Understanding Development Process of Machine Learning Systems: Challenges and Solutions

Elizamary de Souza Nascimento
Institute of Computing
Universidade Federal do Amazonas
Manaus, Brazil
elizamary.souza@icomp.ufam.edu.br

Iftekhar Ahmed
Department of Informatics
University of California
Irvine, USA
iftekha@uci.edu

Edson Oliveira
Department of IT
Secretaria de Estado da Fazenda
Manaus, Brazil
edson.cesar@sefaz.am.gov.br

Márcio Piedade Palheta
Executive Director
Buritech Sistema Inteligentes
Manaus, Brazil
marcio.palheta@buritech.com.br

Igor Steinmacher
Computing and Cyber Systems
Northern Arizona University
Flagstaff, USA
igor.steinmacher@nau.edu

Tayana Conte
Institute of Computing
Universidade Federal do Amazonas
Manaus, Brazil
tayana@icomp.ufam.edu.br

Abstract*— *Background: The number of Machine Learning (ML) systems developed in the industry is increasing rapidly. Since ML systems are different from traditional systems, these differences are clearly visible in different activities pertaining to ML systems software development process. These differences make the Software Engineering (SE) activities more challenging for ML systems because not only the behavior of the system is data dependent, but also the requirements are data dependent. In such scenario, how can Software Engineering better support the development of ML systems? ***Aim*:** Our objective is twofold. First, better understand the process that developers use to build ML systems. Second, identify the main challenges that developers face, proposing ways to overcome these challenges. ***Method*:** We conducted interviews with seven developers from three software small companies that develop ML systems. Based on the challenges uncovered, we proposed a set of checklists to support the developers. We assessed the checklists by using a focus group. ***Results*:** We found that the ML systems development follow a 4-stage process in these companies. These stages are: understanding the problem, data handling, model building, and model monitoring. The main challenges faced by the developers are: identifying the clients' business metrics, lack of a defined development process, and designing the database structure. We have identified in the focus group that our proposed checklists provided support during identification of the client's business metrics and in increasing visibility of the progress of the project tasks. ***Conclusions*:** Our research is an initial step towards supporting the development of ML systems, suggesting checklists that support developers in essential development tasks, and also serve as a basis for future research in the area.

Keywords—Machine Learning Systems, data handling, software development, Software Engineering, challenges

I. Introduction

The number of Machine Learning (ML) systems developed in industry is increasing rapidly [1, 2, 7]. Brynjolfsson et al. [3] showed that the development process of ML systems differs from the process used for developing traditional software. One key difference is in the data handling process for identifying valid, new and useful patterns in existing dataset [7]. Another difference is that ML systems are data-centric and there are multiple feedback cycles between the different stages of the process, such as model training, feature engineering, and so on [2]. Researchers have investigated the development process for ML system and found that currently, there is no common development process as ML system development is still evolving [4, 7].

Lack of process and differences between ML and traditional software development make the Software Engineering (SE) activities more challenging in ML systems [1]. Some of the known challenges include: (a) how to specify and translate the requirements, since many requirements are discovered through the process of data handling [1, 7]; and (b) how to set or choose the performance metrics for these systems in practice [3]. Besides, these challenges raise other issues, like: How can SE process better support the development of ML systems? What techniques and practices can be adopted to facilitate the development of ML systems?

Researchers have been investigating the ML systems' development process, practices and challenges faced by software industry professionals [2, 4, 7] to answer the questions posed earlier. These studies focused on developers working in large organizations. Given that several ML system development companies are either startups or small companies with few developers, it is of utmost importance to understand the needs and challenges of developers working in these small organizations. Thus, we performed a study with developers of three companies specialized in developing ML systems. In this study, we formulated the following research questions:

RQ1: How software developers build ML systems in small companies? Previous research investigated how developers of large companies, such as Microsoft and GitHub, work on the development of ML systems [2, 4, 7]. In addition, these studies show the practices and challenges that they face in the industry. In contrast to these studies [2, 4, 7], we investigated how developers of three local and smaller companies develop ML systems. Among these three companies, two are startups that develop ML systems for retail, and one develops ML systems for the Government. Additionally, we assessed whether the professionals from these three companies follow similar processes and practices that professionals from other companies have been doing [2, 4, 7] or not.

