetings. Other forms that were mentioned within data science teams are Jupyter Notebooks, co-developing with SDEs, and sharing files or code of actual working models.

Fig. 2. Results from survey regarding the roles of team members with whom our participants discussed their models.

Table 2. Results from the survey regarding the question, what aspects of your models did you talk about? Percentage shown is derived from the number of employees in each of the 4 groups.

	Directly working on model n=64		Model Stakeholder n=31	
	Problems n=44	No Problems n=20	Problems n=22	No Problems n=9
Model’s behavior	42 (95.5%)	16 (80%)	12 (54.5%)	5 (55.5%)
Accuracy	31 (70%)	16 (80%)	20 (91%)	9 (100%)
Errors	27 (61%)	15 (75%)	10 (45.5%)	8 (89%)
Privacy	26 (59%)	12 (60%)	9 (41%)	5 (55.5%)
Subset of data	26 (59%)	12 (60%)	9(41%)	4 (44%)
Engineering aspects	25 (57%)	14 (70%)	13 (59%)	4 (44%)
Model metrics	25 (57%)	10 (50%)	9 (41%)	3 (33%)
Accountability	12 (27%)	6 (30%)	4(18%)	-
Fairness	11 (25%)	6 (30%)	9(41%)	3(33%)

4.2.4 Goals. Table 4 shows the main goals of sharing and discussing ML models: explaining the model and influencing decisions to ship or deploy. Although explaining models to an audience is challenging, it is still the top goal of discussing models. In addition to what is listed in Table 4,

Table 3. Results from the survey regarding the question, in what form did you share your models? Percentage shown is derived from the number of employees in each of the 4 groups

	Directly working on model n=64		Model Stakeholder n=31	
	Problems	No Problems	Problems	No Problems
PowerPoint	57%	50%	45%	56%
Email	57%	35%	32%	67%
OneNote	41%	35%	27%	22%
Meeting	66%	70%	54.5%	67%
Casual conv.	48%	40%	50%	44%
Demo	45%	45%	54.5%	33.5%
Shared document	36%	45%	27%	27%
Whiteboard	29.5%	35%	50%	50%

Table 4. Results from the survey regarding the question, what was the main goal of sharing the model? Percentage shown is derived from the number of employees in each of the 4 groups

	Directly working on model n=64		Model Stakeholder n=31	
	Problems	No Problems	Problems	No Problems
Debugging	20%	30%	27%	17%
Deploy decisions	45%	30%	27%	17%
Explain model	64%	65%	54.5%	17%
Explore	27%	20%	36%	8%
Integration	45%	35%	50%	8%
To ship service	60%	75%	60%	25%
UX decisions	29.5%	10%	27%	8%

participants working with legal and sales teams cited giving users confidence in the outputs of the model and brainstorming for future investments

4.2.5 What do you look for in order to assess the quality of models? In addition to trusting the model builder as a sign of the quality of the model, how the model performs on segments of the data and model metrics are the main cues to assess models. When we asked employees who work directly on models about what they look for to assess the quality of a model, the top reported answer was that they look at statistical metrics, such as confusion matrices, AUCs, ROCs, etc. (81% for those facing problems in communication, 85% for those who did not report facing any communication problems). They also reported that they look at how the model performs on subsets of the data, for example, certain customer segments, certain telemetry events, etc. (81% for those facing problems, 80% for those who do not). Another aspect of model quality that was reported was looking at the results of A/B testing of features that use the models (75% for those facing problems, 55% for those who do not). Employees who do not work directly on a model reported that they generally trust the model builders to decide when a model is good enough (63% for those facing problems, 44% for those who do not). They also reported that they look at how the model performs on subsets of the

data (45% for those facing problems, 67% for those who do not). They also look at statistical metrics (41% for those facing problems, 55% for those who do not). Other quality cues that were reported include comparing the model with other similar work and assessing its performance against reality over time. Others also reported looking directly at the code.

5 DISCUSSION AND RECOMMENDATIONS

Our interviews and survey focused on observing what quality means to team members working on ML models and how the quality of ML models is communicated within big teams holding different roles. Through our observations, we identified challenges that team members face in perceiving the quality of models and how they tackled. Some teams overcome the challenges in communicating ML models by having a middleman, usually a PM or SE, to communicate model quality aspects between model developers and other team members who are non-ML experts (e.g., UX designers, legal, sales, etc.). This causes some information to be lost in translation. As some of our participants reported, involving ML developers in meetings with other team members has proven to be more efficient and fruitful to the discussion and overall success of a product. In this section, we start by synthesizing and discussing the main challenges in communicating the quality of ML models to a wide audience. After that, we discuss best practices in communicating quality through five lenses (who and with whom, what, form, and goal). Throughout the discussion, we use the word stakeholders to refer to internal employees who are part of the same software organization but are from different teams, such as UX, legal, sales, etc.

5.1 Challenges in Communicating ML Models

In this section, we discuss the main challenges and best practices around communicating ML models within teams. In the surveys and interviews, participants were asked two questions about challenges: One was aimed at ML experts who develop the models, asking them directly about challenges they face when discussing ML models with non-experts. The other question was directed toward non-ML developers working on ML projects and the challenges they face within their teams when those ML models are discussed. Through another affinity diagram exercise, we iterated over the reported open-ended questions, we identified emerging patterns. We share the following challenges and best practices to inspire future work and design direction for new solutions.

