Metrics for Evaluating Machine Learning Cloud Services

Thesis Report

PAPER WITHIN Software Product Engineering
AUTHOR: Augustin Tataru
TUTOR: Tuwe Löfström
JÖNKÖPING October 2017
This exam work has been carried out at the School of Engineering in Jönköping in the subject area “Metrics for Evaluating Machine Learning Cloud Services”. The work is a part of the two-year university diploma programme, of the Master of Science programme, Software Product Engineering. The authors take full responsibility for opinions, conclusions and findings presented.

Examiner: Ulf Johansson

Supervisor: Tuwe Löfström

Scope: 30 credits (second cycle)

Date: 2017-06-09
Abstract

Machine Learning (ML) is nowadays being offered as a service by several cloud providers. Consumers require metrics to be able to evaluate and compare between multiple ML cloud services. There aren’t many established metrics that can be used specifically for these types of services. In this paper, the Goal-Question-Metric paradigm is used to define a set of metrics applicable for ML cloud services. The metrics are created based on goals expressed by professionals who use or are interested in using these services. At the end, a questionnaire is used to evaluate the metrics based on two criteria: relevance and ease of use.
Summary

ML cloud services aim to make ML accessible and easy to use. Given that there are several such services now, and it is likely that more will appear in the future, users need to be able to evaluate and compare between multiple services. To do this they need to have metrics to measure various aspects of these services. Unfortunately, there aren’t many established metrics to do this. The purpose of this research is to systematically define a set of metrics that can be used to evaluate ML cloud services. The main research question with its sub-questions are:

What metrics can be used to evaluate ML Cloud Services?
- What goals related to the ML Cloud Services the stakeholders have?
- What metrics can be used to determine the effectiveness in meeting the goals?

The research method used to answer the main research question is Design and Creation. This method is suitable to use when artifacts must be designed and created. The method is used in combination with Goal Question Metric (GQM) – a paradigm that describes a structured approach to creating metrics.

To answer the first sub-question, interviews were conducted with professionals using or interested in using ML cloud services. Interview data was analyzed to identify goals they have for ML cloud services. Two types of goals were distinguished: quality goals and functional goals. Most of the quality goals are related to cost, usability, availability, integrability and performance of ML cloud services.

To answer the second sub-question, artifact creation was used. Questions were asked about previously identified goals. After, metrics were created that answer the questions. Some metrics are quantitative while others are Boolean. It was not possible to come up with metrics for certain goals. The created metrics were evaluated using questionnaires based on two criteria: relevance and ease of use. The questionnaires were addressed to the interviewed stakeholders. The questionnaire results showed that the stakeholders find the metrics relevant and not very hard to use.

It was difficult to create quantitative metrics for some aspects of ML cloud services. It was especially difficult to create quantitative metrics based on functional goals. Additional iterations of the GQM could improve the metrics even further.

Keywords

Metrics, Machine Learning, Cloud services, MLaaS, GQM, Evaluation
Contents

List of Figures ................................................................................................................. 2
List of Tables ..................................................................................................................... 2
1 Introduction .................................................................................................................... 3
  1.1 BACKGROUND ........................................................................................................... 3
  1.2 PURPOSE AND RESEARCH QUESTIONS ................................................................. 4
  1.3 DELIMITATIONS .................................................................................................... 4
  1.4 OUTLINE ................................................................................................................ 5
2 Theoretical background ................................................................................................. 6
  2.1 MACHINE LEARNING .............................................................................................. 6
  2.2 DATA SCIENCE ....................................................................................................... 10
  2.3 CLOUD COMPUTING ........................................................................................... 15
  2.4 MACHINE LEARNING CLOUD SERVICES ............................................................ 17
  2.5 SOFTWARE METRICS ............................................................................................ 19
  2.6 DEFINING METRICS USING GOAL-QUESTION-METRIC PARADIGM ...................... 22
  2.7 STAKEHOLDERS IN PROJECTS USING ML CLOUD SERVICES ............................... 25
  2.8 PREVIOUS RESEARCH .......................................................................................... 26
3 Method and implementation ......................................................................................... 30
  3.1 RESEARCH METHOD ............................................................................................. 30
  3.2 IMPLEMENTATION .................................................................................................. 31
4 Findings and analysis .................................................................................................... 35
  4.1 WHAT GOALS RELATED TO THE ML CLOUD SERVICES THE STAKEHOLDERS HAVE? ....... 35
  4.2 WHAT METRICS CAN BE USED TO DETERMINE THE EFFECTIVENESS IN MEETING THE GOALS? 42
5 Discussion and conclusions ......................................................................................... 69
  5.1 DISCUSSION OF METHOD ....................................................................................... 69
  5.2 DISCUSSION OF FINDINGS ................................................................................... 70
  5.3 CONCLUSIONS ....................................................................................................... 71
6 References ..................................................................................................................... 73
7 Appendices ..................................................................................................................... 77
7.1 SET A - INTERVIEW QUESTIONS FOR PEOPLE WITH LIMITED EXPERIENCE WITH ML/DATA MINING OR WHICH ARE NEW TO THE FIELD ................................................................. 77

7.2 SET B - INTERVIEW QUESTIONS FOR PEOPLE WHO ARE EXPERIENCED WITH ML OR DATA MINING 77

7.3 ALE - INTERVIEW WITH A1 (2017-04-18) ........................................................................... 79

7.4 ERI - INTERVIEW WITH B1 (2017-04-25) ................................................................. 86

7.5 SUM - INTERVIEW WITH C1, C2 AND C3 (2017-05-12) ............................................. 93

List of Figures

Figure 1. Data Science Disciplines (Barga, et al., 2015) .............................................................. 11
Figure 2. CRISP Data Mining Process (Provost & Fawcett, 2013) ................................................. 12
Figure 3. The Cloud Stack (Kavis, 2014) ............................................................................. 16
Figure 4. A pyramid model of cloud computing paradigms (Marinescu, 2013) ..................... 16
Figure 5. What are software metrics? (Westfall, 2005) ............................................................ 19
Figure 6. Measurement defined (Westfall, 2005) .................................................................. 20
Figure 7. Metric and property (NIST, National Institute of Standards and Technology, 2015) 21
Figure 8. The structure of GQM (Basili, et al., 1994) ............................................................ 23
Figure 9. The goal’s coordinates (Basili, et al., 1994) ............................................................ 24

List of Tables

Table 1. Abbreviations for interview transcripts ........................................................................ 35
Table 2. Examples of referencing interview transcripts .......................................................... 35
Table 3. Stakeholder roles of the interviewed people .............................................................. 36
Table 4. Quality goals ............................................................................................................. 41
Table 5. Functional goals ....................................................................................................... 41
Table 6. Mapping of rating scores to ordinal values .............................................................. 66
Table 7. Questionnaire results ............................................................................................... 66
Table 8. Statistics for Relevance and Ease of Use ................................................................. 67
1 Introduction

1.1 Background

Cloud computing is now essential for organizations to maintain their competitive edge. It enables companies to reduce costs and increase the business flexibility (Michael, et al., 2010). The main features of using the cloud are elasticity on demand, costs savings and high performance (Deepak, et al., 2015). There are many business applications that can take advantage of the cloud computing. One of these applications is managing extremely large data sets known as Big Data. Big Data is usually coupled with cloud computing because it is typical for companies that have this large volume of data to use the cloud to store and process it (Agrawal, et al., 2011). Then to utilize the dynamic scalability and high performance of the cloud to achieve the benefits of reduced cost, less infrastructure, and increased time to value by focusing on what to do with the data which is what matters the most to the business (McAfee & Brynjolfsson, 2012).

There is a high demand in organizations to reveal patterns, trends, and associations contained in a large amount of data. This process is called Data Mining (Tan, et al., 2006). Using basic data analysis techniques - sums, counts, averages and database queries - is not enough in data mining. The volume of the data is too large for comprehensive coverage, and the potential patterns and associations between the data sources are too much to be observed by the analyst. Therefore, more advanced techniques are required such as Machine Learning (ML) (Witten, et al., 2011). Unlike other types of analysis, the prediction of ML algorithms is usually enhanced with huge data. This predictive capability can help in making effective business decisions.

Many of the ML algorithms require high computational power because they need to operate on large amounts of data. This makes it difficult to use on-premises computers to perform ML (Xu, et al., 2013). On the other hand, the cloud environment offers high computational power and parallel processing techniques such as Map-Reduce (Ekanayake & Fox, 2009). It is possible to create a ML system from scratch and deploy it on the cloud’s infrastructure. However, this requires advanced technical knowledge of ML, developing systems and deploying them on the cloud (Low, et al., 2012).

In recent years, several cloud providers have started to offer ML cloud services (Nketah, 2016). Among them are big cloud providers such as Azure, Amazon Web Services, and Google. But there are also smaller companies such as BigML which have a wide spectrum of options for cloud ML and entered the market even earlier than the big companies. These cloud providers offer tools and services necessary to perform the steps in a ML workflow. Moreover, it is possible to use the created ML models for setting up your own services that for example could offer real-time predictions.

The purpose of the ML cloud services is to make it easy for scientists, developers, companies and other individuals to use ML technologies (Amazon Web Service, 2017) and even integrate them with their own systems. However, the wide
spectrum of choices of providers and the fact that the cloud providers offer the ML services in a different way makes it hard to choose the ML cloud solution. Furthermore, it is very likely that other cloud providers will enter the market offering similar services. The question is: What information is needed to evaluate if an ML cloud service is a right option for a problem?

The cloud service providers have Service Level Agreement (SLA) where they mention the available resources and prices (Microsoft Azure, 2015). Nonetheless, there are many other factors that should be considered when selecting the service. It can be a challenge even for someone who is familiar with ML but has only used on-premises solutions. It would be troublesome to dedicate to one ML solution and later realize that it doesn’t fit the problem. This usually means a waste of money and time. Therefore, more information is required to evaluate different ML cloud services.

1.2 Purpose and research questions

There is little published research regarding metrics and evaluation of ML cloud services. One of the reasons could be that the ML cloud services are quite new. Aside from some exceptions, most of the ML cloud services have appeared in the last 2-3 years, and some of them are still in beta release (Google Cloud Platform, 2017). This gives the problem more importance and leads to the necessity of having well-defined metrics. Therefore, the goal of this research is to systematically define metrics that can be used to evaluate ML cloud services. These metrics can be the basis of an evaluation process used by anyone who intends to do an ML project in the cloud, and it will be mainly used in the initial stages of the project when a decision must be made about what service is suitable for the problem. The main research question is:

What metrics can be used to evaluate ML Cloud Services?