RQ2: What challenges are perceived by developers during the development of ML systems in small companies? In this

978-1-7281-2968-6/19/$31.00 ©2019 IEEE

question, our goal was to understand the challenges when developing ML systems, and to assess how those challenges can impact the ML project. Once again, we assessed whether these challenges are similar to those reported on previous studies [4, 7] or not.

RQ3: Is it possible to help the developers overcome these challenges? Altarturi et al. [1] and Amershi et al. [2] present SE initiatives to support the development of ML systems. In this study, we also proposed SE checklists to support the development of ML systems. Our goal is to assess if it is possible to help the developers facing the challenges. To assess our proposed checklists, we arranged a focus group, so that we can get feedback regarding the proposed checklists and their benefits.

The remainder of the paper is divided into six sections. Section 2 presents research that show how developers work in the industry in the development of these systems. In Section 3, we describe the qualitative study. In Section 4, we present the results obtained in this study. In Section 5, we present our proposal and the results of its initial evaluation. In Section 6, we present discussions on the results obtained. Finally, in Section 7, we present the conclusions and future work.

II. Background

Researchers have been investigating the ML systems' development process, practices, and challenges faced by software industry professionals [2, 4, 7, 8]. Each of these studies presented development processes for data science, data mining, and ML systems contexts. Still, these studies found different numbers of stages in the development process, indicating the lack of a well-defined development process for ML systems [2, 4].

Byrne et al. [4] noted how software developers work in GitHub's data science teams. They showed that ML teams of this company follow ten stages. Besides, the authors found that one practice used by the teams is to have a "stage zero" before starting the project. In this stage they conduct an exploratory analysis of the data to better define the scope and plan the project. Still, the authors identified challenges faced by ML teams such as combining customer success and ML metrics, and having a priori knowledge of the structure and format of the data that will be used in the project.

Amershi et al. [2] analyzed how software teams develop ML systems at Microsoft. The authors identified nine stages followed by the development teams. The stages are primarily divided into "data-oriented" and "model-oriented" stages. For example, in the data-oriented, the first stage is to collect data, and the second stage is to clean data. The model-oriented stage starts with the model requirements, followed by feature engineering, model training, model evaluation, among others. Authors noticed the existence of feedback loops during the ML systems' development. For example, during the evaluation stage of the model, it is possible to return at any previous stage. The authors also presented a set of recommended practices for overcoming challenges while developing ML systems. These practices include integrating ML development support into the infrastructure of other systems development, data collection, cleaning, and management.

Both of the studies presented above [2] and [4] investigated the stages involved in building ML systems. They also identified the practices and challenges that ML development teams have in large organizations. However, there is a gap in understanding how professionals develop ML systems in small and local companies. In our study, we want to understand more closely the needs and challenges faced by the developers in small companies or startups, and how we may support them. Our study aims at fulfilling this gap.

III. Methodology

A qualitative study was carried out to identify how teams on small companies develop ML systems and what are the challenges perceived when developing such systems. We collected data by using semi-structured interviews.

A. Participants and Context

The participants are software developers from three companies located in Manaus, Brazil. Table I details the main characteristics of each company. Altogether, we interviewed 7 professionals, as shown in Table II. All of them were working on the development of ML projects at the time of this study.

TABLE I. Company Characteristics

Comp.	*Description*	*ML Proj.*
Company A	It is a 7-year-old startup. This startup develops ML systems for debt renegotiation and online sales using self-service chatbot, employee and candidate turnover forecasting, and customer experience mapping across retail and hospital groups. At the time of the study, the startup was developing ML projects focused on retail solutions counting with four developers	B
Company B	It is a startup that started its activities in late 2018 and has been developing ML solutions focused on retail credit solutions. Its main products are ML systems for retail credit risk assessment. The ML solutions generate a probability that a customer may be indebted when their purchase history is evaluated for a period of six months. At the time of the study, had a team of five developers who work on these projects.	A
Company C	It is an agency that works directly for the State government. It is composed of a team of IT professionals who work with software development for more than 20 years. However, it started the development of ML projects three years ago, with two professionals in this area. The agency develops ML systems to support the supervision of electronic tax documents for Amazonas State Government.	A