5.1.1 Mismatch between the discussion of user-need-driven and model-driven performance. One of the common problems that has been observed in a variety of ways is that when discussing ML features in a system, data scientists tend to be more model driven as opposed to user-need driven (e.g. they discuss the performance of a model in terms of accuracy or recall, but not in terms of the overall task). This mostly affects those ML model developers who are building customer facing features of models. For example, a software engineer mentioned the challenge he faces when discussing the model as follows: *“Our team focuses on building tools for 3rd party developers, so our work is generally focused on models as collections/generally. It’s often hard to discuss the models when we’re focused on how they perform on a standard task (ImageNet, CoCo) but customers care about their specific problem space”*. Another participant who is a UX designer working on customer facing ML based projects explains the situation with the following metaphor: *“Trying to navigate the ‘cart before the horse’ discussion because it can be a sensitive topic. Cart before the horse = conducting research, creating and training models before having a sense of what real customer problem you’re trying to solve. Can lead to sometimes feeling like we’re shoe-horning ML into products where they *might* not be needed”*. Another UX designer who has been working on AI based projects for more than 5 years indicated that the fact that that priorities for each practitioner in AI based teams is different and often causes challenges in communicating efficiently and described as a “build model

first” process “then find each other” referring to the team and product that better fits the model. He described this as more of a model driven process which causes problems when communicating models later on within the team.

5.1.2 Struggle in problem formulation. Another emerging pattern in our observations from the survey and interviews is due to the lack of knowledge regarding the capabilities and potentials of ML. The problem is divided into two groups: One occurs at the beginning of the ML development workflow due to issues in problem formulation as a result of not knowing what to expect from ML, and the other occurs during validation. For example, a data scientist clearly described the challenge he faces regarding his audience expectations: *“Many non-technical people assume that ML can do anything and that it will tell them what they need to know. But ML is only good when you have 1) good questions, and 2) good data (with good labels!). Conveying both of these things to non-ML people is very hard, [particularly] the need for good, focused questions and goals for statistics and ML. Many people also don’t want to hear about model fit statistics, or error metrics. Business leaders just want to know if something ‘works.’ But understanding your model’s error what it ‘can’t’ do is just as important as understanding what it can do”*. Another data scientist mentioned *“[l]ack of clarity from stakeholders on what they expect from the ML solution”*. A model developer for the gaming industry further explained the challenge: *“The Gaming industry is new to this technology, so they are not aware of the breadth and depth of business questions that could be answered with predictive modeling and unsupervised learning methods. General awareness of what is ML and how can it help them are the biggest challenge. I’d love the stakeholders to get to the level of being interested in the accuracy, bias and fairness of the model”*. Another data scientist attributed this challenge in communication due to the fact that *“non engineers don’t understand math, engineers don’t know business logic or context”*.

5.1.3 The need for education before communication. Not only is the lack of ML knowledge challenging at the start of the workflow during problem formulation, but it also requires extra effort from the model developer to educate and raise awareness in their audience. As one data scientist explained, *“We usually need to put in extra work to achieve the baseline level of knowledge before we can discuss the actual mode/feature”*. Another data scientist mentioned how this mismatch in the level of knowledge forces her to hide some details to avoid misunderstanding. In her words, the challenge is *“[h]ow to convert technical terminology into daily words. Sometimes, in order to explain clearly, I have to ignore some exceptions/details to avoid confusion”*. Another participant mentioned how data scientists have to deal with external customers and partners and have to match their level of knowledge: *“[We often] interact with partners who have no ML knowledge at all, so we need to be able to explain at a basic level. E.g., explain ML model to someone with an MBA”*. This problem not only occurs with business partners but also with PMs and UX when designing ML features and understanding how to design for failure.

To help team members prepare for the discussion and learn at their own pace, some model developers reported sharing documented information about the model before the meeting: *“I usually create a OneNote page, add all the models and their statistical performance measures and share it with anyone I want to discuss the model with and then go to them to start the conversation.”* This practice guarantees a more efficient discussion and ensures that time spent is toward important details rather than educating the audience. One manager explained their process on their team: *“Working within a notebook, making sure people are prepared to answer questions about model performance and potential bias issues. The more people are educated on ML fundamentals, the more time we can get into the model-specific issues”*. Not only is this a burden on the educator (i.e., ML expert), but also those with less ML experience would also appreciate the knowledge shared beforehand to help prepare and focus on the important parts of the discussion as suggested by this SE with no background in ML: *“Before discussion, there should be sent a note of resources which will be helpful for me to have a*

better grasp of the discussion”. Resources shared could be anything from online shared documents to videos about a topic or online notebooks to enable further exploration by the audience.