The steps to define the metrics are based on the approach of Goal-Question-Paradigm (GQM), developed by Basili et al. (1994). This approach is used by modern research to establish a software engineering measurement process which usually includes defining metrics (Becker, et al., 2015). It is based on identifying stakeholders and their goals, then asking questions related to these goals. Finally, metrics are selected to answer the questions. The GQM approach was utilized to answer the main question in the thesis by performing several steps (Westfall, 2006). The steps are established in a structured way to answer the sub-questions below:

• What goals related to the ML Cloud Services the stakeholders have?
• What metrics can be used to determine the effectiveness in meeting the goals?

1.3 Delimitations

The research work does not cover prioritizing the metrics and does not describe a detailed process, such as a framework or methodology, for how the metrics or a subset of them should be combined or used together. Also, the research work does not cover evaluating the use of the metrics in a real-world scenario.
1.4 Outline

The report is divided into the following sections in this order: Introduction, Theoretical background, Method and implementation, Findings and analysis, Discussion and conclusions, References and Appendices.

The main theoretical concepts that were used as a foundation for this research are presented in the Theoretical background section. This section also contains a summary of previous research work related to the topic of the thesis.

The Method and implementation section describes the research method used to answer the main research question. The data generation techniques and how the data is analyzed to produce findings are presented for each research sub-question.

The Findings and analysis section presents the collected data, the analysis process, and the obtained findings. This is the section containing the results of the work.

The Discussion and conclusion section contains a discussion of research method and findings, and most importantly, the conclusion where the main points are summarized.

The References section lists all the references that are used in the paper.

The Appendices contains additional material or documents that were used in the process of research and are referred in the paper.
2 Theoretical background

2.1 Machine Learning

2.1.1 Overview

Machine Learning (ML) is a term we get to hear a lot nowadays. The term Machine Learning was defined by Arthur Samuel in 1959 as “the field of study that gives computers the ability to learn without being explicitly programmed” (Meysman, et al., 2016). This is a simple definition that explains well the general idea behind ML. Basically, in ML, instead of hard-coding rules into the programs, you make the programs learn from already existing knowledge, just like the way a person learns. Samuel (1959) says in one of his papers that “programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort”.

Mitchell’s (1997) definition of machine learning is as follows: “a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”. It is a formal definition which can be easily applied to real-world machine learning problems. For example, the problem of classifying spam email can be expressed through this definition as follows: the class of tasks T is correctly classifying spam email, the experience E constitutes the emails that already have been labeled by humans as spam or not and the performance measure P is how accurate the program can predict if a previously unknown email is spam or not. This example describes a supervised learning approach.

Alpaydin (2014) defines ML as “programming computers to optimize a performance criterion using example data or past experience”. This definition ties the concept of ML to optimization problems and mentions the term “data”. Data is most of the time essential to ML and as Alpaydin (2014) says “where we lack in knowledge, we make up for in data”. The more data you feed into a ML program the better it gets at performing its task. Solving a specific task can be achieved by offering specific data to the algorithm (Meysman, et al., 2016). The quality of the data is therefore very important. When talking about Classification, which is one application of ML, Miroslav (2015) says that “the classifier can only be as good as the data that have been used during its induction”.

ML is part of the artificial intelligence field and borrows techniques from mathematics and statistics (Meysman, et al., 2016). This idea is supported by Alpaydin (2014) who states that “to be intelligent, a system that is in a changing environment should have the ability to learn”. Indeed, a computer capable of learning is a step forward towards achieving artificial intelligence.

2.1.2 The use of Machine Learning

The reader might ask why is ML needed. To solve a problem using a computer, an algorithm is used. The problem is usually described by an input and the desired output. The algorithm is responsible for transforming the input into the output and usually, there are several algorithms that can solve the same problem, some being more efficient than others. However, for some problems, it is not clear how
to transform the input into output which basically means there is no clear algorithm (Alpaydin, 2014). One example of such a problem is making a prediction whether a customer will buy a product or not. Making a prediction, in this case, is not easy because the customer’s decision of buying or not can depend on the combination of many factors. This makes it hard to define rules capable of predicting the customer’s behavior. This is especially true in the case of large amounts of data. However, as previously said, ML algorithms can become better with more data. Machine Learning can provide a solution to these types of problems because it can identify patterns, some of which can prove very difficult to identify by human experts. According to Alpaydin (2014), with ML “we may not be able to identify the process completely, but we believe we can construct a good and useful approximation”. The same author suggests that even though it is just an approximation, it can still provide useful information necessary for example to make predictions or discover patterns which can help understand the process better.

ML can be used in many situations. To name just a few examples: ranking web pages, face recognition, classifying spam email, predicting the price of a product, give a diagnosis for a patient and many others. It is employed heavily in data science. Meysman et. al. (2016) refers to ML as being “ubiquitous within data science “. In data mining, ML algorithms are used to discover valuable knowledge from large volumes of data (Mitchell, 1997).

2.1.3 Types of machine learning

Types of ML can be categorized based on the amount of human effort needed for learning and if the instances in the data are labeled or not (Meysman, et al., 2016). By using this criterion, the following types of ML can be distinguished: supervised learning, unsupervised learning and semi-supervised learning.

**Supervised learning**

“In supervised learning, the aim is to learn a mapping from the input to an output, whose correct values are provided by a supervisor” (Alpaydin, 2014). This type of learning is used when the data is labeled (Meysman, et al., 2016). Besides its features, each instance in the data has a label representing a categorical, binary or continuous attribute. The label is also called the class. The goal of the supervised learning is to build a model that can predict the label value for new unlabeled instances for which the feature values are known (Kotsiantis, 2007).

The types of ML problems that fall into this category of learning are regression and classification. In such problems “there is an input X, an output Y, and the task is to learn the mapping from the input to the output” (Alpaydin, 2014). The input X represents the features of an instance, while the output Ŷ represents the label for classification or a continuous value for regression.

**Unsupervised learning**

In practice, the data is often only partially labeled (Meysman, et al., 2016), which prevents using supervised learning techniques. However, these datasets are still valuable because they can be used to study how the data is distributed within
them, study the structure and values in the data, discover patterns. In this type of learning, there is no supervisor to provide the labels for the data and “the aim is to find regularities in the input” (Alpaydin, 2014). A very common type of problem solved through this type of learning is clustering (Amparo & Wolfgang, 2011).

**Semi-supervised learning**

One significant weakness of supervised learning techniques is that they require labeled data. Labeling big data sets can be expensive and time-consuming. To overcome these issues semi-supervised techniques were created. Semi-supervised techniques use both labeled and unlabeled data to create models with better performance than in a supervised learning approach using the same data (Amparo & Wolfgang, 2011). One semi-supervised learning technique called label propagation uses labeled data to label similar unlabeled instances with the same label. Semi-supervised learning can be used when a small part of the data has labels (Meysman, et al., 2016).

### 2.1.4 The process of ML

The outcome of ML is to produce a model. Its nature can be predictive or descriptive (Alpaydin, 2014). This model can be represented through a software entity which could, for example, predict new values or classify instances. Alpaydin (2014) describes this process as “using a learning algorithm on a dataset and generate a learner”. The learner is the resulting model. A modeling process can have the following steps (Meysman, et al., 2016):

1. Feature engineering and algorithm selection
2. Training the model
3. Model validation and selection
4. Applying the model to unseen data

The last step is not always necessary because sometimes the goal is to extract some insights and patterns from the produced model (Meysman, et al., 2016).

**Feature engineering and algorithm selection**

The first step is very important in the modeling process because the quality of the model will depend heavily on the selected features. At this stage, it might be necessary to consult a domain expert who could say which of the features have a high chance of being useful or not (Meysman, et al., 2016). The reason for having both feature engineering and algorithm selection at step one is probably because these activities can depend on each other. Referring to data mining which employs ML, Tan et.al. (2006) state that “the type of data determines which tools and techniques can be used to analyze the data”.

In other literature sources, the first step is called data pre-processing. This step is necessary because the training data can come with many issues such as the presence of noise and outliers, missing data, inconsistent data, duplicate data (Tan, et al., 2006). When talking about the quality of the data, Tan et. al. (2006) say that “data is often far from perfect”. The same authors also mention different strategies and techniques that can be used for data preprocessing such as aggregation, sampling, dimensionality reduction, feature subset creation, feature
creation and others. Basically, feature engineering is included or is the same as data pre-processing depending on the context.

**Training the model**

This is the phase when the model is trained. Some literature sources prefer to use words such as 'build' or 'generate' instead of 'train', however, they all refer to the same thing.

A definition for model training given by Meysman (2016) is “a model is fed with the data and it learns the patterns hidden in the data”. In this case, there is no distinction being made between the algorithm and model as it considers them as one whole. Rasmussen (2006) mentions a type of models called parametric models which basically “absorb the information from the training data into the parameters” of the model, after which the training data is not needed anymore by the model. In classification, a learning algorithm identifies a model that “best fits the relationship between the attribute set and class label of the input data” (Tan, et al., 2006). The algorithm generates the model.

Once the model is created, depending on its nature, it can be used to perform a certain task (for example, classification) or provide a description of a certain phenomenon under study. However, before using it in the real world it is necessary to check if it is useful.

**Model validation and selection**

Good models should represent well reality and be interpretable (Meysman, et al., 2016). The model validation/evaluation phase is needed to determine if the model produced during the previous step is any good. This can be accomplished by using an error measure and a validation strategy. Different error measures can be used depending on the type of problem. In classification, a common measure used is classification error rate while in regression mean squared error is used (Meysman, et al., 2016). There are also various measures and techniques to evaluate models produced for unsupervised learning type problems like clustering (Tan, et al., 2006).

There are several validation strategies that can be applied to a model. Here are some of them:

- **Holdout method** – the data is split into two partitions, the training data and the test data. The training data is used to train the model, and then the test data is used to evaluate it (Tan, et al., 2006).
- **N-folds cross-validation** (also known as K-folds cross-validation) – the data is split into N subsets. Next, each of the N subsets is used as test data for a model that uses the rest N-1 subsets as training data (Miroslav, 2015).
- **Leave-one-out** – it uses the same principle as N-folds with the exception that the size of each N subset is 1.

An important issue that can appear when doing classification or regression is model overfitting. An overfitted model has very good results when applied to the training data, but offers unsatisfactory results when applied to the validation data.
or new data. There are various techniques to address this issue depending on the model (Tan, et al., 2006).

Alpaydin (2014) mentions that there are additional properties of the model which might need evaluation such as the efficiency of its space representation or time complexity.

**Applying the model to unseen data**

It's finally time to use the model for the purpose it was created in the first place. At this stage, the model is applied to real world data. The data will most likely come unlabeled so we rely on the model to offer the right output (Meysman, et al., 2016).