Legend: **Comp**. – Company code; **ML Proj.** – number of projects focused on the development of ML systems: **A** – Up to 4 ML projects developed; and **B** – More than 4 ML projects developed

TABLE II. Summary of Participants

Company	*Company A*			*Company B*			*Company C*
Part.	P1	P2	P3	P4	P5	P6	P7
SW Exp.	4	10	7	4	8	3	25
ML Exp.	3	2	3	2	6	2	4

Legend: **Company** – Current company where the participant develops ML systems. **Part**. – Participant Code; SW **Exp.** – Years of experience in the industry in the development of traditional software; ML **Exp.**– Years of experience in the development of ML systems.

B. Data Collection

We carried out semi-structured interviews based on the following set of questions:

1) *Talk about your experience in ML projects.*
2) *When a ML project starts, what's the first thing you do? What tasks you perform?*
3) *How do you analyze customer data? What initial steps you perform to receive and analyze these data? Is there any rule or mandatory tasks?*

To meet the ethical requirements, we first explained the purpose of the research and the participants' rights through an informed consent form, guaranteeing the confidentiality of the data provided and the participant's anonymity. Also, we conducted the interviews individually, in a quiet and uninterrupted manner. During data collection, participants could talk about their tasks in the development of ML systems, and their practical experience. We performed the data collection in two phases. In the first phase, we interviewed the participants from Company A. We recorded and transcribed the interviews and started the data analysis process. Then, we interviewed the participants from the Companies B and C, and performed the whole analysis process again. The interviews resulted in 324 minutes of audio data (an average of 46 minutes per interview).

C. Data Analysis

We employed the coding procedures from the Grounded Theory (GT) method to perform the data analysis [10]. According to Strauss and Corbin [10], the researcher can use only some of the procedures to meet their research goals. Therefore, we employed open coding (1st step) and axial coding (2nd step) in our study. In the first step, we created the codes from the analysis of the interviews we did. In the second step, we grouped the codes by categories and subcategories and made the relationships between the created codes. Finally, our coding process was reviewed and discussed with another more experienced researcher until reaching consensus and the final results of the data analysis.

IV. Results

In this section, we report the answers to our targeted research questions that emerged from the data.

RQ1: How software developers build Machine Learning systems in small companies?

We categorized the stages used by the developers of the companies into four stages according to the categories and subcategories that emerged from the data: (1) Problem Understanding, (2) Data Handling, (3) Model Building; and (4) Model Monitoring. In Fig. 1 we show these stages and their respective tasks, that we describe in the following.

1) *Problem Understanding* – to start the project, the team needs to understand the problem and define the goals. The team assesses the business metrics that the customers already use to solve the problem and maps to ML goals. P2 (Comp. A), P5 (Comp. B) and P7 (Comp. C) reported these tasks:

*"The first thing we do is **understand the problem** you're going to address so you can model the data." (P5)*

*"We had a meeting with the client where we had a brainstorming meeting, understanding from the client what kind of **metrics** they already **use** for their problem."* (P2)

*"There are **business metrics**, defined according to the problem, that should be considered in the project."* (P7)

2) *Data Handling* – in this stage, developers conduct data handling to find out what data is necessary to meet the customer's goal. To start the data handling, the developers perform some tasks, such as data acquisition, data exploration, data structuring, and feature engineering. P4, P6 (Comp. B) and P7 (Comp. C) mentioned this:

*"First is the **data acquisition** part, we receive the raw data that the client sent, and there enter any type of extension, as database backups."* (P4)

*"The process of data science is the process of **discovering** the **data**, doing several **experiments** until you find the features that influence the outcome of the goal defined by the customer."* (P7)

*"Usually the bases we receive are not structured and we spend a lot of time **structuring the data**, making filters, adjustments to be able to use ML methods and algorithms."* (P6)

