Brazilian Data Scientists: Revealing their Challenges and Practices on Machine Learning Model Development

João Lucas Correia[1], Juliana Alves Pereira[2], Rafael de Mello[3], Alessandro Garcia[2], Baldoino Fonseca[1], Marcio Ribeiro[1], Rohit Gheyi[4], Willy Tiengo[1], Marcos Kalinowski[2], Renato Cerqueira[5]
[1]Federal University of Alagoas (UFAL), Brazil; [2]Pontifical Catholic University (PUC-Rio), Brazil; [3]Federal Center for Technological Education (CEFET/RJ), Brazil; [4] Federal University of Campina Grande (UFCG), Brazil; [5]IBM Research

ABSTRACT

Data scientists often develop machine learning models to solve a variety of problems in the industry and academy. To build these models, these professionals usually perform activities that are also performed in the traditional software development lifecycle, such as eliciting and implementing requirements. One might argue that data scientists could rely on the engineering of traditional software development to build machine learning models. However, machine learning development presents certain characteristics, which may raise challenges that lead to the need for adopting new practices. The literature lacks in characterizing this knowledge from the perspective of the data scientists. In this paper, we characterize challenges and practices addressing the engineering of machine learning models that deserve attention from the research community. To this end, we performed a qualitative study with eight data scientists across five different companies having different levels of experience in developing machine learning models. Our findings suggest that: (i) data processing and feature engineering are the most challenging stages in the development of machine learning models; (ii) it is essential synergy between data scientists and domain experts in most of stages; and (iii) the development of machine learning models lacks the support of a well-engineered process.

KEYWORDS

Software Engineering, Machine Learning, Practitioner, Empirical Study

1 INTRODUCTION

The adoption of Machine learning (ML) models has been growing as the intelligence behind software systems. These models are used for solving specialized problems in several domains. In the oil and gas industry, ML models are used for mitigating environmental disasters [2, 11]. Governments have been using ML models for monitoring socioeconomic development [14]. In the Ecology domain, ML models have been used to classify animal species [17, 18].

ML models are developed for allowing programs to learn from previous experiences [9]. In this context, professionals working with ML modeling should define appropriate resources to train their models (e.g., data samples, features, algorithms, and parameters), optimizing the learning from different perspectives, and obtaining the model that best fits the desired solution. Typically, the professionals allocated to conduct this development are the data scientists [13]. From these professionals, it is required a multidisciplinary knowledge, including but not limited to data management, mathematics, and software engineering. Besides, they also often need to interact with customers and domain experts to understand the scope of the problem to be solved.

Data scientists may be served from different strategies to develop an ML model depending on the practical problem to be addressed. For instance, if the model's clarity and communication are the most important, it would require data scientists to use interpretable ML models (e.g., decision tree) independent of its accuracy. Otherwise, if the model's accuracy is more important, it would require data scientists to focus on tuning model parameters.

One might argue that data scientists would benefit from adopting classical software engineering disciplines (e.g., systems design, quality assurance, and verification) to properly build their models. However, ML modeling addresses a distinguished development paradigm, requiring proper tools and techniques [8, 19]. The lack of this support may lead data scientists, frequently having a limited background in software engineering, to improvise and frequently perform adhoc activities to overcome the particular challenges involved in engineering ML models. For example, the validation and verification process of stochastic models is challenging. Current tools and engineering disciplines do not fully support the verification and validation of code with random behavior [20].

In this context, we investigate the challenges and practices that emerge during the development of ML models from the perspective of data scientists. We focus our analysis on well-know stages present in the ML workflow [1, 5, 12]: *Model Requirement*, *Data Collection*, *Data Cleaning*, *Data Labelling*, *Feature Engineering*, *Model Training*, *Model Evaluation*, and *Model Deployment*. More specifically, we investigate which stages are considered more challenging by data scientists, and which are the common practices adopted by them to deal with the corresponding challenges.

To perform our study, we conducted individual semi-structured interviews [4] with data scientists from different Brazilian companies. These professionals are experienced in developing back-box and white-box ML models in three main domains: oil and gas, government, and natural resources. We transcribed the interviews and coded it using the open-coding methodology [16]. Next, we grouped these codes into five categories: actor, activity, method, limitation, and challenge. We rely on the categories' limitations and challenges to point out problems and issues faced nowadays by data scientists that deserve attention from the research community. Then, we rely on the categories activity and method to understand common practices followed by data scientists when developing ML models.

The findings of our study indicate that data scientists perceive the *Data Processing* and *Feature Engineering* as the more challenging stages of ML model development. However, data scientists also reported important issues addressing other stages. In general,

SBQS'20, December 2020, Online
2020. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the challenges reported indicate the need to enhance the synergy between data scientists and the other actors involved, including domain experts, customers, and project managers. Besides, we also found that the practice of ML development lacks the support of a well-engineered process. For instance, we did not find in the interviews mentions about techniques for assuring the traceability between the features and the model requirements. Also, the validation of the ML model is often not performed given the difficulty to test back-box ML models. These characteristics reflect on the recurrent rework and on the strong and continuous dependence of data scientists to the domain experts. Thus, we understand that the findings of our study may support future research on designing a comprehensive and engineered process for supporting data scientists on developing ML models.

The remainder of the paper is organized as follows. Section 2 introduces background concepts and discusses the related work. Section 3 introduces our research questions and methodological steps. Section 4 discusses the execution of our methodological steps. Sections 5 and 6 describe our findings and discuss our results, respectively. Section 7 describes the threats to validity. Finally, Section 8 concludes our work.

2 RELATED WORK AND BACKGROUND

Recent studies [1, 8, 10, 19, 21] have indicated a difference between the challenges faced by ML practitioners (in our research, data scientists) and non-practitioners.