### 2.2 Data Science

#### 2.2.1 Overview

With the rise of big data in organizations, the demand for advanced data analysis techniques increases significantly. Using classical data analysis methods becomes insufficient (Meysman, et al., 2016). This is mainly because of the massive size of the data, since the analysis are based on scanning through all the data to gain some insight using manual or partially automated tools (Tan, et al., 2006). The explosion of data shows that the current data management systems, such as Relational Database Management Systems, are not capable of obtaining more value from it, except managing the storing and accessing of data (Meysman, et al., 2016). Therefore, companies realized that with the fast development of computer power and artificial intelligence algorithms, more competitive advantage can be obtained by deeply exploring the data (Provost & Fawcett, 2013).

#### 2.2.2 Data Science Definition and Disciplines

The term Data Science has existed since the 1960s in different types of science (Gerstein & Kiang, 1960), and the movement toward making it an independent discipline started from 1990s (Cleveland, 2001). Data Science can be defined as “the practice of obtaining useful insights from data” (Barga, et al., 2015).

“Data science involves using methods to analyse massive amounts of data and extract the knowledge it contains” (Meysman, et al., 2016). Therefore, Data Science is much more than the traditional data analysis techniques, statistical methods, or database queries. Data science is a multidisciplinary field (Barga, et al., 2015). Its disciplines are shown in Figure 1.

In fact, researchers from each of these disciplines are contributing continuously to the field of data science by developing more efficient and scalable tools that can explore large amounts and different types of data (Tan, et al., 2006). As a result, this field improved commutatively by adopting different methodologies, techniques and algorithms to establish the structure of data science.
2.2.3 Data Mining

Another term that started to be used widely since the late 1990s in the database community is Data Mining (Chen, et al., 1996). Alpaydin (2014) defines data mining as “applications of machine learning to large databases”.

Organizations are using data mining to understand their data and detect more opportunities to grow their businesses. Data mining, in general, refers to the process of gathering all previous data and then try looking for patterns in this data. Data mining utilizes traditional data analysis techniques as well as advanced intelligent algorithms. When new knowledge is obtained, it is validated by testing the detected patterns on new subsets of data, then it is mostly used for predicting future observations (Tan, et al., 2006).

2.2.4 Data Mining Process

Data mining can be used and optimized for different purposes depending on the business goals and the context. Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman, et al., 2000) is a well-defined structure of the data mining process (Provost & Fawcett, 2013). It describes and organizes the required steps in data mining projects starting from understanding the business until the integration with actual systems to make useful decisions (Chapman, et al., 2000). The steps of data mining in CRISP-DM are:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment
2.2.4.1 Business Understanding

The business drives data science and data mining, it is the source of the need to understand and analyze the data (Meysman, et al., 2016). Therefore, it is very important to understand the business and the context to establish the data mining goal (Provost & Fawcett, 2013).

Understanding the business is also a key part when choosing the best data mining model to use for the problem, and how to deploy the results to the business tools, such as the decision support system, and customer relationship management (Tan, et al., 2006). At the end, evaluating the model, and the data mining results before and after deployment is important to be done by a stakeholder who understands the business and its goals.

2.2.4.2 Data Understanding

Understanding the data that is being collected its properties complement the business understanding and forming the goal of the data mining (Provost & Fawcett, 2013). Two main aspects to consider when trying to understand the data are the type of data, and the quality of data.

Types of data

The types of data differ in many ways, such the data storage method or how it is represented. The main types of data can be categorized as (Meysman, et al., 2016):

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
Each data object of the categories above needs to be transformed into more structured instances, even the structured data itself. Each instance is usually represented by attributes, and the attributes can be within four types: nominal, ordinal, interval, and ratio. However, each attribute can have different types of values. These value types can be grouped as discrete or continuous values (Tan, et al., 2006).

“An attribute is a property or characteristic of an object that may vary; either from one object to another or from one time to another” (Tan, et al., 2006).

Understanding the differences between the attribute types is a key success element when preparing the data and building the model in next steps (Chapman, et al., 2000).

Quality of data

Having a good idea about the quality of the existing data and how to improve it is very useful for obtaining the required results from data mining. Potential issues that data can have (Tan, et al., 2006):

- The presence of noise and outliers
- Missing data
- Inconsistent data
- Duplicate data
- Data is unrepresentative of the population that it is supposed to describe

2.2.4.3 Data Preparation

Data preparation is considered the most consuming tasks in the data mining process (Tan, et al., 2006). Although data gathering might be considered as a standalone step, it can yet be considered part of the data preparation in data mining, and usually the first step of this phase. The sources of data are also not necessary to be from one organization, it can be from different domains (Meysman, et al., 2016). The main steps in the data preparation are (Tan, et al., 2006):

1. Merging data from multiple sources
2. Cleaning data, such as removing missing values, outliers, typos and duplicate observations
3. Selecting data instances that are relevant to the data mining goal
4. Data pre-processing. For example, transforming an attribute with continuous value type into an attribute with discrete value type, i.e. length from number of centimeters into “Short, Medium, Long”

2.2.4.4 Modeling

This phase is similar to the process of ML presented earlier. The tasks are performed at this phase are (Chapman, et al., 2000):

1. Select modeling technique.
2. Generate test design - the test design refers to the model validation techniques.
3. Build the model.
4. Assess the model - during this task, the quality of the produced model is assessed. Modeling is an iterative process because building and assessing the model can happen several times until an acceptable model is produced (Chapman, et al., 2000).

2.2.4.5 Evaluation
Now that a model has been produced and assessed, it is time to review all the steps performed to create it and check if it achieves the initial business objectives. Also, at this step other data mining results obtained so far during the project are evaluated. Some of the results might not be tied to the original business objective, however, they could still provide some useful information. (Chapman, et al., 2000).

2.2.4.6 Deployment
Depending on its type, the produced model during the data mining process is used in a different way. If the purpose of the model was to provide understanding or knowledge, then this knowledge should be presented in a way that is accessible to the customer. However, very often “live” models must be applied to an organization’s decision-making process (Chapman, et al., 2000). This could be the case for a bank that uses a model to decide if they should offer a loan to a customer. The tasks performed at this phase are (Chapman, et al., 2000):

1. Plan deployment
2. Plan monitoring and maintenance – the authors emphasize that monitoring and maintenance require special consideration if the data mining results will be used often by the organization. This is because incorrect usage of these results could be detrimental to the organization.
3. Produce final report

2.2.5 Data mining tasks
The data mining tasks describe what can be done with data mining. According to Tan, et al. (2006), the main data mining tasks are:

**Predictive Tasks**
It is based on predicting the value of one attribute based on values of other attributes in an instance of the data. The terms Dependent or Target Variable are often used to refer to the attribute value that needs to be predicted. The terms Independent or Exploratory Variables are used to describe the attributes that used to make the prediction (Tan, et al., 2006).

**Descriptive Tasks**
These are the tasks related to describing patterns, relationships or hidden information in the existing data that was unknown before. Descriptive analysis usually requires a deeper understanding of the data than other parts which will require more domain experience as well (Tan, et al., 2006).

**Association Analysis**
This task is focused on deriving rules based on the discovered patterns in the data. These rules are often called Association Rules (Tan, et al., 2006). It is necessary to
have a good understanding of the complexity of the data in this type of analysis as the retrieval of rules can grow exponentially. Therefore, many algorithms are optimized for this task to get the most interesting rules based on the business goal.

**Cluster Analysis**

It means categorizing the data into different clusters. The clusters obtained through this process are based on common observations in a group of data instances. The observation can be recognized manually. However, with a large amount of data more advanced techniques are required depending on the type and quality of data (Tan, et al., 2006).

**Anomaly Detection**

The detection of significant differences in some instances of the data apart from the rest of the data. The detected instances are often called Anomalies or Outliers (Tan, et al., 2006). Examples of data mining goals that are relevant to this task are fraud detection, and unusual patterns of diseases.

### 2.3 Cloud computing

#### 2.3.1 Overview

National Institute of Standards and Technology (NIST) defines cloud computing as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction” (Mell & Grance, 2011). The cloud model described by NIST “is composed of five essential characteristics, three service models, and four deployment models” (Mell & Grance, 2011):

- **Essential characteristics**: on-demand self-service, broad network access, resource pooling, rapid elasticity, measured service
- **Service models**: Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS)
- **Deployment models**: Private cloud, Community cloud, Public cloud, Hybrid cloud

Erl et al. (2013) define cloud as a “distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources” and state that cloud computing “is a form of service provisioning”.

A cloud service is defined as “any IT resource that is made remotely accessible via a cloud” (Erl, et al., 2013). The author of this definition acknowledges that the term “service” can have a broad meaning in the context of cloud computing. A cloud service may be offered as a web application or something more complex like a remote access point or an Application Programming Interface to various IT resources (Erl, et al., 2013).

Each cloud service model offers a different level of abstraction “that reduces efforts required by the service consumer to build and deploy systems” (Kavis, 2014). Figure 3 presents the so-called “Cloud stack”. The image clearly shows that
as we move from IaaS, PaaS and to SaaS, the effort required from the consumer of the service decreases.

It seems that the cloud service model is an example of a layered architecture where the service levels are built on top of each other. Marinescu (2013) refers to this as the “pyramid model of cloud computing paradigms” and its representation can be seen in Figure 4.

2.3.2 Service Level Agreement

Being a measured service is one of the essential characteristics of cloud computing (Mell & Grance, 2011). Erl et al. (2013) state that a Service Level Agreement can offer information regarding “various measurable characteristics related to IT
outcomes” and this is especially useful for the consumer (customer) who doesn’t know the underlying details of the implementation for the service he is using.

Kavis (2014) defines an SLA as “an agreement between the cloud service provider (CSP) and the cloud service consumer (CSC) that sets the expectation of the level of service that the CSP promises to provide to the CSC”. The author further mentions how critical SLAs are for cloud services because the cloud service provider assumes responsibility towards the consumer.

### 2.3.3 Additional considerations

The main business factors that lead to the creation of cloud computing are capacity planning, cost reduction and organizational agility (Erl, et al., 2013). Cloud computing brings significant changes to the IT industry because it has made possible computing as a utility (Armbrust, et al., 2010). Some of the main benefits of cloud computing are reduced investments and proportional costs, increased scalability, increased availability and reliability (Erl, et al., 2013). Marinescu (2013) says that cloud computing is seen by many as “an opportunity to develop new businesses with minimum investment in computing equipment and human resources”. However, Kavis (2014) suggests that many make the mistake of selecting a cloud vendor they are familiar with instead of choosing based on their needs. Erl et al. (2013) point out that “there is no greater danger to a business than approaching cloud computing adoption with ignorance”.