3) *Model* Building – in this stage the developers perform the ML model training and testing. Next, they evaluate the model. If the results are good enough, the model is deployed. Below are some quotations from the participants

*"After the exploratory analysis, then comes the **training** and preliminary **tests** with the **model**."* (P3)

"*We validate the model in **training** and **test**, we take a sample and train that sample of data (...) measuring the metrics of the model with this set of tests."* (P6)

*"After doing the analysis, optimizations, feature engineering, and testing, we look at the results and, if they are good, its time to **deploy** the model."* (P5)

4) *Model Monitoring* – in this stage, the ML model in the production environment is continuously monitored. Since new data is inserted into the ML model in the production environment, the developers keep monitoring the model performance, checking if it is necessary to change it or create a new one. In some cases, that means returning to earlier stages (model building or data handling). We present the quotations related to this stage below:

*"Since the model is in production, we need to decide whether to change it or whether it should be maintained, and there are **metrics** that **measure** the **quality loss** of the **model** and then need to retrain the model and then change it again or make a new model."* (P6)

*"Business metrics need to be monitored because there is a risk of entering new data that has not been learned by the ML model during training. **Monitoring** the **ML model** may mean that it goes back to the ML Development process to be readjusted and then goes through the whole process again until the evaluation of the model."* (P7)

RQ2: What challenges are perceived by developers during the development of ML systems in small companies?

Based on the understanding of the ML system development process stages uncovered in RQ1, in this subsection we answer RQ2. We found three major challenges mentioned by developers during our interviews.

a) Identifying business metrics is not trivial - in the initial stage of "Problem Understanding," developers need to identify what the customers' business metrics are. However, performing this task is challenging, as stated by P5, when we asked *"how do you identify customer's business metrics? That's a challenge."* Still, the customer wants to have policies to improve their business, but does not understand what metrics and data are required to do so, as evidenced by participants P4 and P6.

"The customer wants to have a credit policy, but he has no idea how to do it, what kind of data he should need to do it. And sometimes he does not even understand the data he needs to do it (...)" (P6)

*"When the customer does not have a metric, how do you do it? I believe this is a **problem** that we still have until today: **metrics**. Because we have academic metrics that are already known in the literature that we use, but*

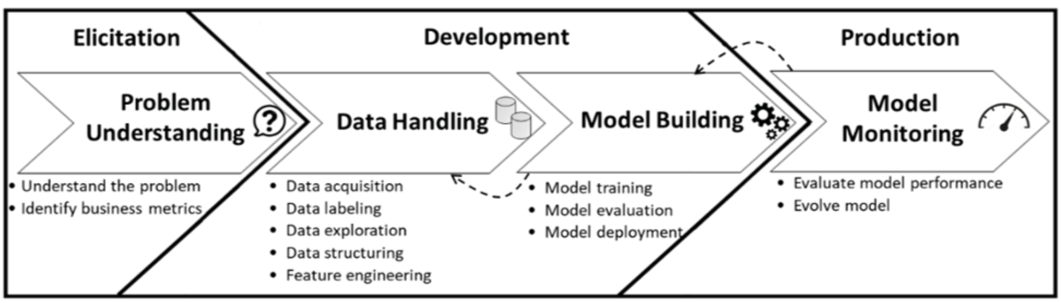

Fig. 1 The four stages of the ML system development process. At each stage, we list the tasks performed by the developers. The dotted arrows illustrate that one can go back to the earlier stages.

to show this to the customer is difficult, because he often does not understand it." (P4)

b) Undefined process – during the "Data Handling" stage, the developer performs various tasks including data preprocessing, which entails checking missing data, verifying inconsistencies, performing feature engineering. As stated by P5: *"At Feature engineering stage, it is important to have insights. Because we know that if we do not do anything in some attributes, the model should discard these attributes in the next stage."*