Nguyen et al. [10] reported an exploratory study across seven companies. They investigated how software engineering processes and practices can be applied to develop systems based on Artificial Intelligence (AI). Their findings revealed that particularities of AI-based applications, such as the uncertainty of predictions, hinder the adoption of traditional software engineering guidelines during the development of those applications, showing different development approaches. In this context, our study focuses on practices and challenges faced by data scientists in ML models development. Besides, we explore companies with development goals driven by business and/or research.

Amershi et al. [1] reported a study involving AI professionals at Microsoft. The authors investigated scientists, researchers, managers, programmers, and other professionals in their daily activities. They observed three major challenges in building large-scale AI applications: *data management*, *reuse*, and *modularity*. Amershi et al. [1] concluded that building AI systems require more effort and expertise from professionals. In this sense, our paper aims to investigate in-depth the efforts of data scientists professionals in ML from five different companies and three different domains.

Wan et al. [19] investigated how the adoption of ML frameworks affects software engineering practices. The authors performed an empirical study through interviews and a survey. They showed statistically significant differences in software engineering practices and teamwork characteristics. They suggested that most of the differences come from the uncertainty present in the inherent randomness of the data and ML algorithms. Notice that the authors explore frameworks employed for ML, while our study examines the whole development process.

We found recent studies [8, 20] investigating and identifying state-of-the-art approaches for supporting the engineering of ML systems. Masuda et al. [8] conducted a survey to discover software engineering approaches to support the development and assure the quality of ML systems. They analyzed 78 research papers and pointed out that existing software engineering practices may be inappropriate to deal with ML's uncertainty. Zhang et al. [20] investigated 138 research papers looking for approaches to test and debug the ML code. Their findings indicated that only a few contributions focus on testing interpretability, privacy, or efficiency. Zhang et al. [20] focus on analyzing exclusively the stage of *Model Evaluation*, instead we target all ML stages.

All these studies [1, 8, 10, 19, 21] highlight the need for improving the ML development process. In this context, a few recent works [1, 3, 7, 15] have been proposing ML workflows to guide ML developers. These workflows are composed of typical stages addressing the development of ML models. Among them, the workflow proposed by Amershi et al. [1] is the most recent work found in the literature, presenting the most comprehensive and up to date workflow. This workflow is composed of the following stages:

- *Model requirements.* This stage establishes the basis for an agreement between stakeholders about how the ML model should work. In this stage, stakeholders decide which data to consider and what types of ML models are most appropriate for the given problem.
- *Data processing.* This stage involves collecting all relevant data; cleaning biased and irrelevant data; and data labelling (for supervised learning).
- *Feature engineering.* This stage covers the process of modifying the selected data (e.g., by encoding features and extracting new features) in order to better suit the particularities of the chosen model and improve its accuracy.
- *Model training.* In this stage, the ML model chosen is trained and tuned on the (labeled) data.
- *Model evaluation.* In this stage, metrics are used to evaluate the created model on new (non-labeled) data.
- *Model deployment.* This stage covers the deployment of the model in the customer environment and the activities performed for monitoring and maintaining the model.

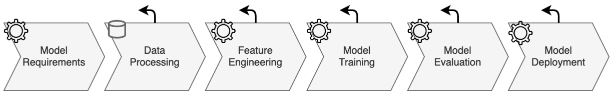

Figure 1: Machine Learning modeling workflow. Adapted from [1].

In this work, we used the Amershi et al. workflow (Figure 1) as a common reference to understand the ML modeling process followed by the data scientists interviewed. For this propose, we introduced the workflow to the data scientist at the beginning of each interview, asking him/her for identifying to what extent the workflow addresses his/her developing practice (see Section 4). The arrows above each stage in Figure 1 indicates the possibility of a *feedback loop* across the stages.

3 STUDY SETTINGS

Our study aims at characterizing the main challenges faced by data scientists on developing ML models as well as the practices adopted by them to deal with these challenges. More specifically, we address the following research questions:

RQ_1 *What are the most challenging stages faced by data scientists on developing ML models?*

RQ_2 *What are the practices adopted by data scientists to deal with the most challenging stages?*

To answer these questions, we conducted an exploratory study based on semi-structured interviews. By challenging stages, we mean the stages with complex, time-consuming, and error-prone activities. By practice, we mean the activities performed by data scientists, as well as the tools and methods they use to perform these activities.

To address RQ_1, we started the interviews by stimulating data scientists to reflect on the stages of the ML workflow (see Figure 1). Then, we asked data scientists to objectively indicate which stages they consider more challenging and why. To address RQ_2, we applied open questions to stimulate the participant to describe in detail how they perform the more challenging stages. After coding all the answers given by the data scientists, we identified code categories and distributed the coded data among these categories.

Target population and sample. This study's target population comprises experienced data scientists in developing ML models for complex and customer-oriented solutions. These models may be supervised or not. We recruited a small but diverse sample of eight data scientists from five different companies (see Table 1). The participants' experience with ML ranges from two to 40 years. Besides, we selected participants working in three different domains within our industrial collaboration network: three data scientists ($d1$, $d2$, $d4$) from the oil and gas domain; three data scientists ($d6$, $d7$, $d8$) of natural resources; and two ($d3$, $d5$) data scientists working for the government. Although we do not directly explore projects in the academy, six of the eight data scientists ($d3$, $d4$, $d5$, $d6$, $d7$, $d8$) are postgraduate students, having experience in using ML models for research. The frameworks for developing ML models frequently employed by the data scientists are TensorFlow[1], Keras[1], Scikit-learn[1], and PyTorch[1].

Instrumentation. All the authors of this paper were involved in the design and validation of the interview questions. Two authors described the interview questions and protocol, and the others validated it. At the beginning of each interview, the interviewer introduced to the data scientist the workflow that we centered our research (Figure 1). Then, we asked the participant to report any difference in this workflow with his/her practical experience. Once we focus on the high-level stages, data scientists did not have any disagreement with the workflow or with the terminologies used. Next, we applied *general questions* and *specific questions* regarding each stage of the workflow, as described in Listing 1. The interview' *general questions* aim at exploring the background of the data scientist, the role of people in his/her team, and how data scientists execute and verify activities in a given stage. Besides, we also ask the data scientists which workflow stages they perceive as most challenging. Otherwise, the *specific questions* explore aim to explore in more depth the activities performed by the development teams in the most challenging stage(s). With this, we aim to (1) optimizing the interviews' time, limited to 1 hour, and (2) gathering more reliable data for supporting our analysis.