5.1.4 Conflicting documentation and standards. . One of the interesting patterns in the results is how a lack of standards and documentation that are universal across team members is a hurdle in the communication process because no common language is available. One data scientist explained, *“The biggest challenge I’ve run into on our team is folks having a common language to use when discussing models in terms of functionality.”* He then followed up with how it is less of a challenge due to his team process *“Fortunately our API consumes the ONNX model format which has helped improve these conversations.”* Another data scientist mentioned, *“There is no standard checklist for checking the model’s accuracy. When you go and talk to different data scientists, even seniors they have their own mind map, so going through the model with different people is different”*. Because some teams have not moved to any standardized ML development life cycle, some data scientists expressed that it is challenging to establish common ground regarding what they communicate and had to propose metrics for the ML models because they do not exist. A software engineer mentions *“ML models are created by a group and used by another, but the accuracy information and other implementation details are lost in transition. We have to find the hard way about a Model’s inefficiency”*. While lack of standards imposes a challenge, having too many standards and practices from different disciplines causes another communication challenge as described by this designer *“Too many disciplines are involved. Over saturated practices that turn off people and deviate from the truth”*.

5.1.5 Failure to see ‘The Big Picture’. Another challenge around communicating models arises from the fact that an ML model is part of an ecosystem in which conversations are bidirectional. The challenge arises out of the fact that some data scientists are not aware of the deployment and engineering practices. For example one PM mentioned: *“Modelers don’t always know what it takes to take a model to production”*. An SE confirmed this by stating: *“Scoping work can be difficult; it’s not always easy to know how many models we’ll need to try and how long we’ll need to iterate”*. The challenge in communication is not necessarily caused by data scientists trying to explain their models but from their need to know how it fits into the environment where it belongs. As one data scientist described in the survey: *“Engineering presents more issues: many data scientists are not as familiar with engineering and deployment aspects”*.

5.1.6 Struggle to explain and understand common model metrics in context. Variations in metrics and their subjectivity raises a challenge in discussions within teams. For example one a data scientist mentioned, *“Most people do not understand that it’s difficult to compare metrics across models - sometimes an AUC of 0.9 is good enough, sometimes it is not”*. Another senior data scientist mentioned something along those lines: *“Sometimes a slightly low accuracy does not ruin the UX, it could be ignored or maybe it does not have impact. It is hard to convey that to our stakeholders”*. Another issue related to communicating meaningful metrics is how they are being presented and discussed, because some data scientists choose to bias their presentation toward those metrics that work. As a data scientist expressed: *“Not everyone knows what metrics really mean. Some people just blindly follow the metrics. Authors of models usually highlight the examples where the model performs well”*. A designer mentioned that in his experience data scientists are usually *“optimistic”* and do not usually expose the failures. For example, in a model with %95 accuracy rate, the builder would emphasize on the %95, but what the designer really cares about is the %5: Who are the affected users? What is the cost of not supporting them?

5.1.7 Struggle to explain and interpret model behavior. There is often a difficulty in *“explaining”* a model’s behavior, (i.e., When the model works well, why? When the model is not working as

expected, why?) As three data scientists mentioned, respectively, “[It is] hard to explain why [an] ML model performs in a particular way (some models are not very interpretable and hard to explain)”, “Model explanation should be the most challenge when I discuss with other team members”, “explaining cases when the model might not perform well”. A PM mentioned that in his discussions with data scientists, he cares about the explicable nature of the model and that it is mostly challenging for data scientists: “How explicable is this model? It helps to have a nice story around the model, what’s the story and how will it perform in the wild? The person has to [tell] a story”.

5.1.8 Intimidation by the perceived complexity, or is it too much trust? An obvious pattern in our interviews with data scientists is the issue of trust in the sense that a model consumer often trusts that the model builder knows their job; therefore, there is not as much discussion around the model as there might need to be. As one UX designer mentioned, “I have no formal education related to ML models. So basically, I just rely on the adhoc understanding of the model developer”. A PM also mentioned that, “stakeholders don’t care about model metrics they just want to know if the model is good or not, Usually I just share one number (aggregate measure like accuracy) and no one ever asked for details. They just want to see data related to the requirements”. However, one data scientist explained this as a challenge and expressed their desire for their audience to be more engaged and curious about important model metrics such as errors or data quality. For example, one data scientist explained: “Many people also don’t want to hear about model fit statistics, or error metrics. Business leaders just want to know if something works. But understanding your model’s error what it *can’t* do is just as important as understanding what it can do.”

5.2 Who Discusses the Model Quality (and With Whom)

As shown in 1, employees with more years of working experience face fewer problems when communicating their models within teams. In addition, having more experience on ML-based projects makes communication easier. Longer experience entails previously working in a variety of roles. For example, from our interviews, we learned that some senior PMs were originally data scientists or statisticians before their current position. As a result, they are now PMs who are heavily involved in the model-building process. In the survey, the same pattern was observed even with other roles. For example, one SE working on models stated, “we are all considered software engineers, even though I make changes in the trainer code and cost function for example”. This is not surprising as more employees are opting to learn ML, job roles and titles become blurred [1, 3, 47]. However, although job roles are more fluid, employees still face challenges, especially for software engineers which echoes existing work [16, 36, 37, 46]. Most reported challenges in communication occurred when data scientists communicated with SEs, given the close and recurring interaction between data scientists and SEs in the company.