As stated by P5, failure to perform these tasks can result in poor model and performance. Since all companies do not have a defined development process for ML systems, each developer may or may not do the Data Handling tasks, as reported by P3: *"This part is very handmade. There are projects that I do one thing, and there are others that I do not do, simply because I* ***forgot*** *it. If I had a defined checklist, it would help, and there is* ***no well-defined process*** *here".*

c) Difficulty to design the database structure –In the "Data Handling" stage, developers also structure the data. Developers have reported that this task is a major challenge because it requires time and technical knowledge, and it is initially a manual process. As quoted by participants P5 and P6 (Comp. B):

"The greatest difficulty is always the base. It is hard to get a ***well-structured base****. Generally, the bases are not, and we spend a lot of time structuring the data, doing filters, adjustments, to be able to use ML methods and algorithms."* (P5)

"This part takes more time because we need to ***structure the data*** *right or else the model will not be able to make a good prediction and will end up failing (...) But if the data is in the correct format, it becomes automatic and if it arrives non-standardized or wrong, you should need to adjust it manually"* (P6)

V. Proposal to support the development of ML Systems

In this section, we answer the last research question (RQ3) based on the understanding of the challenges faced by professionals in RQ2. We developed two checklists to support developers to overcome the three challenges encountered. The first checklist (CheckBM) was designed to support the challenge of *identifying business metrics*. The second checklist (CheckDP) was designed to support the challenges of *undefined process* and *design the database structure*, related to the "Data Handling" stage. We emphasize that the tasks and criteria included in each checklist are based on the qualitative analysis of the data. Thus, the criteria included in the checklist are a combination of the practices used and know by some developers we interviewed. Also, we used the literature [12, 13] as the basis for building the checklists. The full version of the checklists can be found in the technical report [14].

RQ3. Is it possible to help the developers overcome these challenges?

A. Checklist to support Business Modeling (CheckBM)

We designed CheckBM to support developers overcome the challenge related to "*identifying business metrics*". We begin by organizing the CheckBM to match the tasks performed by the developer at this stage. We created a set of verification criteria for each task based on the qualitative data. In Table III, we show an extract in which the "Task" column identifies the name of the task executed by the developer, and the "Criteria" column contains the items that the developer should check for each task. We considered some rules proposed by Zinkevich [13] to support the construction of CheckBM. Table III presents an extract of our checklist for the *identify business metrics* task.

TABLE III. Example of the Checklist items to Support in Business Modeling

Task	Criteria
Identify business metrics	1) Check if the customer already has some kind of metric or established rule that s/he use to solve his/her problem. If so, which one? a) Describe the heuristic that the client uses b) Identify how the client applies the heuristics c) Create scenarios in which the heuristics are applied by the client d) What percentage is accepted by the customer?

B. Checklist to support in Data Processing (CheckDP)

We designed CheckDP to support the *undefined process* and the *difficulty of designing the database structure*. CheckDP was developed to support the *structuring the data* task previously presented in RQ1. In this stage, the developer performs different subtasks (*e.g.*, verifying that the data is complete). Thus, we create the verification criteria based on each subtask identified in our qualitative analysis. In Table IV, we present "missing data" subtask as an example. The "Criteria" column shows a description of the items that need to be verified by the developer. In addition, we relied on Witten et al. [12] about how academics and professionals structure the data, to help us defining the subtasks. In practice, when a developer finds an issue related to a criterion, for instance, "missing data in the database," the developer should take different actions to handle the problem. CheckDP suggests some possible actions to be performed. One of the actions is, for example, to verify if it is possible to acquire data to populate instances with null values.

TABLE IV. Example of the Checklist items to support structuring data

Subtasks	Criteria
Incomplete Data (Missing Data)	2.1. Check, in the database, how many fields are missing data, i.e., how many instances are with null values 2.2. Check if fields that are missing data may be kept or removed from the database

According to the stages presented in Fig. , the developer can use CheckBM to support the "Problem Understanding" stage and CheckDP to support the "Data Handling" stage. However, not all of the tasks we report at each stage can be performed on all projects. Therefore, it is up to the developer to choose at which stage each checklist can serve as support. We used the checklist format because it has been proven to be useful in SE activities like inspection, and tends to assist less experienced professionals in this activity [11].