The complete form with all the specific and general questions is available in our supplementary material [6]. Notice that a few additional questions appeared according to the flow of each interview.

Listing 1: General and Specific Questions.

1. **General Questions**
 1.1. Background information
 1.2. Who is involved at each stage
 1.3. Activities to accomplish a stage
 1.4. Stage(s) which the data scientist has more expertise[2]
 1.5. Most challenging stage(s)
2. **Specific Questions**
 2.1. *Model Requirements*
  2.1.1. Requirements specification
  2.1.2. Functional and nonfunctional requirements
  2.1.3. Completeness, correctness and testability
  2.1.4. Verification and migration
 2.2. *Data Processing*
  2.2.1. Data Collection
  2.2.2. Data Cleaning
  2.2.3. Data Labeling[3]
 2.3. *Feature Engineering*
  2.3.1. Feature selection and transformations
  2.3.2. Importance of an expert in this stage
 2.4. *Model Training*
  2.4.1. Training, testing and validation sets
  2.4.2. Algorithms and hyperparameters
  2.4.3. Data drift
 2.5. *Model Evaluation*
  2.5.1. Metrics for evaluation
  2.5.2. Importance of an expert in this stage
  2.5.3. Overfitting, underfitting, robustness
  2.5.4. Interpretability
 2.6. *Model Deployment*
  2.6.1. When a model is ready for deployment
  2.6.2. Monitoring data and model quality
  2.6.3. Maintenance

4 EXECUTION

The eight interviews were taken between November 2019 and April 2020. Each interview took, on average, 45 minutes. To be able to collect different viewpoints and perspectives, two researchers conducted each interview. One researcher played the role of the main interviewer, applying the planned questions. Another researcher predominantly played the role of observer, taking notes

[1] https://tensorflow.org/, https://keras.io/, https://scikit-learn.org/, https://pytorch.org/

[2] To measure the expertise, we consider the number of projects or the time the practitioner worked on that stage.

[3] Only for supervised learning.

Table 1: Characterization of the data scientists.

Company	Projects Context	Data Scientist	Experience	Technical Background		Frameworks
$c1$	Oil and gas	$d1$	2 years	Supervised Learning	Image Detection	TensorFlow Keras
		$d2$	40 years	Supervised Learning	Image Detection	TensorFlow Keras
$c2$	Government	$d3$	12 years	Supervised Learning	Discriminative models	TensorFlow Keras Scikit-learn
$c3$	Oil and gas	$d4$	4 years	Supervised Learning	Discriminative models	Scikit-Learn
$c4$	Government	$d5$	3 years	Unsupervised Learning	Clustering	Scikit-Learn
$c5$	Natural resources	$d6$	4 years	Supervised Learning	Knowledge Representation	TensorFlow Scikit-Learn PyTorch
		$d7$	4 years	Supervised Learning	Image Detection	TensorFlow
		$d8$	12 years	Supervised Learning Unsupervised Learning	Image Detection	TensorFlow Scikit-Learn

about the participants' behaviors and asking additional questions to gather more information.

4.1 Context of the Answers Given

All the study participants found correspondence between the workflow introduced by the interviewers (Figure 1) and their practice, confirming that their development activities follow the same flow: *definition of model requirements*, *data processing*, *feature engineering*, *model training*, *model evaluation*, and finally *model deployment*. Besides, most of the participants argue that several iterations may happen during the development of ML models, mainly due to issues lately identified. For instance, data scientist $d4$ describes cases in which was necessary to go back in earlier stages of the workflow.

> *"Very often during model training we identify problems having to return to the stage of feature engineering. Similarly, it also occurs in the stage of model evaluation, once the resulting model presents bad accuracy we have either to reprocess the data or reengineer the features." - (Data scientist d4)*

Our interview was composed of general and specific questions (see Section 3). We applied the general questions to all stages where the data scientist was experienced, and the specific ones to the most challenging stage(s). Table 2 presents the distribution of the general and specific questions answered by the data scientists.

One can see that none of the data scientists answered general questions about *model requirements* once they are not experienced in this stage. Besides, they did not classify this stage as challenging. However, it does not mean that the participants not provided information about this stage. For instance, $d8$ described how requirements for a model arises during the model idealization.

> *"I think the way we process the data and engineer the features comes a lot from the customer (the stage of Model Requirements). What does the customer want? What task would he like to solve applying ML? What is the problem he is facing [...] I think there is a lot of customers' input here." - (Data scientist d8)*

Table 2: Distribution of the data scientists' answers by stage.

Stage	General Questions	Specific Questions
Model Requirements		
Data Processing	$d1$,$d2$,$d3$,$d4$,$d8$	$d1$,$d2$,$d3$,$d4$
Feature Engineering	$d2$,$d3$,$d4$,$d6$,$d7$	$d2$,$d3$,$d4$,$d8$
Model Training	$d1$,$d5$,$d6$,$d7$	$d5$,$d6$,$d7$
Model Evaluation	$d6$,$d7$	
Model Deployment	$d5$	

4.2 Generated Codes

During data analysis, the first author performed the transcription of the recorded interviews. Then, he conducted the *open-coding* process [16] of each quotation (raw transcription) to support the analysis. The open-coding process, was executed as illustrated in the example that follows:

> **Raw Transcription:** *"Geologists and geophysicists bring knowledge in geology and geophysics. They often warn that data-driven solutions, merely based on data, are minor problems once there is a whole physics background supporting this data."*

Table 3: Example of codes from the raw transcription.