5.2.1 Co-development with different team members. To improve team dynamics and lower the barriers in explaining and discussing ML models, continuous feedback from “with whom” the model is shared with helps get everyone in the team on the same page, “Present your work in peer review sessions to get inputs from peers in the team.”. In our surveys, a team lead pointed out that the skill of asking for feedback and making the effort to be part of those conversations is a crucial skill: “Feedback handling is critical – folks often want to help but need to know how.”. Discussing ML models continuously is also an important part of the model’s life cycle, as one SE explained “There are daily stand-ups for communication. Creating models in isolation is generally much harder to deploy.”.

As model developers, PMs and SEs can co-develop and work in the same room to have more organic conversations that allow for iterations and continuous feedback: “Discussion about whether info is reasonable is easier with the modeler in the room.”. Developing models in a silo will only make discussing those models with stakeholders more difficult: “My team’s best practice is to identify a

team where ML can really help and form a small v-team from my AI team and the dev team to work on that ML problem and co-develop models. There are daily standups for communication. Creating models in isolation is generally much harder to deploy". A more inclusive development cycle would pair the ML model developer with a user experience designer or researcher, specially when this model is critical and customer facing. A PM described his process in his team: *"Modelers are involved in the discussions with the user/ Watch users is highly motivating. Getting exposure second hand is less motivating. Show the modeler that the job isn't done".* Partners or customers could be involved as well creating a more model "participatory design" process which would help increase understanding and the sense of ownership over the products: *"Ensure that internal partners or internal customers of these ML models have input into these models, to gain a sense of co-ownership with the product"*

5.3 What to Discuss About Models' Quality

We have observed that practitioners discuss models from different angles (e.g. errors, data, ROI, UX and functionality, etc.). Whereas model builders look at quality in a more technical way, those who reported that they were successful in communicating their work usually avoid presenting technical metrics. The representation of model's quality is tailored to each type of audience not tied to the model itself. For example, with legal teams, privacy and accountability are reported to be discussed more often. Another example is PMs, who reported discussing the fairness of the models. However, current tools and practices make discussing fairness challenging [12, 12, 29]. From our results, we found that those who have no problems in discussing models, whether they are presenting a model or a model is presented to them, do not talk about technical model metrics.

5.3.1 Standards. There is no one way to discuss model quality as there are no standards that are clearly understood by non-ML developers, unless we look at some forms of model documentation that could, when implemented properly within the team, serve as discussion facilitators. As we mentioned previously in the challenges, the lack of universal accepted standards between different roles in the team is a hurdle when it comes to discussion. However, when we look at specific aspects within some experienced teams, for example data scientists communicating with each other or other ML trained engineers, some standards are becoming more accepted and used. For instance, one data scientist reflected on his experience: *"The biggest challenge I've run into on our team is folks having a common language to use when discussing models in terms of functionality. Fortunately our API consumes the ONNX model format which has helped improve these conversations."* ONNX is a standard for exchanging deep learning models which is designed to provide framework interoperability. [11] Another standardization that could help in communicating error metrics for more advanced audiences is to have a better understanding of the cost functions : *"Standardize the error metrics and cost functions to account for weighted cost functions"*. Standardization in data is also encouraged, for example standardizing test data sets helps in communicating the results fairly : *"Standard blind evaluation sets. Easy for anyone to measure themselves and reduce variance (e.g. force teams to train 5 variations with different random seeds and pick the medium quality numbers)"*. In general, whether data or models are being discussed, using a common language should be the goal before communicating the technicalities, as simply put by this PM: *"Have a common ML framework and/or a deep learning explainer/terms glossary to be sure we are speaking the same language."*

5.3.2 It's all about the data! Another important aspect to discuss when communicating the quality of ML models is data. As a data scientist put it *"Do we have the right data that is suitable to be fed to the model?"*. The right data includes discussing how the data is collected, the source of the data and diversity of the data. As mentioned by another two data scientists *"setting clear expectations around data collection and cleaning is important. The model is a tiny part of the whole process."* They also stated that it is important to *"[f]ocus on the input/content on how it is leveraged rather than*

feature/model engineering". With GDPR and privacy rules being enforced and followed, many team leads stated that privacy is a concern and a priority in their workflows, so discussing data related issues, especially related to end-users, is an important part of the model discussion. Sometimes it is hard to separate the data from the model, but for the sake of discussion emphasizing on the data is crucial. For example, a team lead suggested, *"Talk about freshness, accuracy of results and how to handle corner cases"*. The corner cases could be due to the nature of the data (e.g. marginalized groups).

5.3.3 Workflows. Sharing the process that led to the solution is also crucial in promoting confidence in the quality of the solution. Some details about ML models seem to be important to the audience, such as the reason for choosing a specific class of models or how the parameters were tuned and so on. For instance, a data scientist mentioned what to discuss around ML models: *"Problem being solved and why we picked the ML model? Your thought process behind picking the model"*. A software engineer also suggested a more engaging discussion in which the results and the journey to the solution matter to non-ML developers *"show how you've arrived at the results and how tweaking certain parameters can impact the model output (in short show the journey and not just the final outcome of the model)"*. Regarding how to share those workflow, a data scientist recommended the following *"Ideally, start with the high-level model, the motivation and applicability to the problem and the experimental results. Then, show the demo if applicable. Finally, go into the nitty-gritty details about how the model is constructed and trained. If you start with too many details about the model, people lose the overall big picture."*. Another way to discuss the journey to a solution is by discussing ML pipelines as recommended by a software engineer *"I would recommend discussing ML models in terms of pipelines. ML pipelines involves the entire evolution of transforming data into intelligence through mathematical transformations. When we talk about ML model in terms of pipeline we cover end to end value add of the model to a business solution"*.