C. Evaluation of the Checklists with Focus Group (FG)

To evaluate the initial version of the checklists, we conducted a Focus Group (FG). The goal was to assess them regarding the applicability and benefits for characterization purposes, regarding the context of ML systems, from the point of view of software developers. For the FG session, we selected two participants who would be available to attend to the session (P1 and P8). P8 was not interviewed previously (Table II), but he was invited to participate in this section to bring an external perspective. P8 works for Company A and has 2 years of software development experience in addition to 1.5 years developing ML systems.

The FG was conducted in two phases, with a duration of 1.5 hour each. In the first phase, the "CheckBM" checklist was evaluated, followed by the "CheckDP". At each moment, the participants received the checklist and took approximately 15 minutes to go through it. The moderator then encouraged them to talk and discuss their point of view on the following topics:

a) Applicability - applicable or not applicable to its context and why?

b) Benefits (Utility) - what benefits are applicable to your context, or does it not benefit your context, and why?

The moderator recorded all the interactions. Participants shared experiences about the applicability and benefit that each checklist brings to their context, as described below:

i. Checklist to support Business Modeling (CheckBM)

For the **(a) applicability** of this checklist, the participants stated that it should be used at the beginning of the project.

*"At the **beginning of the project**. Because this is the starting point for the beginning of the project."* (P1 and P8)

*"The **client metric** (business) has to be **defined at the beginning.** (...) the client says, we have this today and you can improve it? And then we have to try to make a model that should get many more features, more insights than the customer is doing manually."* (P8)

Regarding the **(b) benefits**, the participants made the following comments: *"The benefit of having a metric set at the beginning of the project is that you're going to have **something to measure**. Is the client's business doing well? For example, it's 50% of something the customer already has, but the model is making 60%."* (P8)

ii. Checklist to support in Data Processing (CheckDP)

For the **applicability** of the second checklist, P1 mentioned that it should be applied in data pre-processing stage, i. e. before training the model: *"In the data **pre-processing** stage, i.e., before applying tests and training the model. In fact, you do this before and then use it in the model."* (P1)

Regarding the **benefit** of this checklist, P1 found it interesting to improve the visibility of the work, avoiding forgetting some steps: *"(...) is interesting and helps the team to know where they are working, and this **facilitates** the **visibility** of their **work**, of the team, this is something very positive. And it also helps you to **not forget** a few **stages**, besides **maintaining the organization**."* (P1)

VI. Discussion

A. Stages in the ML system development

The development of ML systems starts with the "Problem Understanding" stage, similar to the "model requirements" and "problem definition" stages reported by [2] and [4] respectively, as they show that the developers define which problem is addressed and the goal of the project should be achieved. We verified that the task "identify business metrics" at this stage may be identified by the roles of product manager and application engineer in [4]. In our study, we noticed that at least one startup has this role of product engineer performing this task.

In the second stage of "Data Handling", the tasks of data acquisition, data labeling, data exploration, and features engineering are performed so that the data is transformed to meet ML methods. These tasks are similar to the stages reported in [2] and [4], performed in separate stages. For example, the task of data acquisition evidenced in our study, in [2] this task is done in the stage "data collection" and in [4] in the stage "incoming data".

In the third stage "Model Building", the tasks of training, testing and evaluating the model are similar to the stages presented in [2] and [4]. In Fig. 1, we show that there are feedback loops since one may return to previous stages if the model does not meets the customer's goal. These feedback loops are also demonstrated by Amershi et al. [2] and Byrne [4]. For exemple, in [2] the return to previous stages may be after the stage "model evaluation". In [4] the return may be from the stage "model prototyping" for the first stage "problem definition" to review the definition of the problem and what is efficient for the client. Also, Byrne [4] presented additional stages before the model building, such as "build infrastructure", "production model" and "outgoing data". In the fourth stage "Model Monitoring," the model is already in production and there is a risk of entering new data that have not been previously trained. Therefore, the performance of the model is monitored, and it's possible to return to previous stages, such as "Data Handling", to improve or make a new model. This stage is similar to the stages of [2] and [4]. For Byrner [4], model monitoring is done during the "evaluation" stage and may return to the "production model" or "define success metrics" stages. Despite having few differences between the works mentioned, the stages described have in common the essence of the data-centered ML system development process and also the feedback moments between the stages. Although the number of stages in small companies seems smaller, the developer performs tasks similar to the previous works. For instance, in the stages of data handling and model building, the tasks are similar.