Code	Content
Code 1:	domain expert presence
Code 2:	domain expert indicates rules that must be respected
Code 3:	data-driven solutions are not always possible
Code 4:	the model must respect domain particularities

After coding the whole raw data (see example in Table 3), the first author analyzed the codes, aiming at identifying an initial set of categories for grouping these codes. After discussions and refinements on initial codes and categories, all authors reached a consensus for establishing the final set of categories. Then, the first author redistributed the codes into the new set of categories. Next, another author who did not attend the interviews, validated the codes and their corresponding categories. They disagreed in 43 of 447 codes (9.6%). Then, we allocated a third author for solving the points of disagreement. At the end of the process, it has emerged the following categories:

- *actor*: person or team involved in a stage;
- *activity*: process or task performed by an actor;
- *method*: tools or methodology used by the actor;
- *limitation*: expected behaviors or activity limitations;
- *challenge*: challenges faced by an actor.

In particular, codes from Table 3 were classified in the following categories: Code 1 ⟶ *actor*; Code 2 ⟶ *activity*; Code 3 ⟶ *limitation*; Code 4 ⟶ *limitation*.

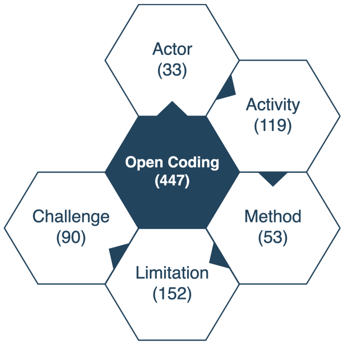

Figure 2: Categories emerged from open-coding with their frequencies.

In total, the analysis for all interview data resulted in 447 codes, (see Figure 2) distributed into the five categories: actor (33); activity (119); method (53); limitation (152); and challenge (90). With the aim to access a diverse number of interviews, in the analysis process the authors ponder three factors:

(1) The codes convergence to a similar main set of categories.
(2) The amount of data to answer our research questions (i.e., did all available interviews cover our research questions?)
(3) The availability (1-hour interviews) and access (industry partnership) to interviewees.

To provide some rationale about the context of the codes, Table 4 shows examples of code found by category. Our analysis shows that even the answers of the most and the least experienced data scientists, respectively *d*2 and *d*1, do not diverge from the central topics of discussion addressed by other interviews. All codings and interview transcriptions are available at our supplementary material[4].

After establishing the categories and distributing the codes among them, we mapped each code with its corresponding workflow stage (Figure 1). As an example, Table 5 describes the main actors (data scientist, domain expert, customer, project manager and infrastructure team) involved in each stage. Codes without a clear and specific association with the workflow stages was classified as *general*.

Table 4: Examples of the categories emerged.

Category	Example of relevant codes
Actor	Presence of the client in Model Evaluation
Activity	Data scientist define hyperparameters according to his experience
Method	Use of framework for model development
Limitation	One data scientist responsible for all stages
Challenge	Uncertainty about model quality for real scenarios

5 REVEALING CHALLENGES AND PRACTICES

In this section, we answer the research questions of the study, revealing the challenging stages and common practices in the ML model development.

5.1 RQ$_1$: Most Challenging Stages

To answer this research question, we benefit from two distinct moments First, we analyzed the number of data scientists that considered each stage as more challenging. Second, we analyzed the distribution of codes categorized as challenges in each stage during the *open coding* process. In our classification, we consider the number of data scientists indicating a stage as challenging more important than the number of challenging codes per stage, since individual interviews may unbalance the code distribution. Table 6 shows the stages analyzed in our study, the number of participants that reported the stage as challenging, and the number of codes in the category *Challenge* associated with each stage.

We found that *Data Processing* and *Feature Engineering* are the two stages more frequently perceived as challenging. Six and seven participants reported these stages as challenging, respectively. Besides, we also found that the *Data Processing* and *Feature Engineering* are the first and third stages with more challenge codes associated. While the *Data Processing* challenges are more related to the difficulties in choosing proper data to compose the dataset, the *Feature Engineering* challenges are frequently associated with the management of features for training the model.

Table 6: Distribution of the challenges reported.

Stage	Data Scientists (Total)	Challenges Reported
Model Requirement		0
Data Processing	*d*1,*d*2,*d*3,*d*4,*d*5,*d*8 (6)	19
Feature Engineering	*d*2,*d*3,*d*4,*d*5,*d*6,*d*7,*d*8 (7)	23
Model Training	*d*1,*d*5,*d*6,*d*7,*d*8 (5)	22
Model Evaluation	*d*2,*d*4,*d*5,*d*6,*d*7 (5)	12
Model Deployment	*d*5 (1)	2

[4] https://github.com/sbqs2020/sbqs2020

Table 5: Participation of actors by stage from the perspective of the data scientists.

	Data scientist	Domain expert	Customer	Project manager	Infrastructure team
Model Requirements	*d*2,*d*3,*d*4,*d*6,*d*7,*d*8	*d*6	*d*5,*d*4,*d*6,*d*7,*d*8	*d*5	
Data Processing	*d*1,*d*2,*d*3,*d*4	*d*1,*d*3,*d*4,*d*6,*d*7	*d*8		
Feature Engineering	*d*1,*d*2,*d*3,*d*4,*d*5,*d*6,*d*7,*d*8	*d*1,*d*3,*d*4,*d*6,*d*7,*d*8			
Model Training	*d*1,*d*2,*d*3,*d*4,*d*5,*d*6,*d*7,*d*8				
Model Evaluation	*d*1,*d*2,*d*3,*d*4,*d*5,*d*6,*d*7,*d*8		*d*2,*d*4,*d*6,*d*7	*d*5	
Model Deployment	*d*5,*d*6				*d*4,*d*6,*d*8

Finding 1: *Data Processing* and *Feature Engineering* are perceived as the most challenging stages.