5.4 Forms to Show Quality

In this section we discuss the forms used to communicate ML models (process, workflows, results etc.).

5.4.1 Presentation tools. Looking at the results of Table 3, we see some practices and tools that are currently used to present models' quality to elicit launching decisions or any of the other goals shown in Table 4. PowerPoint and OneNote were popular in communicating models, which presents a potential design opportunity. The current state of these tools is that they are disconnected from data scientists' workflows. When these tools are used, results are shared in a static way. Integrating features to explain models in such presentation tools would be ideal in supporting discussions around models. On the other hand, sometimes tools are used to bypass discussion altogether. For example, as one team lead explained *"We have also made a system to create simple models with online training for parameter tuning so that non-ML SDEs can start to create their own models."*

5.4.2 Demo of the model. Another form of presenting the quality of the model is sharing a prototype to demo the behavior of the model. Prototyping ML systems demonstrates the model in a more expressive and meaningful way to a diverse set of audience, making the model more approachable to non-ML experts. [29] As one data scientist mentioned, *"I built a UI to show what features played a role in making the ML algorithm worked the way it worked. This helped the users a lot to understand when the algorithm produced wrong results. They also understood the limitation of the technology and also suggested new feature engineering ideas to further improve the solution"*. Another data scientist who is a manager mentioned a best practice they follow on his team: *"Demos on showing the model input, output, parameters and results on different segments sets a great context both for debugging and*

explaining." In this case, demos were used to focus on a subset of the data to show how the model behaved.

Another important advantage of prototyping a demo as a form of communication, is building confidence and trust in the audience. A data scientist manager who has been working on applying ML to developer tools to make coding a better experience for developers mentioned that a "demo is very powerful to give confidence to the decision makers". The fidelity of the demo depends on the task; for example, one data scientist mentioned that one should "[b]uild good ux around the model in order to elicit more relevant feedback". In our interview with this data scientist he mentioned that he struggled to elicit good feedback when he built a low fidelity prototype at first, in which it was more confusing as he was not driving the discussion he aimed for (i.e. explaining and getting feedback.). For that reason he iterated over his prototype and provided his team with a high fidelity demo which was a successful attempt.

5.4.3 Quality shines with context and visuals. In presenting models, having the right form of presentation that relates the model to a context is a helpful exercise. As one PM mentioned regarding his approach when data scientists are presenting their models to him: *"How explicable is this model? it helps to have a nice story around the model, what's the story and how will it perform in the wild? the person has to [tell] a story."* In addition to stories around the model, "walk through examples" and relevant use-cases are also helpful. A manager who used to be a data scientist mentioned *"Keep it simple. Imaging you are talking to someone with just a business degree and no advance math. Make analogies or use examples they can relate to."* The discussion should be tailored to the type of audience and the problems. For instance, a solution for an image recognition system when presented with visualization of the segment of the data in which the model fails helped the audience clearly understand how it might behave in the wild. Visualizations are also helpful in shedding light on the data for exploratory and debugging purposes to convey quality issues to other data scientists and engineers as observed from our interviews. For example, *"using dashboards to show the model input, output, parameters and results on different segments sets a great context both for debugging and explaining the model."*

5.5 Goal of Discussion

As shown in (Table 4), explaining models was one goal that was mentioned by many of our participants. Given the diverse background of team members, some of them are not trained in ML, which causes some issues for model builders. As some of our participants reported, a mentorship model is helpful in teaching new skills and raising the level of ML expertise within a team. For example, this best practice was suggested by an SE to facilitate the discussion on his team *"Formal training or mentorship from an experienced person in ML"*. However, this might be time consuming for team members [35, 61].

5.5.1 Risk mitigation. Another aspect of quality is risk assessment, which is important to discuss. Just like any organization needs to mitigate risks, it is very important to emphasize risks associated with delivering ML solutions as this directly impacts revenue, productivity, UX etc. Some of the risks that should be highlighted are related to data privacy and service production: *"[you] [n]eed to set the right expectation to people for the ML models. Discuss early on about every possible issues that would block shipment, e.g. privacy concerns"*. It is important to educate the stakeholders about the uncertainties associated with model results and metrics [34], the numbers produced using testing sets do not necessarily represent how the model will behave in the wild [4]. As one data scientist suggested: *"Under promise and over deliver. Work with stakeholders to determine an acceptable level of correctness (ex. accuracy); remind them that we may not be able to achieve that level in practice"*.