B. Challenges

We discuss the difficulties of our participants about the three major challenges they reported during the development of ML systems and how they deal with these challenges. In addition, some of the related work's challenges are compared with the evidence found in this study.

The first challenge is concerned with the difficulty of "*identifying business metrics.*" In this study, we observed that business metrics may not be well defined for some customers, and, in this case, it's up to the development team to identify them. This may pose additional challenges such as searching for related works, identifying similar issues and verifying which type of metrics should be used. In addition, we note that combining the metrics identified with the customer's business objective is a challenge, similar to [4]. Making this combination is important to allow the team to create ML models and come to valid conclusions for the customer.

The second challenge concerns the "undefined process" to support developers during the ML system development process. For example, when structuring the data, the developers need to perform some subtasks, such as checking if the data is incomplete, duplicate, inconsistent, and so on. This challenge may occur because less experienced professionals do not have a guide that support them throughout the subtasks that need to be performed during this task. To address this challenge, our participants research examples, techniques, and methods of data structuring used in the literature or by other professionals from other companies. Amershi et al. [2] show that to deal with this challenge, Microsoft issued a set of principles around uses of artificial intelligence in the open world. All teams at Microsoft have been asked to align their engineering practices and the behaviors of fielded software and services in accordance with these principles.

The third challenge concerns the "*difficulty to design the database structure*", it is similar to the challenges cited by [7] on "data quality" and "data preparation". This is closely related to the data structuring task, performed during the data handling stage. We found that this task takes time, and this can occur because the database provided by the client does not have the structured data. Therefore, the developers need to structure the data manually. This task was the most reported challenge, regardless of the developers' level of experience. In addition, the data availability challenge reported in [7] is a challenge for our participants, because in smaller companies there are security policies that do not allow developers to access the database. To address this challenge, our participants need the support of a person who is experienced in the domain to provide help with the database. In larger companies, this challenge may be overcome by teams that have prior knowledge and access to the database [4, 7].

Finally, we found that the professionals of the three companies assume more than one role. For example, the person who analyzes the data is the same person who collects the data, and will be the same person who will develop ML solutions, while in large companies, this scenario is different, because there may be specialized teams or roles that act in each stage of the project [4, 7].

C. Limitations

The participants of our study work in three companies that develop ML solutions. Among these companies, two startups develop ML solutions for retail, and one develops ML solutions to support financial management. The experience of professionals from other companies developing ML solutions in other problem domains may be different. Another threat is that perceptions of the participants could be biased towards their own beliefs. These beliefs could cause some distortions when interpreting reality. To reduce this threat, the chosen software developers were those who had more experience in their organizations. Also, the participants represent a small sample, which limits the conclusion of our results. Another threat we considered was the questions in our interview script, initially emphasized the initial steps, and may have influenced our results. However, to reduce bias, we included open-ended questions that allowed the developer to report on the activities performed in the ML system development process from start to finish.

VII. Conclusion

In this paper, we performed a qualitative study in three companies located in Manaus, Brazil. We interviewed the professionals of these companies to understand how they develop ML systems. We identified the main stages and tasks that these professionals perform in their projects, as well as the challenges they face, and if these can be minimized with the use of our proposed two checklists (CheckMB and CheckDP). Besides, we evaluated our checklists by conducting a focus group to assess the applicability and benefits brought to the ML project in the company.

We found that the ML systems' development process of these companies follows, in general, four stages: (i) Understanding Problem; (ii) Data Handling; (iii) Model Building; and (iv) Model Monitoring. Compared with the stages presented in [2] and [4] we notice that the stages are similar, even though our number of stages is smaller. However, the tasks performed by the developers of our study in the data handling and model building stages are similar to the stages presented in [2] and [4].