Besides, five participants reported the *Model Training* and *Model Evaluation* as challenging. While the challenges in *Model Training* has 22 codes associated, *Model Evaluation* has only 12 codes. In *Model Training*, the challenges frequently address concerns with time-consumption, the definition of artifacts such as parameters and algorithms, and the verification of the fitness of the chosen artifacts. The *Model Evaluation* challenges predominantly address the definition of the best metrics for evaluation and the model's quality assurance.

A single data scientist reported the *Model Deployment* as a challenging stage. He reported two challenges regarding the need to master specific technologies during deployment. To overcome this challenge, we observed that an infrastructure team frequently conducts the model deployment, as reported by data scientists *d*4, *d*6, *d*8 (see Table 5). Finally, none of the participants reported the *Model Requirements* as one of the most challenging stages. It was an expected result once although most of the data scientists recognized their presence in this stage (Table 5), most of them reported low experience on this stage.

5.2 RQ$_2$: Practices Adopted by Data Scientists

In the previous section, we identified that the *Data Processing* and *Feature Engineering* are the most challenging stages from the data scientists' perspective. Now, we identify how data scientists deal with these stages. To do so, we analyze the actors, activities, methods and limitations related to these stages.

5.2.1 Data Processing.

Actors. The data processing involves three distinct actors: *data scientist*, *domain expert*, and *customer*, as described in Table 5. The data scientist is the professional involved in all the activities related to the ML model development. The customer is a representative of the company or professional aware of the company's interests. Finally, the domain expert is a professional having sufficient scientific knowledge in the field of the application. For example, we observed development teams containing electrical engineers and geologists. Depending on the availability of data sources, there are different approaches to perform *Data Processing*. For instance, if the company has the data required to build an intelligent model, it is not necessary external sources and, therefore, data processing can be performed in-house (at the company).

The data processing in-house may be interesting, mainly in situations in which the company contains confidential content that cannot be shared, or even when the company has a large number of data stored in different ways, and not all data is relevant to build the model. In both cases, the data scientists need the support of the customer for data gathering, since they may not have access to the confidential content. Besides, in some cases, they may not be aware of the company's internal data infrastructure as well as the most pertinent data. Thus, the conduction of data processing would be very costly. In any case, the synergy between the data scientists, customers, and domain expert is needed to avoid unnecessary efforts, as described in the quotation below:

> *"The Data Processing stage ends up requiring a lot of inputs from the customer. Although the customer give us some instructions to collect the data, often the data are not in the right format (or in a good shape) for us to proceed, thus we often need several interactions to get the desired data." - (Data scientist d8)*

On the other hand, external data sources are used when the customer does not have the required data to build an intelligent model. In such cases, data processing is conducted by the data scientist and domain expert. The domain expert plays a crucial role, analyzing which data is relevant to meet the requirements. As we will see forward, one of the most challenges from data scientists is the demand for domain knowledge, which is mitigated by the participation of the domain expert in the activities. For instance, the subject *d*3 report difficulties in the data processing:

> *"Very often you will have to get the data from different sources and extract the relationship behind it, thus you will need to use your own conventions for that." - (Data scientist d3)*

Activities. Our results indicate that the first challenge activity of this stage is the *instance labeling* by the domain expert. In supervised models, the data scientist must provide a sample of instances and their labels to the training process. In more complex domains, such as natural resources, the labeling activity is not trivial, requiring several and detailed dataset analysis. For instance, the following statement was made by a data scientist working with image classification, he recognizes that he cannot label its instances due to the lack of expertise in the domain.

> *"I always need to label new images to train the classification model, however very often I do not know what they mean, so I heavily depend on a specialist who will annotate the images for me." - (Data scientist d1)*

Another challenging activity of data processing reported by data scientists is the *data enrichment*. Data scientists tend to perform data enrichment when they recognize that the number of data available lacks size and/or quality. We observe that this scenario is common in the context of projects involving image processing. In these cases, the data scientists reported that the small datasets lead them to artificially creating new images through techniques such as rotation, contrast increasing and noise addition.

Method. The unique data processing method reported by data scientists was the use of charts, such as box plots and histograms, for supporting the verification of data quality. The motivation behind the adoption of charts addresses the need for visual tools to avoid the use of inappropriate data. For instance, if a data scientist identifies an error in the dataset only further in the *Feature Engineering* stage, he/she needs to re-execute *Data Processing*, which implies in higher development costs.

Limitation. We found one main limitation for data processing; the data scientists depend on the domain expert and the customer. As shown in our results, data scientists believe that data processing considerably affects the next workflow stages. Therefore, aspects such as gathering relevant data, label data correctly, data quantity, and quality assurance should be well performed in this stage to mitigate the impact on the following stages.

Finding 2: The synergy between data scientists, domain experts, and customers during data processing is essential to perform data collection, data enrichment, and instance labeling. In this stage, charts are commonly used for the verification of data quality.

5.2.2 Feature Engineering.

Actors. In this stage, our results indicate the involvement of two actors: the data scientist and the domain expert (see Table 5). As previously reported, the data scientist is involved in almost all stages once he/she is responsible for building the entire intelligent model. The domain expert assists data scientists to better understand complex features composing the dataset.

Activity. Feature Engineering addresses preparing features and selecting those most relevant for the model training. Our results identify three activities at this stage. The first one is the *feature analysis*, activity in which data scientists explore the whole dataset aiming at characterizing feature's parameters such as value range, correlation, distributions, and independence degree.

The second activity is *data transformation*. It consists of executing operations over features for removing inadequate values, increasing representativeness, and converting types and values. For instance, neural network algorithms require numeral features with values in the range between zero and one. Thus, it may be necessary to convert the data type and value of some features. One special case of transformation is the feature combination, in which two or more features are combined into a new one more valuable for the learning process.

Data transformations are usually performed only by data scientists, except by the feature combination, typically performed by the data scientist with the assistance of the domain expert. It can be explained by the fact that identifying opportunities and performing these transformations requires understanding the semantics of features used and resulted from transformations.