5.5.2 Explain errors. Highlighting the risks and uncertainties also included making sure that when reporting numbers such as errors, your stakeholders are aware of what those numbers actually mean in terms of the business goal or customer experience: *“Always report error. And understand what your errors actually *mean*. E.g., just reporting an R^2 metric of F1 score does not tell you the types and magnitudes of the errors it makes, or what the errors mean in real, practical terms. This is incredibly important”*. Errors are very subjective and in some scenarios having more errors on a subset of the data could be culturally or politically unacceptable, so having these clear conversations with stakeholders is very important. Furthermore, showing and explaining where models fail is an approach that been appropriated from practices in the industry such as failure modes and effects analysis (FMEA) [54]. Some tools such as in [8, 9, 49, 55] received good feedback when used to present models to explain and give confidence to a larger set of audience and could be incorporated into model discussion workflows.

6 LIMITATIONS AND CONCLUSION

Our results were elicited from observing the communication process through our interviews and surveys. Our questions were inspired by Lasswell’s model of communication [25] to provide a comprehensive analysis of the communication process. During our interviews, participants shared artifacts they used to discuss ML models (e.g., emails sent to other team members, presentations, projecting their work, etc.), which helped us in the elicitation process. However, a longer ethnographic shadowing would be beneficial in adding to the challenges and best practices we have identified. Our recruited interview and survey participants were all employees of a software company with large multidisciplinary teams. Most of our participants had some exposure to ML because the company is actively advertising events and courses to learn ML. Therefore, some of our observations might not generalize to less experienced teams. Furthermore, the total number of data scientists and software engineers was higher than the numbers of any other user group reflected in the software company population. We tried to randomize the sample by randomizing the questions regarding communication to different roles.

Our contribution in this paper is an investigation on how different roles in a team communicate about the quality of ML models. We conducted a mixed-methods study at a large, international technology company, first interviewing 15 employees in a variety of roles, then surveying 168 employees to broaden our understanding. In our results, we talked about the quality of ML models, which includes aspects beyond technical issues of model evaluation, such as accuracy. We then presented the challenges that teams face in communicating the quality of ML models, as well as work practices to mitigate those challenges. We found several challenges, including a mismatch between user-focused and model-focused notions of performance, misunderstandings about the capabilities and limitations of evolving ML technology, and difficulties in understanding concerns beyond one’s own role. We found several mitigation strategies, including the use of demos during discussions to keep the team customer-focused. We also found that whereas model builders look at quality in a more technical way, those who report no problems in communication usually present fewer model-related technical metrics.

7 ACKNOWLEDGMENTS

We thank our interview and survey participants for their time and valuable input.

REFERENCES

- [1] [n.d.]. 2017 Kaggle ML & DS Survey. <https://kaggle.com/kaggle/kaggle-survey-2017> Library Catalog: www.kaggle.com.
- [2] [n.d.]. How to Prevent Discriminatory Outcomes in Machine Learning. <https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-in-machine-learning/> Library Catalog: www.weforum.org.

- [3] 2014. 3 Data Careers Decoded and What It Means for You. <https://blog.udacity.com/2014/12/data-analyst-vs-data-scientist-vs-data-engineer.html> Library Catalog: blog.udacity.com Section: Career Guidance.
- [4] 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (Oct. 2018). <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>
- [5] 2018. Tutorial: 21 fairness definitions and their politics. https://www.youtube.com/watch?v=jIXluYdnnyk&ab_channel=ArvindNarayanan
- [6] Jeroen C. J. H. Aerts, Keith C. Clarke, and Alex D. Keuper. 2003. Testing Popular Visualization Techniques for Representing Model Uncertainty. *Cartography and Geographic Information Science* 30, 3 (Jan. 2003), 249–261. <https://doi.org/10.1559/152304003100011180>
- [7] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, and Harald Gall. [n.d.]. Software Engineering for Machine Learning: A Case Study. ([n.d.]), 10.
- [8] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 337–346. <https://doi.org/10.1145/2702123.2702509> event-place: Seoul, Republic of Korea.
- [9] Gennady Andrienko, Natalia Andrienko, Steven Drucker, Jean-Daniel Fekete, Danyel Fisher, Stavros Idreos, Tim Kraska, Guoliang Li, Kwan-Liu Ma, Jock Mackinlay, Antti Oulasvirta, Tobias Schreck, Heidrun Schmann, Michael Stonebraker, David Auber, Nikos Bikakis, Panos Chrysanthis, George Papastefanatos, and Mohamed Sharaf. 2020. Big Data Visualization and Analytics: Future Research Challenges and Emerging Applications. <https://hal.inria.fr/hal-02568845>
- [10] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. 2018. FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity. *arXiv:1808.07261 [cs]* (Aug. 2018). <http://arxiv.org/abs/1808.07261> arXiv: 1808.07261.
- [11] author. [n.d.]. ONNX: Open Neural Network Exchange Format. <https://onnx.ai/>
- [12] Ricardo Baeza-Yates. 2016. Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science (WebSci '16)*. Association for Computing Machinery, New York, NY, USA, 1. <https://doi.org/10.1145/2908131.2908135>
- [13] Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. 2017. How Data Workers Cope with Uncertainty: A Task Characterisation Study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM Press, Denver, Colorado, USA, 3645–3656. <https://doi.org/10.1145/3025453.3025738>
- [14] George E. P. Box. 1976. Science and Statistics. *J. Amer. Statist. Assoc.* 71, 356 (Dec. 1976), 791–799. <https://doi.org/10.1080/01621459.1976.10480949> Publisher: Taylor & Francis _eprint: <https://www.tandfonline.com/doi/pdf/10.1080/01621459.1976.10480949>.
- [15] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley. 2017. The ML test score: A rubric for ML production readiness and technical debt reduction. In *2017 IEEE International Conference on Big Data (Big Data)*. 1123–1132. <https://doi.org/10.1109/BigData.2017.8258038>
- [16] Carrie J Cai and Philip J Guo. [n.d.]. Software Developers Learning Machine Learning: Motivations, Hurdles, and Desires. ([n.d.]), 10.
- [17] Nan-Chen Chen, Jina Suh, Johan Verwey, Gonzalo Ramos, Steven Drucker, and Patrice Simard. 2018. AnchorViz: Facilitating Classifier Error Discovery through Interactive Semantic Data Exploration. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval - IUI '18*. ACM Press, Tokyo, Japan, 269–280. <https://doi.org/10.1145/3172944.3172950>
- [18] Comet.ml. [n.d.]. Comet.ml - Supercharging Machine Learning. <https://www.comet.ml/>
- [19] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM Press, Denver, Colorado, USA, 278–288. <https://doi.org/10.1145/3025453.3025739>
- [20] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Paweł W. Woźniak, and Albrecht Schmidt. 2019. The Role of HCI in Reproducible Science: Understanding, Supporting and Motivating Core Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19*. ACM Press, Glasgow, Scotland Uk, 1–6. <https://doi.org/10.1145/3290607.3312905>
- [21] Richard Finger and Ann M. Bisantz. 2002. Utilizing graphical formats to convey uncertainty in a decision-making task. *Theoretical Issues in Ergonomics Science* 3, 1 (Jan. 2002), 1–25. <https://doi.org/10.1080/14639220110110324> Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/14639220110110324>.
- [22] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Dauméé III, and Kate Crawford. 2018. Datasheets for Datasets. *arXiv:1803.09010 [cs]* (March 2018). <http://arxiv.org/abs/1803.09010> arXiv: 1803.09010.
- [23] R. Stuart Geiger, Nelle Varoquaux, Charlotte Mazel-Cabasse, and Chris Holdgraf. 2018. The Types, Roles, and Practices of Documentation in Data Analytics Open Source Software Libraries. *Computer Supported Cooperative Work (CSCW)*