### 2.4 Machine Learning Cloud Services

#### 2.4.1 Overview

Developing ML solutions used to require advanced knowledge and expensive resources (Nketah, 2016). Because of this ML was accessible to mostly large companies who had such capabilities. However, for smaller companies or individual IT professionals, it was hard to harness the power of ML. One way to overcome the mentioned issues would be to have a platform providing ML as a Service (MLaaS) which would be able to offer computational resources on-demand and a clearly defined interface to ML processes. Such a platform would allow the users to focus on the problem they are trying to solve instead of dealing with implementation details (Ribeiro, et al., 2015). In short, MLaaS can be used to “build models and deploy in production” (Baskar, et al., 2016).

In recent years, many cloud providers started offering services which allow IT professionals to perform machine learning easily and with reduced cost (Nketah, 2016). Examples of services are Amazon ML (Amazon Web Services, 2017), Google Cloud ML Services (Google Cloud Platform, 2017), Azure ML (Microsoft Azure, 2017). ML services are not being provided only by big names such as Amazon, Google or Microsoft. Other smaller companies such as Big ML (Big ML, 2017) do it as well. Ribeiro et al. (2015) say that “the increasing demand for machine learning is leveraging the emergence of new solutions” which implies that new similar services will appear. According to a news article in M2 Presswire (2016), “the MLaaS market size is estimated to grow from USD 613.4 million in 2016 to USD 3,755.0 million by 2021, at a Compound Annual Growth Rate (CAGR) of 43.7% from 2016 to 2021”.

17
2.4.2 Examples of ML Cloud Services

Microsoft Azure (2017) defines its Azure ML as “a fully managed cloud service that enables you to easily build, deploy and share predictive analytics solutions” and says that it allows to “go from idea to deployment in a matter of clicks”. It is a complete cloud service which does not require users to purchase hardware and software or take care of deployment and maintenance issues (Mund, 2015). Creating a ML workflow is done using a web interface where the user just adds and connects various elements. Once the model is built it can be deployed very quickly as a web service which can be consumed from different platforms such as desktop, web or mobile (Barga, et al., 2015).

Google Cloud Platform (2017) defines its Cloud ML Engine as a “managed service that enables you to easily build machine learning models, that work on any type of data, of any size”. The trained model can be used right away as a web service with the global prediction platform which is highly scalable. The users focus on building their models while the cloud platform behind the service takes care of the rest (Google Cloud Platform, 2017).

Amazon Web Services (2017) defines its Amazon ML as a “managed service for building ML models and generating predictions, enabling the development of robust, scalable smart applications”. This service aims to help developers of all skill levels to use ML without the need to know the details behind advanced machine learning techniques or manage the infrastructure that will power the prediction models. The documentation of the service does mention that the quality of the created models depends on the training data (Amazon Web Services, 2017).

Some recurring ideas can be noticed by looking at the short descriptions provided above for each of these services. First, all of them are described as “managed” services. Second, all the services allow to build and deploy machine learning models. Third, it seems that all these 3 services have a focus around predictive analysis. The fact that they have similar aspects makes it interesting to compare them based on these aspects. For simplicity, ML cloud services will also be referred in this document as ML as a Service (MLaaS). Based on the documentation of these services, the general workflow of using MLaaS involves the following steps:

1. Import the data
2. Analyze and preprocess the data (optional)
3. Create the model
4. Evaluate the model (optional)
5. Deploy the model as a predictive web service (optional)
6. Use the model

2.4.3 Benefits of using ML Cloud Services

The documentation for the ML Cloud services mention some of the main benefits they have (Amazon Web Services, 2017), (Google Cloud Platform, 2017), (Microsoft Azure, 2017):
• Proven ML technology solutions
• Easy and fast creation of ML models
• Deployment of ML models as web services
• High scalability and computational performance (benefits of Cloud Computing in general)
• Tools for data preprocessing and visualization
• Integration with other cloud services of the same provider, such as storage services
• No need to manage any infrastructure
• Reduced cost

2.5 Software metrics
2.5.1 Overview
The term metrics or software metrics can often be encountered in the software engineering discipline. Goodman (1993) defines software metrics as “the continuous application of measurement-based techniques to the software development process and its products to supply meaningful and timely management information, together with the use of those techniques to improve that process and its products”. Figure 5 explains the relationship between the different concepts mentioned in the definition. This definition is related to traditional software development back in the days when cloud computing did not exist. Nevertheless, it is generic enough to also be applied to development projects using cloud services.

Besides metrics, another important concept used in metrology is the measurement. Fenton (1991) defines measurement as “the process by which numbers or symbols are assigned to attributes or entities in the real world in such a way as to describe them according to clearly defined rules”. As Figure 6 shows, during measurement, numbers and symbols are mapped to features and properties of entities.
NIST (2015) provides definitions to several metrology concepts in the context of cloud services. Here are some of the definitions:

- A cloud service property is “a property of the cloud service to be observed”.
- The context is defined as “the circumstances that form the setting for an event, statement, or idea, in which the meaning of a metric can be fully understood and assessed”.
- A metric is “a standard of measurement that defines the conditions and the rules for performing the measurement and for understanding the results of a measurement”.

NIST (2015) defines measurement as a set of operations that produces a value which expresses an assessment of a property of an entity. The same source suggests using the term “measurement result” for the value produced by the measurement instead of using the term “measure”. This is because “measure” has multiple definitions, which can cause confusion.

Other important concepts are observation and unit of measurement. NIST (2015) defines an observation as a “measurement based on a metric, at a point in time, on a measurement target” and a unit of measurement as a “real scalar quantity, defined and adopted by convention, with which any other quantity of the same kind can be compared to express the ratio of the two quantities as a number”. Notice that the definition for “unit of measurement” emphasizes that is should be a real scalar quantity. According to the same source, this does not prevent metrics from using a qualitative scale with nominal or ordinal values.

It is important to make a clear distinction between metric and measurement. A measurement uses a metric. Figure 7 shows the relationships between some of the concepts defined earlier. The measurement is not present in the image but it can be though as the combination of metric, observation and measurement result. During the measurement, a metric is used through an observation to obtain a measurement result. The metric together with the measurement result provide knowledge about the measured property.
2.5.2 The use of metrics

Software metrics are very useful in software engineering because they can provide valuable information required by both engineers and managers to make better decisions (Westfall, 2005). This is also emphasized by McWhirter & Gaughan (2012) who say that “metrics should be a constant staple for all decision making” and can “give stakeholders confidence in the use and performance of their services”.

One of the essential characteristics of cloud computing is being a “measured service” (Mell & Grance, 2011). According to NIST (2015), in order “to describe a measured service, one needs to identify the cloud service properties that have to be measured and what their standards of measurement or metrics are”. The same source also says that “a metric provides knowledge about characteristics of a cloud property through both its definition and the values resulting from the observation of the property”.

In the context of cloud services, metrics are important for selecting the cloud provider so that the customer’s expectations are met (Bardsiri & Hashemi, 2014). Garg, et al. (2011) state that considering the diversity of cloud services, it can be a challenge for a customer to decide what cloud service would best satisfy his requirements. Besides selecting the cloud service, metrics can provide support for decision making in activities such as “defining and enforcing service agreements, monitoring cloud services, accounting and auditing” (NIST, National Institute of Standards and Technology, 2015).

2.5.3 Objects of measurement

“To measure, we must first determine the entity” (Westfall, 2005). This is a valid statement because it is necessary to first identify the entity for which some property needs to be measured and then select the adequate metric.

According to Basili, et al. (1994), “objects of measurement can be products, processes and resources”. Resources are the objects necessary as input for the processes. Processes are the development activities which result in producing some products. The products are the artifacts produced during development. This paper was written in 1994 and it is obvious that the author is referring to software
development projects. Nevertheless, his approach is still valid for modern software development in the context of cloud computing where there are entities which could be classified as resources, processes and products.

In a more recent paper, Westfall (2005) proposes a similar approach inspired by the one belonging to Basili, et al. (1994). The model of “input – process – output” is used to distinguish types of software entities. Input entities represent all the resources used for software development and research. Process entities can be various software activities and events that take place during the software development lifecycle. The author states that very often process entities are related to a time factor. Therefore, temporal aspects of processes entities could be measured. Finally, the output is made of artifacts, documents and other deliverables. Westfall implies that each software entity can have multiple properties that can be measured and therefore it is necessary to have a strategy for selecting what metrics to use.

2.5.4 Good metrics

It seems that metrics can provide many benefits. However, creating a metric, utilizing it for measurement and analyzing measurement results requires effort and resources. A metric without a purpose is not a useful metric. If no one is going to use the metric then there is no point producing it in the first place (Westfall, 2005).

NIST (2015) referring to cloud computing services, states that metrics definitions should be reusable. This would allow building composite metrics from other metrics previously defined. It is an efficient way to reduce the amount of duplicate information.

Not all software metrics are related to quality, only a subset of them are. Very often the software quality metrics are more concerned with the process and product than with the project (Kan, 2002).

2.6 Defining metrics using Goal-Question-Metric paradigm

Defining software metrics systematically is an important part of the software measurement process (Westfall, 2006). It requires prior planning to establish a well-defined structure to be followed throughout the overall project. This ensures effective software measurement process and more relevant metrics to the domain of study (Basili, et al., 1994). The concepts and best practices used when applying software measurement is usually based on some of the already established software quality factors (McCall, 1994) and mechanisms (Haag, et al., 1996).

In a paper published in 1994 (Basili, et al., 1994), the authors present an approach called Goal Question Metric as a more efficient mechanism to extract software metrics. It was developed originally for evaluating defects for a set of projects in NASA Goddard Space Flight. The GQM is based on the principle that purposeful measurement is based on clearly defined goals. Therefore, an organization should first identify its goals and only after select metrics and measure how these goals are being achieved (Basili, et al., 1994). The GQM approach allows the definition and extraction of software metrics in a top-down strategy by focusing on, why we
are defining the metrics, rather than focusing on what software metrics should we define. There are too many software metrics, thus, it is more efficient to let the goals drive the definition and selection of metrics. Basili et al. (1994) state that research on applying metrics and models in the industry has shown that effective measurements should be focused on precise goals.