The next activity is the *feature selection*. It identifies the best features from the entire dataset for training the model. Previously mentioned activities influence directly in the quality of feature selection, since observation in *feature analysis* and *data transformation* will impact the data scientists' judgment regarding the best features to train the model. The *feature selection* is conducted by the data scientist and the domain expert, since each one cumulate knowledge about features. We observe in our results participants saying that in some cases, the domain expert reveals excellent features according to his/her expertise. However, data scientists noticed some cases that only their feature analysis without the support of a domain expert revealed excellent features (not noticed by domain experts):

> *"For example, although the expert claims a feature is not that important, [...] after trying out the model with the feature it may turn out to be very valuable. Also, the opposite may happen. The algorithm shows us that a specific feature is not that important, but [...] it turns out to be very important for the domain expert, and we cannot simply ignore it." - (Data scientist d4)*

Methods. Our codings revealed two main methods for performing the *Feature Engineering*: the use of *statistical methods* in *data analysis*, and the use of *automatic feature selectors* in *feature selection*.

Statistical methods were mentioned as being widely used to assist the *data analysis* process, since they provide functions and tools for helping data scientists on observing the data behavior. Data scientists did not mention specific tools for data analysis. However, they reported the use of the Python language and its libraries, e.g. pandas [5], to run statistical operations.

According to the participants, the use of automated feature selector is associated with deep learning, because algorithms for this propose automatically learns the best features to problem solution and model training, discarding the need of the data scientists for performing this activity:

> *"It is worth mentioning that in today's deep learning era, we skip this stage (Feature Engineering), once we have a model that learns the features, and also learns the task." - (Data scientist d8)*

On the other hand, when algorithms for deep learning are not used, the developer performs the feature selection manually. In this case, the role of the domain expert is essential once he/she will point out the appropriate features to train the model based on ground truth. Although the domain expert knows the relevant features of the problem, the data scientist must execute operations such as feature scoring to ranking features based on relevance. It should happen once the building of ML models is an exploratory process, which may reveal new patterns or behaviors even not know previously by experts.

[5] https://pandas.pydata.org/

Limitations. Our findings show that limitations in this stage are related strictly to the need for understanding about the application domain, for this reason the domain expert is so essential:

> *"Feature engineering is a very complicated process. First, you have to understand the (application) domain and, after you understand this domain, you should identify those aspects that will make the statistical model more efficient in understanding the data representation." - (Data scientist d6)*

When data scientists talk about the *feature selection*, they frequently report being not sure on whether selected features are the best options to train the model, mentioning that the stage that indicates this feature adequacy is the model evaluation. Therefore, it suggests that at this point, any mechanism grants confidence to the data scientist about their work on *Feature Engineering*. The next quote exposes how data scientists deal with the correctness of *Feature Engineering*:

> *"Very often we have to turn again to the stage of Feature Engineering, especially when we are dealing with data that we do not know very well [...] We first evaluate the model and then if the accuracy is not good enough, we have to make changes in the way the features are being considered in order to improve accuracy." - (Data scientist d7)*

Finding 3: Data scientists and domain experts are the main actors involved in feature engineering, performing feature analysis, data transformation, and feature selection. Automatic feature selection is the most common method at this stage, but data scientists still need the domain expert to better understand the application domain.

6 DISCUSSION

Table 7 describes the central topics of discussion for each stage of the workflow of Figure 1. The table also presents the frequency of the central topics in each stage grouped by data scientists' experience with ML models (low experience and high experience). We consider as having low experience those data scientists with less than five years of experience in ML development. The data scientists with high experience are those having more than five years of experience.

First, we group similar topics from all coding referring to a single stage. Then, we produced a list of the most discussed topics. Table 7 presents the top-3 hot topics (except for the stage *Model Requirements* which had an insignificant number of discussions for a single another topic). Second, for all listed topics referring to each stage, we compute how often these topics are discussed by low and high experienced data scientists. As an example, notice in the first row of the table that the central topic discussed in the stage of *Model Requirements* was the *"Presence of internal and external actors"*. We observe that 67% of the comments were about this topic. From these comments, 58% were made by low experienced data scientists, while the others were made data scientists with high experience in ML development. Considering this summary, three things are worth noticing:

(1) Presence of internal and external actors. The presence of internal and external actors is essential in most of the stages, except for *Model Training*. In *Model Training*, data scientists write code, optimize model parameters, and train the model. Thus, this stage requires skills in ML algorithms, hence only data scientists act in this stage. However, for the remaining stages, we observe the need for synergy between the data scientists and the external actors.

(2) Challenging stages. The problems of the challenging stages address the need to understand the domain area of the ML model. Both most challenging stages (*Data Processing* and *Feature Engineering*) directly depend on the quality of the requirement specification, which strongly relies on the knowledge of external actors. Therefore, data scientists need to properly extract knowledge of the domain and the external actors' needs to perform a good requirements specification. However, data scientists often associate the need for this interaction with problems. For instance, the quotations below illustrate the discussion about the stage of *Data Processing* with the data scientist $d1$:

- *"Poor data implies in non-accurate model"*
- *"Difficulty in ensuring the correctness of the dataset"*
- *"We are unable to work without the support of external actors"*

The quotations below illustrate the discussion about the stage of *Feature Engineering* with the data scientist $d4$:

- *"Difficulty in identifying suitable transformations for the data"*
- *"The stage of feature engineering impacts the other stages"*
- *"Uncertainty whether the chosen attributes are the most suitable for the model"*

Although *Model Training* does not directly depend on external actors, it was considered mainly by less experienced data scientists (90%) as problematic due to the uncertainly about the quality of the model. The quotations below illustrate the discussion about the stage of *Model Training* with the data scientist $d7$:

- *"Uncertainty whether the training data represents real data"*
- *"Uncertainty whether the model will be good in practice"*
- *"Random characteristic of the model makes its verification difficult"*

Finally, data scientists also faced problems in the stages of *Model Evaluation* and *Model Deployment*. More exactly, they reported difficulties in choosing appropriate and sufficient metrics for evaluation:

- ($d4$) *"Overfitting is often observed"*
- ($d6$) *"Classic metrics are usually not enough"*
- ($d7$) *"Uncertainty about whether the chosen metrics are the most appropriated."*

Notice that in *Model Training* and *Model Evaluation* even though data scientists mentioned problems, they discussed about the general aspects of the use of metrics and model accuracy. We describe a few problems with *Model Deployment*.