- 27, 3 (Dec. 2018), 767–802. <https://doi.org/10.1007/s10606-018-9333-1>
- [24] Philip Guo. [n.d.]. Data Science Workflow: Overview and Challenges. <https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext> Library Catalog: cacm.acm.org.
- [25] Lasswell Harold D. 1948. The structure and function of communication in society. In *The Communication of Ideas*. Harper's, New York, N.Y.
- [26] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, Barcelona, Spain, 392–402. <https://doi.org/10.1145/3351095.3372831>
- [27] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 579:1–579:13. <https://doi.org/10.1145/3290605.3300809> event-place: Glasgow, Scotland Uk.
- [28] 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:1805.03677 [cs]* (May 2018). <http://arxiv.org/abs/1805.03677> arXiv: 1805.03677.
- [29] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–16. <https://doi.org/10.1145/3290605.3300830>
- [30] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–26. <https://doi.org/10.1145/3392878> arXiv: 2004.11440.
- [31] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. 2019. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 903–913. <https://doi.org/10.1109/TVCG.2018.2864889>
- [32] Alex Kale, Matthew Kay, and Jessica Hullman. 2019. Decision-Making Under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 202:1–202:14. <https://doi.org/10.1145/3290605.3300432> event-place: Glasgow, Scotland Uk.
- [33] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (Ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5092–5103. <https://doi.org/10.1145/2858036.2858558> event-place: San Jose, California, USA.
- [34] Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 2015. How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 347–356. <https://doi.org/10.1145/2702123.2702603> event-place: Seoul, Republic of Korea.
- [35] Claire Kayacik, Sherol Chen, Signe Noerly, Jess Holbrook, Adam Roberts, and Douglas Eck. 2019. Identifying the Intersections: User Experience + Research Scientist Collaboration in a Generative Machine Learning Interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19*. ACM Press, Glasgow, Scotland Uk, 1–8. <https://doi.org/10.1145/3290607.3299059>
- [36] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The emerging role of data scientists on software development teams. In *Proceedings of the 38th International Conference on Software Engineering (ICSE '16)*. Association for Computing Machinery, Austin, Texas, 96–107. <https://doi.org/10.1145/2884781.2884783>
- [37] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2018. Data Scientists in Software Teams: State of the Art and Challenges. *IEEE Transactions on Software Engineering* 44, 11 (Nov. 2018), 1024–1038. <https://doi.org/10.1109/TSE.2017.2754374> Conference Name: IEEE Transactions on Software Engineering.
- [38] Sean Kross and Philip J. Guo. 2019. Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 263:1–263:14. <https://doi.org/10.1145/3290605.3300493> event-place: Glasgow, Scotland Uk.
- [39] Peter Kun, Ingrid Mulder, and Gerd Kortuem. 2018. Design Enquiry Through Data: Appropriating a Data Science Workflow for the Design Process. In *Proceedings of the 32Nd International BCS Human Computer Interaction Conference (HCI '18)*. BCS Learning & Development Ltd., Swindon, UK, 32:1–32:12. <https://doi.org/10.14236/ewic/HCI2018.32> event-place: Belfast, United Kingdom.