The GQM approach is used by modern researchers to establish a software engineering measurement process which includes defining metrics (Becker, et al., 2015). It is mainly based on identifying stakeholders’ goals, then asking questions related to these goals. Finally, metrics are selected to answer the questions. Accordingly, the measurement model has three levels as described by Basili et al. (1994):

1. The conceptual level: which is the goal defined usually for a specific object, such as, products, processes, and resources. Different points of views can be taken into consideration when defining the goals. General goals are not restricted to a specific point of view.
2. The operational level: the questions that are used to describe the object of measurement according to a quality factor or quality issue as the authors of the paper refer to it.
3. The quantitative level: it is mainly the answer to the questions from the previous level in a quantitative way. This will form the required software metric to measure the effectiveness of reaching the specified goal.

Figure 8 illustrates how the GQM approach is structured and the relations between each level in the model.

Figure 8. The structure of GQM (Basili, et al., 1994)

Identifying the goals is the starting point in creating the metrics. Basili et al. (1994) propose a structured approach to specifying the goals. In this approach, the goal has a purpose and three coordinates which are the issue, the object, and the viewpoint. The issue can be thought of as a quality attribute. The object is the entity to which the goal applies. The viewpoint is a stakeholder that expressed the goal. Figure 9 illustrates that the purpose is based on the three coordinates. An example of a goal defined in this way could be:

- Purpose: Minimize
- Issue: File upload time
- Object: Data analysis tool
- Viewpoint: Data scientist
Figure 9. The goal’s coordinates (Basili, et al., 1994)

For each a goal a quantifiable set of questions is generated. The questions should be focused on measurement. It is possible for a question to be common for multiple goals. Reasoning can be provided for each question to facilitate better understanding of its utility. (Loconsole, 2000)

The questions help to refine the goals and make them more focused. Also, the questions ensure that the measurement process will collect only data that shows the progress towards achieving the goals. A goal is not viable for measurement if it’s impossible to come up with questions or if it’s impossible to collect data for it. Such goals should be discarded. (Baumert & McWhinney, 1992)

To answer the questions in a quantitative way it is necessary to have metrics that would generate quantitative information (Loconsole, 2000). There can be objective and subjective metrics. What metric to use depends on the object that is measured. Basili et al. (1994) suggest to use objective metrics on more mature objects of measurement while subjective metrics can be used on informal objects or unstable objects.

The metrics provide information necessary to make intelligent decisions. Therefore, metrics selection or creation should be practical and realistic. Clearly defined metrics minimize misunderstandings about how the metrics should be used. A metric can perform one of four functions: Understand, Evaluate, Control and Predict (Westfall, 2005). In this paper, we are interested in metrics that help to understand and evaluate ML cloud services. Terminology used for the metrics should be consistent.
A very important part in creating the metrics is choosing the measurement function or formula. Basic metrics, also called metric primitives, are directly measured and they consist of a single variable. There are also complex metrics which represent a mathematical combination of several base metrics or complex metrics. A mathematical combination is basically a formula that uses multiple variables. A metric also has a unit of measurement. The function and the unit of measurement make up a measurement model. It’s possible to create your own measurement models or use existing ones. (Westfall, 2005)

The GQM process is iterative in its nature because “questions refine goals, metrics refine questions, ability to obtain data refines metrics” (Dow, 2007). The GQM is only a set of guidelines to define metrics, however, it doesn’t have explicit rules for when the metric creation/definition process should be stopped (Locomobile, 2000).

2.7 Stakeholders in projects using ML cloud services

ML cloud services combine two concepts which are ML and Cloud computing. This means that in a project using MLaaS there might be stakeholders concerned with both ML and Cloud computing. Stakeholders in a ML or data mining projects that don’t involve MLaaS are also potential stakeholders for projects that use these services.

A simple search of the term “machine learning engineer” on the job search engine website www.indeed.com on date 2017-05-20 returned almost 8351 hits. Many of these jobs also contain the words “developer” or “software engineer” in their heading. Some of them also contain the word “data scientist”. Looking at few job descriptions some common requirements for a ML engineer is to have experience with ML or data science, to have programming skills and to know technologies that can be used to implement ML solutions.

There is a lot of confusion regarding the data scientist role (Rose, 2016). The reason for this is that data science is a multidisciplinary field. The data scientist role is hard to define. The data scientist is involved in all steps of a data science project such as defining the business problem, getting and preparing the data, developing the model, deploying the model and monitoring how the model performs. He must understand well the data, know some statistics and math, apply machine learning techniques, know how to write code and have a hacker mindset. And very important he should be able to ask interesting questions. It is very hard to find a person who is able to be proficient at everything, so it’s a good idea to have a data science team with complementary skills (Barga, et al., 2015). Rose (2016) even suggests that the tasks of a data scientists could be split across several roles.

The CRISP-DM (Chapman, et al., 2000) mentions multiple stakeholders involved in various steps of a data mining project such as data mining engineer, business analyst, data analyst, database administrator, domain expert, statistician, system administrator etc. Unfortunately, there is no detailed description of what most of these stakeholders are supposed to do. The focus is on the data mining engineer, who is basically involved in all the steps of the CRISP-DM process such as
understanding the business, understanding the data, preparing the data, modeling, evaluating the results and deploying the model. Notice that it is basically the same process that a data scientist follows.

It’s intuitive to call someone who works on a ML project a ML engineer, on a data science program a data scientist, and someone who works on a data mining project a data mining engineer. These roles often mean the same thing because ML, data science, and data mining are related to each other to the point the terms are used interchangeably. Professionals who perform these roles are probably the main potential users and stakeholders for ML cloud services. To avoid confusion, in the rest of the paper, there will be no strict distinction between a ML engineer, data scientist, and data mining engineer.

Very often it is expected from a ML engineer to have knowledge about software development because he’ll have to apply ML into existing or new software systems. MLaaS target software developers as potential users by making ML easy to use (Nketah, 2016). Software developers can be involved in ML/data science/data mining projects even if their knowledge about ML is limited. Therefore, a software developer can be an important stakeholder as well.

Professionals who work with cloud computing such as technology architects or cloud resource administrators (Erl, et al., 2013) are also potential stakeholders and can have goals for MLaaS. These are just some example of professionals who can act as stakeholders in a project using MLaaS. Anyone who intends to use MLaaS is a stakeholder and can express goals regarding the MLaaS that would serve as a basis to develop the metrics. Furthermore, someone who is using MLaaS is actually doing the job of a ML engineer/data scientist/data mining engineer even if his official role is different. This is true even if the person doesn’t have advanced competence in the field of ML. After all, one of the main goals of MLaaS if to make ML accessible even to those with limited experience.

2.8 Previous research
2.8.1 Creating metrics for Cloud Services
Defining metrics and measuring cloud services is not a trivial task. The need for having an industry standard method for measuring and comparing cloud services has prompted the development of The Service Measurement Index Framework. The framework was created by the Cloud Services Measurement Index Consortium (CSMIC) and its purpose is to help organizations compare cloud services from multiple providers or even compare cloud-based vs non-cloud services. The SMI has a hierarchical structure. The first level consists of seven main characteristics of cloud services which are accountability, agility, assurance, financials, performance, security and privacy, and usability. The second level contains attributes of cloud services which can be measured. Each attribute belongs to one of the seven characteristics previously mentioned. Therefore, the characteristics serve as categories to split the attributes. Metrics can be defined for each attribute. The framework does not contain any explicit metrics, the
responsibility of developing the metrics being on the users of SMI. (Siegel & Perdue, 2012) (Cloud Services Measurement Initiative Consortium, 2014)

Garg et al. (2011) use the SMI to create the SMICloud framework which can be used to select the appropriate cloud service based on the quality of service requirements. The SMICloud defines metrics for some of the quantitative attributes present in the SMI. A description is provided for each metric. Some of the metrics also have an explanation for how they were created. The framework describes a mechanism for ranking the cloud services using the measurements taken for different SMI attributes.

Becker et al. (2015) utilize the GQM approach to derive metrics for quality properties of cloud services such as scalability, elasticity, and efficiency. The process of deriving the metrics is performed in a systematic way by following the steps in the GQM approach with a small addition. The addition is an example scenario of a cloud application, including its requirements for scalability, elasticity, and efficiency. The scenario is used to derive exemplary metrics. A general goal related to all the properties mentioned above is set. Next, several questions are defined that can help achieve the goal. The questions are grouped based on the quality property that they address. Then, exemplary metrics are defined to answer the questions in the context of the example scenario. General metrics applicable to cloud services are derived from the exemplary metrics. These metrics are based on visible external properties of the cloud services. All the metrics result in ordinal numbers to allow quantifying the quality properties. At the end, the authors do a review of related work to discover more metrics for scalability, elasticity, and efficiency. For illustrative purpose, they identify some metrics that answer the questions defined with GQM.

Most of the evaluation work on cloud services delivers benchmarking results for various metrics (Li, et al., 2013). However, this does not necessarily make it easier for customers to understand the general picture for a cloud service. It would be useful to have a summary of all the benchmarking results. To address this issue, Li et al. (2013) propose the Boosting Metrics approach using inspiration from ML field. Metrics that are used directly to measure various aspects of the cloud services are combined to create a boosting metric. A boosting metric uses the results from other metrics to produce a single result. The purpose of the boosting metric is not to replace the metrics for individual aspects but to complement them. Sometimes it can be useful to compare different cloud services by using a single metric. A boosting metric can also be used for measuring a complex feature with many properties. Some examples of boosting metrics are the mean (arithmetic, geometric etc.) and the radar plot (Li, et al., 2013).

MLaaS is a form of SaaS. This idea is supported by Pop (2016), who did a survey on machine learning and cloud computing SaaS solutions and categorized services such as BigML and Google Prediction API as SaaS. Therefore, it might be interesting to look at research concerned with defining metrics for SaaS.

Due to demand for measuring the quality of SaaS, Lee et al. (2009) present a quality model for SaaS comprised of several metrics. They initially identify key
features of SaaS by evaluating some research papers on cloud computing. These key features are reusability, data managed by providers, customizability, availability, scalability and pay-per-use. Next, they derive the following quality attributes from the features: reusability, efficiency, reliability, scalability, availability. The multiplicity mapping between key features and quality attributes is many-to-many. Several metrics are defined for the attributes. Each metric is described along with the formula, value range and how it should be interpreted. The quality metric model is validated by conducting an assessment using IEEE 1061.

2.8.2 Metrics and Evaluation of ML Cloud Services

At the time of writing this paper, there is not much research published regarding ML Cloud Services and evaluation of these types of services. The reason for this could be that many of these services are new and have appeared in the last 2-3 years. Nevertheless, it was possible to find materials that cover the topic.

Ribeiro et al. (2015) propose an architecture to create a MLaaS platform with the focus on predictions. Their architecture would facilitate the MLaaS to be scalable, flexible and non-blocking. The authors emphasize the usefulness of having such a platform considering the increased demand for data analysis.