- ($d5$) *"Difficulty to deploy the model due to business rules"*
- ($d4$, $d6$) *"Data Scientists depend on the customer's infrastructure team"*

These study findings suggest that all stages are somehow problematic, mainly due to expert knowledge dependence. However, most of the problems reported regarding the last three stages may be partially explained due to lack of experience. Table 7 shows that

Table 7: Central topics of discussion per stage, with the frequencies of discussion by data scientists' experience.

Stage	Central topics (frequency of discussion)		Experience	
			Low	High
Model Requirements	Presence of internal and external actors	(67%)	58%	42%
	Others	(33%)	($d4$, $d6$, $d7$)	($d3$, $d8$)
Data Processing	Data processing as problematic	(42%)		
	Data scientist has difficulties to assess data quality	(21%)	46%	54%
	Presence of internal and external actors	(18%)	($d1$, $d4$, $d5$, $d6$, $d7$)	($d2$, $d3$, $d8$)
	Others	(19%)		
Feature Engineering	Feature engineering as problematic	(22%)		
	Use of classic techniques are not enough	(21%)	45%	55%
	Presence of internal and external actors	(20%)	($d4$, $d5$, $d6$, $d7$)	($d2$, $d3$, $d8$)
	Others	(37%)		
Model Training	Training as problematic	(26%)		
	Empirically chosen algorithms and hyperparameters	(25%)	90%	10%
	Use of framework for model development	(22%)	($d1$, $d4$, $d5$, $d6$, $d7$)	($d2$, $d3$, $d8$)
	Others	(27%)		
Model Evaluation	Process of choosing a metric for model validation	(41%)		
	External actors also evaluates the model	(19%)	76%	24%
	Model accuracy	(14%)	($d1$, $d4$, $d5$, $d6$, $d7$)	($d2$, $d3$)
	Others	(26%)		
Model Deployment	External actor presence	(41%)		
	Internal actor presence	(18%)	92%	8%
	Data Scientists has difficulties to deploy the model	(18%)	($d4$, $d5$, $d6$, $d7$)	($d8$)
	Others	(23%)		

the problems with *Model Training*, *Model Evaluation* and *Model Deployment* are more frequently reported by less experienced data scientists (76% and 92%, respectively). On the other hand, the table shows that the distribution of problems is more balanced among the groups of data scientists for the first three workflow stages (*Model Requirements*, *Data Processing* and *Feature Engineering*).

(3) Exploratory analysis. The most common problems of data scientists in both stages of *Model Training* and *Model Evaluation* addresses the diversity and complexity of alternatives for implementation. Beyond coding, there are many different frameworks and types of models available. For each model selected, there is an infinite number of parameters, metrics, and validation designs. Since the model selection is not so straightforward, data scientists usually choose over exploratory analysis. We present below a few examples of discussions around this issue:

- ($d6$, *Training*) *"Parameter optimizers are time-consuming."*
- ($d5$, *Evaluation*) *"Metrics have no defined target values"*
- ($d6$, *Evaluation*) *"Random selection of training, testing and validation sets."*

Notice that these stages are even more error-prone due to the way that data are made available in the three first stages: *Model Requirements*, *Data Processing* and *Feature Engineering*. Data scientist $d3$ argued that *"... the difficulty in improving the model accuracy may be resultant from problems in the stages of Data Processing and Feature Engineering"*. Besides, data scientist $d7$ commented that *"...the model is created in a loop involving all stages"*. Also, data scientists' answers indicate that the resulting data quality from these stages is usually not verified. For instance, data scientist $d4$ mentioned *"...the difficulty in verifying whether the chosen attributes are the most suitable for the model"*. Since these stages are very exploratory, domain experts do not end up being heavily involved in them. On the other hand, data scientists end up being uniquely responsible for understanding the data. As a consequence, problems are only discovered late in the process. Thus, while the accuracy of the model is not satisfactory, the stages are reviewed.

Given the mentioned particularities of ML model development, our findings reveal gaps in the developing stages performed by the data scientists. The different personal processes adopted by data scientists in their companies are non-linear, requiring too much rework to satisfy customers' needs. Besides, our observations do not show activities addressing the verification and the validation of the artifacts generated during the workflow stages. Based on these findings, we understand that new and engineered eye on the development of ML models is required. We believe the software engineering values such as traceability and quality assurance should be continuously addressed by concrete, planned, and structured practices. For instance, we see that customized inspection techniques should be developed to support the verification of ML features and models. Thus, we believe that by following software engineering practices since the early ML modeling stages, companies would reduce rework and the dependence of the domain experts, also leveraging the maintainability of ML models. Consequently, they would mitigate recurrent feedback loops in the process by anticipating problems and saving resources.

7 THREATS TO VALIDITY

One threat of our study address the restricted sample size. For mitigating this threat, we interviewed data scientists with diverse background and working contexts. Each data scientist is from a different team, distributed among five different companies located in Brazil. These companies distinguish themselves in terms of size, project domains, and geographical locations. Although our results suggest a great variety of ideas and agreement between responses, we are aware of the chance of geographic bias among the culture of ML practice in the country.

We execute the data analysis process using the open-coding method, which is a subjective task. To avoid bias, we carefully involve distinct authors in the coding, refinements, validation, and discussions since authors have a different interview perspective. Still, to further ensure the validity of our results, a fourth researcher who did not write Section 5, traced back the developers' quotations to their source and none traceability problem was found.