The main challenges faced by the professionals we interviewed are: (a) difficulty in identifying business metrics, (b) lack of a defined process, and (c) difficulty to design the database structure. We realize that these challenges can impact the outcome of the project. For example, the ML model may perform poorly because it does not have enough data to achieve good results, or the data does not show quality (inconsistency, incompleteness).

Based on the understanding of the challenges evidenced by the professionals, we developed two checklists and evaluated with a focus group. We assessed the applicability and benefits of using these checklists in the development of these systems, and whether they can support developers overcome the challenges. Regarding applicability, CheckMB is applicable at the beginning of the project in the understanding problem stage and can assist in identifying business metrics. CheckDP can help in data structuring in the data handling stage, the visibility of the progress of the development stages and keeping the project organized.

With the increased development of ML systems in industry, it has become necessary to discuss the SE practices and methodologies for ML systems. Prior work by various researchers [2, 4, 7, 8] has shed light on some aspects. This paper is an initial step towards supporting and suggesting checklists, which serve as support in the development of ML systems. As future work, we consider applying the checklists in the ML project in industry. We intend to evolve our checklists and assess whether the challenges faced by developers may be overcome in practice. Besides, we will observe at the actual stages of an ML project, and show in more detail the artifacts generated. Finally, we intend to evaluate how the cognitive dynamics occurs in the development teams of ML systems.

Acknowledgment

We would like to thank the financial support granted by CNPq through processes number 430642/2016-4, 423149/2016-4, 311494/2017-0, and 204081/2018-1/PDE. FAPESP through process number 2015/24527-3 and CAPES - PROAP 001 and 175956/2013, National Science Foundation through number 1900903 and FAPEAM through process number 002/2018 POSGRAD 2018 – UFAM.

References

[1] H. H. Altarturi, K. Y. Ng, M. I. H. Ninggal, A. S. A. Nazri and A. A. A. Ghani, "A requirement engineering model for big data software," Int. Conf. on Big Data and Analytics (ICBDA), Beijing, China, pp. 111-117, Mar 2017.

[2] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi and T. Zimmermann, "Software Engineering for Machine Learning: A Case Study," Int. Conf. on Software Engineering (ICSE), Montréal, QC, Canada, pp. 291-300, May 2019.

[3] E. Brynjolfsson and T. Mitchell, "What can machine learning do? Workforce implications," Science, v. 358, n. 6370, pp. 1530-1534, 2017.

[4] C. Byrne, Development Workflows for Data Scientists, 1st ed., O'Reilly Media, 2017, pp. 1-28.

[5] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, v. 349, n. 6245, pp. 255-260, 2015.

[6] F. Khomh, B. Adams, J. Cheng, M. Fokaefs and G. Antonio, "Software Engineering for Machine-Learning Applications: The Road Ahead," IEEE Software, v. 35, n. 5, pp. 81-84, 2018.

[7] K. Miryung, T. Zimmermann, R. DeLine and A. Begel, "Data Scientists in Software Teams: State of the Art and Challenges". IEEE Transactions on Software Engineering, v. 44, n. 11, pp. 1024-1038, 2017.

[8] L. A. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining process models," The Knowledge Engineering Review, v. 21, n. 1, pp. 1-24, 2006.

[9] P. Louridas and C. Ebert, "Machine Learning," IEEE Software, v. 33, n. 5, pp. 110-115, 2016.

[10] A. Strauss and J. Corbin, Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 2nd ed., SAGE Publications, London, 1998.

[11] S. Tiwari and A. Gupta, "A systematic literature review of use case specifications research". Information and Software Technology, v. 67, pp. 128-158, 2015.

[12] I. Witten, F. Frank, M. Hall and C. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Morgan Kaufmann, pp. 285-334, 2016.

[13] M. Zinkevich, Rules of machine learning: Best practices for ML engineering. Google.[Online]. Available: https://developers. google. com/machine-learning/guides/rules-of-ml. Accessed: June 1st, 2017.

[14] E. Nascimento, I. Ahmed, E. Oliveira, M. P. Palheta, I. Steinmacher, T. Conte, Technical report available at: <https://doi.org/10.6084/m9.figshare.8309561.v1>.