Dorard (2015) compares Amazon ML, Google Prediction API, PredicSis and BigML. The comparison was done using a real-world dataset to build the models. The dataset is from Kaggle “Give me some credit” challenge. His comparison was focused on the model training and prediction using the trained model. He used the free tier for all 4 services. The services were compared across 3 metrics: Area Under the Curve (AUC), Time for training in seconds, Time for predictions in seconds. There was no service winning across all 3 metrics. However, the author concludes that “PredicSis offered the best trade-off between accuracy and speed by being the second fastest and second most accurate”. An interesting disclaimer that the author mentions is that the results might be different if another dataset is used and the users should test them with their own data to find out which is better for their needs.

Baskar et al. (2016) do some experiments to compare Azure ML and Amazon ML. They analyzed the services in terms of scalability, robustness and performance. The metrics they used are AUC (%) and time in minutes to build and validate the model. Their experiments showed that Amazon ML has a higher AUC compared to Azure ML, however, it is slower when building and validating the model. In their research, they also briefly describe Google Prediction API and IBM Watson Analytics.

Nketah (2016) performs a comparison of three ML services which are Amazon ML, Google Prediction API and Azure ML Studio to supply information that could help developers select which service fits them best. The author does a quantitative and qualitative analysis of the services. Aspects of the services such as mode of operation, data processing, prediction, model creation, cost and algorithm are considered. The same dataset is used for experiments to run predictions. Some of the metrics used during these experiments are AUC,
Training time, Predictions time. In the conclusion of his research, the author acknowledges that it is likely to have different results if a different dataset is used. Also, he concluded that no service was an obvious winner. However, he enumerates multiple factors in chapter 4 of his work that could be important when selecting a service. One example of such a factor could be the size of the training dataset accepted by the service.
3 Method and implementation

The following section describes the research method and implementation to answer the main research question.

3.1 Research method

**Design and Creation** is a research method that can be used when an IT artifact must be developed (Oates, 2006). The artifact can represent a construct, model, method or working system. The new knowledge is the artifact itself or the lessons learned during the development process. Also, the artifact can be a “vehicle for something else” (Oates, 2006) in which case the new knowledge will be learned from that “something else”, for example, if the main purpose of the research is to see how the created IT artifact can help its users. Design and Creation involves the following stages which can be performed iteratively (Oates, 2006):

- **Awareness** – the stage where the researcher identifies and formulates the problem. The source of the problem can be either literature or industry.
- **Suggestion** – the stage where the researcher comes up with an idea to solve the problem. It’s basically the plan of how the researcher is going to deal with the problem.
- **Development** – at this stage the researcher implements the idea. The researcher must use an appropriate development methodology based on the type of artifact that he intends to create. The development methodology is different than the research methodology. It is possible to use already established development methodologies.
- **Evaluation** – at this stage the researcher must perform an evaluation of the created artifact. The criteria by which the artifact is evaluated depends on the purpose of the research. The evaluation process can result in conclusions about the created artifact or about the development process itself. During the evaluation, the researcher might realize that the artifact or the development process needs additional improvements.
- **Conclusion** – this is the final stage where the researcher analyzes the results, presents the new knowledge and suggests further research.

For a project using Design and Creation to be considered as academic research, it must include analysis, justification, critical evaluation and it must produce new knowledge. Design and Creation can use multiple data generation methods. For example, interviews or questionnaires can be utilized to extract requirements from users. Same methods can also be used to evaluate the developed artifacts, for example, by asking the users opinion about the artifacts.(Oates, 2006)

To answer the main question of the thesis, “What metrics can be used to evaluate ML Cloud Services?”, the research method Design and Creation was used. The rationale behind this choice is that the new knowledge, which is the metrics to evaluate ML cloud services, were considered as artifacts that had to be designed and created.
Here is an explanation of how the research fits into the structure of the Design and Creation method:

- The researcher realized that there is a lack of metrics dedicated for ML cloud services. The problem was formulated as one main research question with two sub-questions. This is the Awareness stage.
- The researcher described how the questions were going to be answered and which data collection methods were used. The GQM paradigm was selected as a development methodology to develop the metrics. This is the Suggestion stage.
- The researcher collected the necessary data and developed the metrics using GQM. More details about this in the Implementation section. This is the Development stage.
- The researcher addressed questionnaires to professionals. The questionnaires aimed to evaluate the metrics based on two simple criteria: relevance and ease of use. The researcher also compared the created metrics with existing metrics in literature research. This is the Evaluation stage.
- Finally, the researcher presented the conclusion and discussion of the created metrics and suggested further research. This is the Conclusion stage.

3.2 Implementation

3.2.1 What goals related to the ML Cloud Services the stakeholders have?

Relation to GQM:

According to the GQM approach, defining the goals is the starting point in creating the metrics. A goal has a purpose and three coordinates. Information that makes up the goal such as issue, object and purpose can be extracted from the stakeholders. Normally, the stakeholder represents the viewpoint. However, the viewpoint was not specified when defining the goals. This is because the purpose of the research was not to create specific metrics for each type of stakeholder, but instead, create a general set of metrics applicable to ML cloud services that any stakeholder can use. Defining the goals is part of the Conceptual Level in the GQM process. Once the goals are defined, the Conceptual Level in the GQM process is considered complete.

Data generation method:

Interviews were conducted with professionals from companies who are using or are interested in using ML cloud services. Such professionals represent stakeholders for ML cloud services. Companies were contacted to learn if the topic is relevant to them and if they were willing to do an interview. More details about the interviews follow.

Interview target: Given the limited time and resources, it was not possible to exhaustively interview all types of stakeholders for a project using MLaaS. Therefore, the focus was on stakeholders who are directly using or are interested in using ML Cloud Services. These are the users or potential users for MLaaS. It is
not strictly necessary for the interviewed stakeholders to have extensive experience with ML because it is also interesting to interview people who are just beginning with ML to find out how they would like to use MLaaS. Stakeholders who are indirectly affected by the usage of MLaaS, for example, a business analyst, were not interviewed.

**Interview purpose:** Identify the goals that a stakeholder has for MLaaS in a data mining or ML project. The interview contains questions about how the stakeholder uses ML in general because this information is also relevant for MLaaS and would help to define goals. A goal is usually defined for a specific object: resource, process, product. The MLaaS, their inputs, and outputs resemble resources, processes, and products.

**Interview structure:** The interview is semi-structured. There are several topics covered during the interview and the questions are grouped by topic. The topics and questions are based on the theoretical framework, mainly on Machine Learning, CRISP-DM, and Cloud Computing. There are two sets of interview questions. The first set is addressed to people who have limited experience with ML/data mining or are new to the field. The second set is addressed to people who are experienced with ML or data mining. The decision to choose a set was made based on the answers received from the introduction questions which gave the interviewer an idea about the knowledge level of the person being interviewed. The interviews were audio recorded with permission of the interviewee.

Introduction questions:
1. What is your role at the company?
2. What experience do you have with ML, data mining or data analysis?

Depending on the answers received from the introduction questions, Set A or Set B of questions was used. Set A is contained in Appendix 1 and Set B is in Appendix 2.

**Findings:**
Stakeholders’ answers to the interview questions (interview data).

**Analysis:**
A qualitative analysis was performed on the collected data to identify goals that the stakeholders have. The collected data was in the form of audio records and/or interview notes. To make the analysis process easier, all audio data was transcribed as text. Interview transcripts were divided into several sections. Each section was labeled depending on the information that it contains. This information and additional notes were used to define the goals.

Goals were expressed according to the GQM's format of 1 purpose – 3 coordinates. For each goal, the issue, object and purpose was specified. The viewpoint was not specified to keep the goals general. The data such as transcript sections or notes based on which the goal was defined is referenced.

**Outcome of analysis:**
The list of identified goals.
3.2.2 What metrics can be used to determine the effectiveness in meeting the goals?

Relation to GQM:

Once the goals were identified, it was possible to proceed to the Operational Level of GQM, which is to ask questions about the goals. A question describes the object of a goal in regard to the quality issue. There can be several questions for one single goal. Next, comes the Quantitative Level, that is creating metrics which answer the questions in a quantitative way. The multiplicity relationship between questions and metrics is many-to-many.

The expected steps to define the metrics are:
1. Ask questions that need to be answered to ensure that each goal is being achieved.
2. Define metrics that can answer the questions.
3. Use questionnaires to evaluate identified metrics.

Data generation methods:

Artifact creation. The process of Design and Creation method results in creating some artifact which can be considered as a form of data. In this case, the produced artifacts were a list of metrics. The created metrics can be applied to ML Cloud Services. The metrics are linked to the goals through the questions. The structure for the metric definition is:

- Name of metric.
- Metric definition – the formula and unit of measure.
- Reused metrics (optional) – other metrics that are part of the formula. This is applicable in the case of complex metrics that reuse base or complex metrics.
- Influencing factors (optional) – additional parameters that influence the results obtained with the metric.
- Comments (optional).
- Examples (optional).

Questionnaires addressed towards previously interviewed professionals in which they evaluated the created metrics based on two criteria: relevance and ease of use.

Findings:
The list of created metrics and the stakeholders’ evaluation of the metrics.

Analysis:

A qualitative analysis was performed on the created metrics and the questionnaire results to draw conclusions about them. The metrics were compared to similar metrics encountered in the research literature. To make the list of metrics more comprehensible, a systemization of metrics was performed. The metrics were divided into categories and sub-categories.
A qualitative analysis was performed on the questionnaire results to draw conclusions about the stakeholders’ opinion of the created metrics. It was not possible to do a quantitative analysis on the questionnaire data because the number of people who took the questionnaire was too small.

**Outcome of analysis:**

A systemization of the metrics and conclusions regarding them.
4 Findings and analysis

4.1 What goals related to the ML Cloud Services the stakeholders have?

4.1.1 Interview details

To arrange interviews, the researcher contacted companies whom he had known are using ML or are interested in using it. Most of the contacted companies are in Jönköping area because the researcher also resides there. The researcher explained to them the purpose of the research and asked to interview a professional from their company who is working with ML or maybe is interested in it. Three different companies replied positively. An interview was set with each of them. For confidentiality purposes, the names of the companies are kept secret. They are referred as company A, B and C. The full name of the individuals is also kept secret. Instead of their real name, the name of the company followed by a number is used. For example, the 3 people from company C that were interviewed are referred as C1, C2, and C3. Their professional role at the company is mentioned.

All the interviews were audio recorded with the permission of the people being interviewed. The audio interviews were transcribed as text and are annexed to this paper as appendices. Each transcript has line numbers to make it easy to refer to a specific a line or fragment of the interview. Table 1 shows how the interview transcripts are abbreviated.