To identify possible differences between workflows followed by data scientists, we asked them to compare their workflows with the one used as a reference in this paper (see Section 2). In this way, we intentionally did not present very detailed stages in the workflow of Figure 1 since they could generate considerable disagreement and difficult data grouping. In the end, we perceived that our strategy stimulated the data scientists to feel free to discuss the granularity level of the workflow stages, detailing activities at a lower level.

ACKNOWLEDGMENTS

This work is supported by CNPq (140185/2020-8, 426005/2018-0, 311442/2019-6, 421306/2018-1, 309844/2018-5, 427787/2018-1) and CAPES(117875, 175956, 88882.452283/2019-01, 88887.373933/2019-00, 373892/2019-00). The authors thank ExACTa/PUC-Rio, LES/PUC-Rio, and IBM Research/Brazil to support the industrial case study.

8 CONCLUSION

This paper presented a study to understand how data scientists develop ML models in practice. To accomplish our goal, we investigated which stages are considered more challenging from the data scientists' perspectives and the practices involved in the most challenging stages. To perform our study, we conducted semi-structured interviews with data scientists from five different Brazilian companies. Then, we applied a code technique on six hours of transcripted interviews. As a result, we obtained 447 codes that address *actors*, *activities*, *methods*, *limitations*, and *challenges* related to the ML stages analyzed by our study.

Our findings reveal that data scientists perceive the *Data Processing* and *Feature Engineering* as the most challenging stages during the ML model development. Although, they also mention important issues on the *Model Training*, *Model Evaluation* and *Deployment*. These findings indicate lacks in the support of an engineered process to the practice of developing ML models. For instance, we did not find in the interviews techniques for assuring the traceability between the features and the model requirements. Besides, different from recommended in software engineering, the ML tests are typically not planned and do not have its coverage measured.

As future work, we intend to perform interviews by considering professionals from other companies located in different countries. Moreover, we plan to perform a action research to understand more deeply the challenges and practices during the development of ML models. At last, we also intend to extend the ML workflow by considering the findings of our interviews and experiments.

REFERENCES

[1] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *International Conference on Software Engineering: Software Engineering in Practice*. 291–300.

[2] A. Baghban, M. Bahadori, M. Lee, A. Bahadori, and T. Kashiwao. 2016. Modelling of CO2 separation from gas streams emissions in the oil and gas industries. *Petroleum Science and Technology* 34, 14 (2016), 1291–1299.

[3] Ciara Byrne. 2017. *Development Workflows for Data Scientists*. O'Reilly Media.

[4] D Cohen and B Crabtree. 2006. Qualitative research guidelines project.

[5] C. Hill, R. Bellamy, T. Erickson, and M. Burnett. 2016. Trials and tribulations of developers of intelligent systems: A field study. In *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 162–170. https://doi.org/10.1109/VLHCC.2016.7739680

[6] Correia J. 2020. Data Scientists: Revealing their Challenges and Practices on Machine Learning Model Development. https://github.com/sbqs2020/sbqs2020.

[7] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2017. Data scientists in software teams: State of the art and challenges. *IEEE Transactions on Software Engineering* 44, 11 (2017), 1024–1038.

[8] Satoshi Masuda, Kohichi Ono, Toshiaki Yasue, and Nobuhiro Hosokawa. 2018. A survey of software quality for machine learning applications. In *2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 279–284.

[9] Tom M. Mitchell. 1997. *Machine Learning*. McGraw-Hill, New York.

[10] Anh Nguyen-Duc, Ingrid Sundbø, Elizamary Nascimento, Tayana Conte, Iftekhar Ahmed, and Pekka Abrahamsson. 2020. A Multiple Case Study of Artificial Intelligent System Development in Industry. In *Proceedings of the Evaluation and Assessment in Software Engineering*. 1–10.

[11] Abdulmujeeb T Onawole, Ibnelwaleed A Husseinl, Mohammed A Saad, Musa EM Ahmed, and Hassan I Nimir. 2018. Computational Screening of Potential Inhibitors of Desulfobacter postgatei for Pyrite Scale Prevention in Oil and Gas Wells. *BioRxiv* (2018), 327957.

[12] Kayur Patel, James Fogarty, James A Landay, and Beverly Harrison. 2008. Investigating statistical machine learning as a tool for software development. In *SIGCHI Conference on Human Factors in Computing Systems*. 667–676.

[13] DJ Patil. 2011. *Building data science teams*. " O'Reilly Media, Inc.".

[14] Alessandro Piscopo, Ronald Siebes, and Lynda Hardman. 2017. Predicting sense of community and participation by applying machine learning to open government data. *Policy & Internet* 9, 1 (2017), 55–75.

[15] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2018. Data lifecycle challenges in production machine learning: a survey. *ACM SIGMOD Record* 47, 2 (2018), 17–28.

[16] Anselm Strauss and Juliet Corbin. 1998. *Basics of qualitative research techniques*. Sage publications Thousand Oaks, CA.

[17] Michael A. Tabak, Mohammad S. Norouzzadeh, David W. Wolfson, Steven J. Sweeney, Kurt C. VerCauteren, Nathan P. Snow, Joseph M. Halseth, Paul A. Di Salvo, Jesse S. Lewis, Michael D. White, et al. 2019. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution* 10, 4 (2019), 585–590.

[18] Jana Wäldchen and Patrick Mäder. 2018. Machine learning for image based species identification. *Methods in Ecology and Evolution* 9, 11 (2018), 2216–2225.

[19] Zhiyuan Wan, Xin Xia, David Lo, and Gail C Murphy. 2019. How does Machine Learning Change Software Development Practices? *IEEE Transactions on Software Engineering* (2019).

[20] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* (2020).

[21] Tianyi Zhang, Cuiyun Gao, Lei Ma, Michael Lyu, and Miryung Kim. 2019. An empirical study of common challenges in developing deep learning applications. In *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 104–115.