- [40] Matthew Lease. 2011. On quality control and machine learning in crowdsourcing. In *Proceedings of the 11th AAAI Conference on Human Computation (AAAIWS'11-11)*. AAAI Press, 97–102.
- [41] Lezhi Li, Yunfeng Bai, and Yang Wang. 2019. Manifold: A Model-Agnostic Visual Debugging Tool for Machine Learning at Uber. (2019). <https://www.usenix.org/conference/opml19/presentation/li-lezhi>
- [42] John Lofland and John Lofland (Eds.). 2006. *Analyzing social settings: a guide to qualitative observation and analysis* (4th ed ed.). Wadsworth/Thomson Learning, Belmont, CA.
- [43] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R. Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How Data Scientists Work Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (Dec. 2019), 1–23. <https://doi.org/10.1145/3361118> arXiv: 1909.03486.
- [44] Lauren Kirchner Surya Julia Angwin Mattu, Jeff Larson. 2016. Machine Bias. *ProPublica* (2016). <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=pB6i06Iyo0OLwmE2vfYUQBGseZmS8U0E>
- [45] 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 - FAT* '19*. ACM Press, Atlanta, GA, USA, 220–229. <https://doi.org/10.1145/3287560.3287596>
- [46] Michael Muller, Melanie Feinberg, Timothy George, Steven J. Jackson, Bonnie E. John, Mary Beth Kery, and Samir Passi. 2019. Human-Centered Study of Data Science Work Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–8. <https://doi.org/10.1145/3290607.3299018>
- [47] Syed Sadat Nazrul. 2018. DevOps for Data Scientists: Taming the Unicorn. <https://towardsdatascience.com/devops-for-data-scientists-taming-the-unicorn-6410843990de> Library Catalog: towardsdatascience.com.
- [48] Azadeh Nematzadeh, Giovanni Luca Ciampaglia, Filippo Menczer, and Alessandro Flammini. 2018. How algorithmic popularity bias hinders or promotes quality. *Scientific Reports* 8, 1 (Dec. 2018), 15951. <https://doi.org/10.1038/s41598-018-34203-2> arXiv: 1707.00574.
- [49] Gagan Bansal Besmira Nushi and Ece Kamar. [n.d.]. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. ([n.d.]), 9.
- [50] Lace Padilla, Matthew Kay, and Jessica Hullman. 2020. *Uncertainty Visualization*. preprint. PsyArXiv. <https://doi.org/10.31234/osf.io/ebd6r>
- [51] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, Atlanta, GA, USA, 39–48. <https://doi.org/10.1145/3287560.3287567>
- [52] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 136:1–136:28. <https://doi.org/10.1145/3274405>
- [53] Kayur Patel, James Fogarty, James A. Landay, and Beverly Harrison. 2008. Investigating statistical machine learning as a tool for software development. In *Proceeding of the twenty-sixth annual CHI conference on human factors in computing systems - CHI '08*. ACM Press, Florence, Italy, 667. <https://doi.org/10.1145/1357054.1357160>
- [54] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *arXiv:2001.00973 [cs]* (Jan. 2020). <http://arxiv.org/abs/2001.00973> arXiv: 2001.00973.
- [55] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D. Williams. 2017. Squares: Supporting Interactive Performance Analysis for Multiclass Classifiers. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan. 2017), 61–70. <https://doi.org/10.1109/TVCG.2016.2598828>
- [56] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Muvkavilli, Konrad P. Kording, Carla Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2019. Tackling Climate Change with Machine Learning. *arXiv:1906.05433 [cs, stat]* (June 2019). <http://arxiv.org/abs/1906.05433> arXiv: 1906.05433.
- [57] Marck Harlan Harris Vaisman, Sean Murphy. [n.d.]. Analyzing the Analyzers - O'Reilly Media. <https://www.oreilly.com/data/free/analyzing-the-analyzers.csp> Library Catalog: www.oreilly.com.
- [58] April Yi Wang, Anant Mittal, Christopher Brooks, and Steve Oney. 2019. How Data Scientists Use Computational Notebooks for Real-Time Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–30. <https://doi.org/10.1145/3359141>
- [59] Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J. Smith, Kalyan Veeramachaneni, and Huamin Qu. 2019. ATMSer: Increasing Transparency and Controllability in Automated Machine Learning. In *Proceedings of*

- the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 681:1–681:12. <https://doi.org/10.1145/3290605.3300911> event-place: Glasgow, Scotland Uk.
- [60] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, Barcelona, Spain, 1–18. <https://doi.org/10.1145/3351095.3372833>
 - [61] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 585–596. <https://doi.org/10.1145/3196709.3196730> event-place: Hong Kong, China.
 - [62] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18*. ACM Press, Hong Kong, China, 573–584. <https://doi.org/10.1145/3196709.3196729>
 - [63] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *arXiv:2001.06684 [cs, stat]* (Jan. 2020). <http://arxiv.org/abs/2001.06684> arXiv: 2001.06684.

Received August 2020; revised November 2020; accepted December 2020