Table 1. Abbreviations for interview transcripts

<table>
<thead>
<tr>
<th>Company</th>
<th>Interview</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Interview with A1</td>
<td>ale</td>
</tr>
<tr>
<td>B</td>
<td>Interview with B1</td>
<td>eri</td>
</tr>
<tr>
<td>C</td>
<td>Interview with C1, C2 and C3</td>
<td>sum</td>
</tr>
</tbody>
</table>

Some examples of how interview transcripts will be referenced can be seen in Table 2.

Table 2. Examples of referencing interview transcripts

<table>
<thead>
<tr>
<th>Fragment of interview text</th>
<th>Referencing notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 12 from the interview with A1.</td>
<td>ale12</td>
</tr>
<tr>
<td>Line 45 to 48 from the interview with B1.</td>
<td>eri45-48</td>
</tr>
<tr>
<td>Line 9, line 70, and line 110 to 115 from the interview with A1.</td>
<td>ale9, 70, 110-115</td>
</tr>
</tbody>
</table>

The first interviewed person was A1 from company A. His official role at the company is Senior Consultant and what he does is “anything and everything with regards to data”. To be more precise, he is an experienced Database Administrator (DBA) and a Data Analyst (ale23-28). His experience with MI is limited but he is interested in it and he has put together a few ideas regarding what could be done with ML at his company (ale32-44). Based on this information, Set A of interview questions was used.
The second interviewed person was B1 from company B. His official role at the company is Business Intelligence (BI) Consultant. His work is focused on data visualization and ML (eri10-14). They are just starting with ML at the company. However, he has been taking an ML course in the past one and a half year and he also has a personal project where he is using ML (eri17-22). Based on this information, Set B of interview questions was used.

The last interview was a group interview because three persons from the company C participated. Their names are C1, C2 and C3. C1 is the Chief Technology Officer (CTO) of the company C. C2 is a Software (SW) Developer focused on high-level programming (sum24-26). C3 is an Embedded Software (SW) Developer who works “close to the metal, to the hardware” (sum17-23). They’ve taken some courses in ML and they understand the business use of ML. They are working to add ML capabilities to their systems and services. Based on this information, Set B of interview questions was used.

The interviewed people are considered as potential stakeholders in a project involving MLaaS. The professional roles for the interviewed people are summarized in Table 3. If a person is using ML cloud services directly then it is reasonable to say that they he is doing the job of a ML engineer and the goals he expresses would be from the viewpoint of a ML engineer. To avoid confusion, the viewpoint was not specified for the goals. Therefore, the goals are not restricted to a specific stakeholder/viewpoint.

<table>
<thead>
<tr>
<th>Company</th>
<th>Name</th>
<th>Stakeholder role</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A1</td>
<td>Data Analyst/DBA</td>
</tr>
<tr>
<td>B</td>
<td>B1</td>
<td>BI consultant</td>
</tr>
<tr>
<td>C</td>
<td>C1</td>
<td>CTO</td>
</tr>
<tr>
<td>C</td>
<td>C2</td>
<td>SW Developer</td>
</tr>
<tr>
<td>C</td>
<td>C3</td>
<td>Embedded SW Developer</td>
</tr>
</tbody>
</table>

By analyzing the interview data, a set of goals was identified. Many of the goals were common for multiple stakeholders. Supporting evidence and explanation for how each goal was identified is provided. There are quality goals and functional goals. The quality goals are concerned with a quality attribute of the service, while the functional goals are concerned more with a specific functionality or feature of the service. The quality goals are expressed using the 1 purpose – 3 coordinates structure. The functional goals are expressed as simple sentences.

4.1.2 Quality goals
C3 requires that the ML cloud service should be easy to use (sum190-192, 208). A1 suggests that such a service can be useful for prototyping and it should be easy to get started with (ale218-221). B1 already is using Azure ML for prototyping (eri226-231). The goal here is concerned with the ease of use of the ML service. The goal can be formulated as:
GQ1. Maximize (purpose) usability (issue) of the ML service (object).

A1 says if the service is not available then it is useless. That’s the tradeoff you have with the cloud, you rely completely on it (ale248, 225-226). C3 emphasizes that the service “should always be there” because otherwise it’s not useful and this would be detrimental to their customers (sum208, 219-223). The goal can be formulated as:

GQ2. Maximize (purpose) availability (issue) of the ML service (object).

A1 mentions one important aspect regarding cloud services in general: integration. Very often they will prefer to use services from one cloud provider just because they are all integrated and work seamlessly together. Therefore, integration capabilities of the ML service with other cloud services is important (ale182-184, 192-200). This goal can be formulated as:

GQ3. Ensure (purpose) integrability with cloud services (issue) for the ML service (object).

A1 wants to be able to import data from different types of data sources (ale220). By different types of data sources, it is assumed he meant relational databases, NoSQL databases, files etc. His goal can be formulated as:

GQ4. Ensure (purpose) compatibility with data sources (issue) ML service (object).

A1 says that the volume of their data can range up to hundreds of terabytes which is no small number (ale57). C1 says that they get a lot of continuous real-time data so the volume of data is quite high (sum51-53). The goal that can be identified here is that the ML service should be able to deal with big volumes of data. The goal can be formulated as:

GQ5. Maximize (purpose) accepted size (issue) of the data (object).

A1 says that he has a lot of data located on premise. This data needs to be imported into the ML service tool to use in the next steps. This process should be as quick as possible (ale87-88, 91). The goal that can be formulated as:

GQ6. Maximize (purpose) performance (issue) of the data import (object).

A1 want the creation of the ML model to be fast (ale131). The goal can be formulated as:

GQ7. Maximize (purpose) performance (issue) of the ML model creation/training (object).

B1 emphasizes multiple times that the comprehensibility of the models is important to their customers. He firmly states that they need to have an explanation for how the model produces the output. They can’t give their customers a black-box and expect them to trust it (eri103-110, 165-167). When asked if he would prefer a white-box model to a black-box model A1 says that some of his clients would need to see what’s inside the box as well (ale150, 150-156). The goal here is:
**GQ8. Facilitate (purpose) comprehensibility/understandability (issue) of the ML model (object).**

All the interviewed persons want to integrate the created ML models into systems (ale165-171, 181-185) (sum164-165) (eri185-187). This goal can be formulated as:

**GQ9. Ensure (purpose) integrability (issue) of the ML model (object).**

C3 want to be able to train the model in the cloud and then deploy it to their sensor, which is also a tiny computer (sum56, 118). A goal that results from this information is:

**GQ10. Ensure (purpose) portability (issue) of the ML model (object).**

B1 referring to predictive models says that it is nice to have good numbers for precision, recall, and accuracy. He is interested in checking the quality of the model. However, he admits that at the end it’s the customer who decides if the results are good or not (eri148). One potential goal here can be:

**GQ11. Evaluate (purpose) the quality (issue) of the ML model (object).**

Referring to the predictive web services that can be created with the MLaaS, B1 says directly that such a service must be fast. He doesn’t want to wait too long for the “magic” (eri213-214, 196-198). The people from company C say that they need to get a quick response from a predictive web service as the real-time data comes in (sum231, 235-239). The goal can be formulated as:

**GQ12. Maximize (purpose) performance (issue) of the predictive web service (object).**

C3 says that scalability is important because you can start small and you don’t have to make big upfront investments. At their company C, they are dealing with real-time data, so the volume can get quite high, which means that they expect the predictive web service to be able to deal with that (sum195). This goal can be formulated as:

**GQ13. Maximize (purpose) scalability (issue) of the predictive web service (object).**

B1 says that the reason he prefers to use Azure is because of good prices (eri207). One of the most important criteria when evaluating cloud services for the company C is the price (sum185-189, 215-217). Minimizing the cost is a general goal that applies to the usage of the whole MLaaS. The goal can be formulated as:

**GQ14. Minimize (purpose) cost (issue) of the ML service usage (object).**

There are several sub goals related to the cost factor. A1 states his concerns regarding the cost of transferring big amounts of data in case the data is imported from on premise or from other cloud storage services (ale57, 192-200, 203, 204). The goal in this case is:

**GQ15. Minimize (purpose) cost (issue) of data import (object).**

The interviewed stakeholders emphasized that cost can be a decisive factor (eri207) (sum185-189, 215-217). The model training/creation is a very important
step in the whole ML workflow. Therefore, the cost for this individual task should also be as low as possible. The goal in this case is:

**GQ16. Minimize (purpose) cost (issue) of model training (object).**

B1 says they will make requests to the predictive web service and even though they don’t expect to do many requests in the beginning, minimizing the cost is desired (eri174-176). The goal here is:

**GQ17. Minimize (purpose) cost (issue) of the predictive web service usage (object).**

C2 and C3 don’t want to bother too much with the maintenance once they start using the cloud because now they are doing everything on their own servers and must deal with all the maintenance (sum193-194, 209). Their goal can be formulated as:

**GQ18. Reduce (purpose) maintenance (issue) for ML service (object).**

### 4.1.3 Functional goals

A1 says he is interested in predictive analysis because it can be useful in production and manufacturing industries (ale59). C1 say that they want to use it for predictive maintenance (sum38-40). B1 needs to make predictions (eri28-34). The goal is:

**GF1. Perform predictive analysis.**

A1s says that clustering is useful in the retail industry (ale61). This idea is also supported by B1 who says that clustering is good for customer segmentation (eri28-31). C2 says they use clustering techniques as part of their anomaly detection so they would expect that as well from the MLaaS (sum124-125). The goal is:

**GF2. Perform clustering analysis.**

A1 says anomaly detection is useful for bank transactions (ale60). People from company C say that the first thing they would like to do with ML is anomaly detection. They stress out this several times during the interview (sum37, 40, 163, 217). The goal can be formulated as:

**GF3. Perform anomaly detection.**

A1 mentions that he might need to collaborate with other people when working on a project (ale103-104). His goal is:

**GF4. Allow collaboration on ML project.**

A1 needs tools to check for data issues and deal with them. The tools can be part of the ML service (ale109-111). B1 says that he might need to do feature engineering to get the results he wants (eri68-71, 96-99). The goal here is:

**GF5. Perform data preprocessing.**
B1 is using now some R libraries to visualize the data (er10, 135-136). C1 mentions using some tools for data visualizations (sum75). A goal that these stakeholders can have for the ML cloud services is:

**GF6. Allow data visualizations.**

A1 says that a graphical interface for the ML service can be useful for prototyping (ale129-130). This idea is also supported by B1 (eri117-119). This common goal can be formulated as:

**GF7. Use ML service through a graphical interface.**

A1 want to be able to use code for the ML workflow to make the model creation process “quick, efficient and continuous” (ale132-134). The goal is:

**GF8. Use ML service through code.**

B1 mentions several algorithms to create ML models. For example, a decision tree is easy to explain to customers. He is successfully using Bayes algorithm for a personal project (eri106-108). C3 says k-NN (k nearest neighbors) is a candidate algorithm they consider for anomaly detection (sum110). It would be useful if the ML service allows selecting the algorithm for model creation. A goal that can be identified here is:

**GF9. Have multiple algorithms for model creation**

Company C would like to update/retrain their ML models with new data at certain time intervals (sum231, 235-237). In an email, they also mentioned that they would like to add features to the training set without losing the ML model training. So, it’s obvious that they would like to retrain the model as new data becomes available. The goal is:

**GF10. Retrain ML model.**

A1 mentions that evaluating the ML model is required, however, he does not mention any specific criteria for the evaluation (ale139-142). In his projects, B1 uses both cross-validation and normal validation (holdout method) to get a “wider angle” (eri108, 159-164). The goal can be formulated as:

**GF11. Evaluate ML model.**

A1 says he wants to use the ML service for predictions (ale59-60, 222-235). B1 thinks that it’s convenient to create a ML model and “store” it as a web service. This makes it easy to later integrate it with a system (eri129-130, 172-173). C2 would like to have a REST API to integrate it with their services (168-169, 227). The goal identified here is:

**GF12. Create predictive web service.**

### 4.1.4 Summary of all goals

The quality goals are summarized in Table 4 and the functional goals are summarized in Table 5. As previously mentioned, the functional goals are
expressed as simple sentences and do not comply with the 1 purpose – 3 coordinates structure of the GQM.

Table 4. Quality goals

<table>
<thead>
<tr>
<th>ID</th>
<th>Purpose</th>
<th>Issue</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>GQ1.</td>
<td>Maximize</td>
<td>Usability</td>
<td>ML service</td>
</tr>
<tr>
<td>GQ2.</td>
<td>Maximize</td>
<td>Availability</td>
<td>ML service</td>
</tr>
<tr>
<td>GQ3.</td>
<td>Ensure</td>
<td>Integrability with cloud services</td>
<td>ML service</td>
</tr>
<tr>
<td>GQ4.</td>
<td>Ensure</td>
<td>Compatibility with data sources</td>
<td>ML service</td>
</tr>
<tr>
<td>GQ5.</td>
<td>Maximize</td>
<td>Accepted size</td>
<td>Data</td>
</tr>
<tr>
<td>GQ6.</td>
<td>Maximize</td>
<td>Performance</td>
<td>Data import</td>
</tr>
<tr>
<td>GQ7.</td>
<td>Maximize</td>
<td>Performance</td>
<td>ML model creation/training</td>
</tr>
<tr>
<td>GQ8.</td>
<td>Facilitate</td>
<td>Comprehensibility/understandability</td>
<td>ML model</td>
</tr>
<tr>
<td>GQ9.</td>
<td>Ensure</td>
<td>Integrability</td>
<td>ML model</td>
</tr>
<tr>
<td>GQ10.</td>
<td>Ensure</td>
<td>Portability</td>
<td>ML model</td>
</tr>
<tr>
<td>GQ11.</td>
<td>Evaluate</td>
<td>Quality</td>
<td>ML model</td>
</tr>
<tr>
<td>GQ12.</td>
<td>Maximize</td>
<td>Performance</td>
<td>Predictive web service</td>
</tr>
<tr>
<td>GQ13.</td>
<td>Maximize</td>
<td>Scalability</td>
<td>Predictive web service</td>
</tr>
<tr>
<td>GQ14.</td>
<td>Minimize</td>
<td>Cost</td>
<td>ML service usage</td>
</tr>
<tr>
<td>GQ15.</td>
<td>Minimize</td>
<td>Cost</td>
<td>Data import</td>
</tr>
<tr>
<td>GQ16.</td>
<td>Minimize</td>
<td>Cost</td>
<td>Model training</td>
</tr>
<tr>
<td>GQ17.</td>
<td>Minimize</td>
<td>Cost</td>
<td>Predictive web service usage</td>
</tr>
<tr>
<td>GQ18.</td>
<td>Reduce</td>
<td>Maintenance</td>
<td>ML service</td>
</tr>
</tbody>
</table>

Table 5. Functional goals

<table>
<thead>
<tr>
<th>ID</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>GF1.</td>
<td>Perform predictive analysis</td>
</tr>
<tr>
<td>GF2.</td>
<td>Perform clustering analysis</td>
</tr>
<tr>
<td>GF3.</td>
<td>Perform anomaly detection</td>
</tr>
</tbody>
</table>
4.2 What metrics can be used to determine the effectiveness in meeting the goals?

Using the goals identified during the previous step, it’s possible to ask questions about them and then define metrics that can answer the questions. Some of the defined metrics use well-known units of measurement (UoM), for example, minutes. It is up to the user of the metric to decide if he wants to use a smaller or larger unit than the one specified for the metric.

4.2.1 Questions and metrics for quality goals

**GQ1. Maximize usability of the ML service**

**Q1. How easy is it to use the ML service?**

**M1. Usability rating of a user**

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = A$ value ranging from 1 to 5, with 1 being very easy to use and 5 being very hard to use.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>Integer value in [1, 5]</td>
</tr>
<tr>
<td>Comments</td>
<td>Usability is a very subjective attribute. That is why the metric is also subjective. It can still be useful because people ask other people's opinion all the time. The issue with this metric is that only one user is asked, therefore it is not very reliable.</td>
</tr>
<tr>
<td>Examples</td>
<td>Max is using Amazon ML. He thinks it is quite easy to use it. His usability rating $M = 2$.</td>
</tr>
</tbody>
</table>

**M2. Average usability rating of a group of users**

<table>
<thead>
<tr>
<th>Formula</th>
<th>$N$ – total number of users $M = \frac{\sum_{i=1}^{n} \text{Usability rating of a user } i}{N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>Rational value in [1, 5]</td>
</tr>
<tr>
<td>Reused metrics</td>
<td>M1. Usability rating of a user</td>
</tr>
</tbody>
</table>
Q2. Does the ML service provide a graphical interface?

M3. Has graphical interface

<table>
<thead>
<tr>
<th>Formula</th>
<th>[ M = \text{True if the ML service has a graphical interface, False otherwise.} ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric shows if the ML service has a graphical interface which can be used for the whole ML workflow.</td>
</tr>
</tbody>
</table>

Q3. Is there a comprehensive documentation for the ML service?

M4. Has documentation

<table>
<thead>
<tr>
<th>Formula</th>
<th>[ M = \text{True if the ML service has documentation or user manual, False otherwise.} ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>A documentation is expected by default from any cloud service nowadays. It can be very helpful to any user of the service, no matter his skill level.</td>
</tr>
</tbody>
</table>

GQ2. Maximize availability of the ML service

Q4. How often is the ML service unavailable?

M5. ML Service outage frequency

<table>
<thead>
<tr>
<th>Formula</th>
<th>[ M = \frac{\text{Number of outages}}{\text{Time period}} ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>outage/week</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric shows how often was the service unavailable during a fixed time period. This metric is only concerned with how often a ML service has outages, but not with how much they last. Getting a big number for this metric would mean that the ML service is unstable.</td>
</tr>
<tr>
<td>Examples</td>
<td>During 4 weeks, the ML service has 10 outages. The following value is obtained using the metric: [ \frac{10 \text{ outage}}{4 \text{ week}} = 2.5 \text{ outage/week} ]</td>
</tr>
</tbody>
</table>

Q5. How long is the ML service unavailable?

M6. ML service availability

<table>
<thead>
<tr>
<th>Formula</th>
<th>[ M = \frac{T_{\text{available}}}{T_{\text{available}} + T_{\text{unavailable}}} \times 100% ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_{\text{available}} - time period the service is available</td>
<td></td>
</tr>
<tr>
<td>T_{\text{unavailable}} - time period the service is unavailable</td>
<td></td>
</tr>
</tbody>
</table>
UoM | % (percentage)
---|---
Comments | This is a standard metric used for the availability of cloud services in general. The users of this metric should consider checking the SLA agreement offered by the cloud provider. Most likely they will specify an availability value that they guarantee, for example, 99.9% availability.

**M7. ML service maximum outage duration**

| Formula | \( T_i - time \text{ duration of an outage } i \)
| --- | ---
| \( M = \max(T_i) \)
---|---
UoM | minute (min)
Comments | Applying the metric involves monitoring the service during a fixed time period, for example, during one month. This metric shows the longest outage experienced in that time period.
Examples | During one month, the ML service had 3 outages: first lasted 1 min, the second lasted 20 min and the third one lasted 3 min. 
\( T_1 = 1 \text{ min}, T_2 = 20 \text{ min}, T_3 = 3 \text{ min} \)
\( M = \max(1 \text{ min}, 20 \text{ min}, 3 \text{ min}) = 20 \text{ min} \)

**Q6. Is the ML service completely or partially unusable during outages?**

**M8. Availability of a specific service functionality during outages**

| Formula | \( TF_{\text{available}} - time \text{ period while the service functionality was available during several outages} \)
| --- | ---
| \( T_{\text{outages}} - time \text{ period while the outages lasted} \)
| \( M = \frac{TF_{\text{available}}}{T_{\text{outages}}} \times 100\% \)
---|---
UoM | %
Comments | This metric checks if a specific functionality of the ML is still usable even during an outage of the ML service. An outage of the ML service means that the whole ML service or some of its individual functionalities are not available. For example, during an outage, it is not possible to train models but the predictive web services are still working.
Examples | During one month time, there were several outages that lasted for total 72 min. During the outages, the predictive web services were still available for 20 min.
\( TF_{\text{available}} = 20 \text{ min}, T_{\text{outages}} = 72 \text{ min} \)
\( M = \frac{20 \text{ min}}{72 \text{ min}} \times 100\% = 0.2777\% \)
GQ3. Ensure integrability with cloud services for the ML service

Q7. Can the ML service fetch the data from storage services of cloud providers?

M9. Integration with storage services of a specific cloud provider

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = \text{True if the ML service can fetch data directly from the storage services of a specific cloud provider; False otherwise.}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>It's convenient to be able to fetch the data directly from cloud storage services. The result produced by this metric can be a critical factor for someone who must decide between several ML services.</td>
</tr>
<tr>
<td>Examples</td>
<td>Check if the Amazon ML can fetch data directly from Azure SQL databases.</td>
</tr>
</tbody>
</table>

Q8. Does the ML service provide a REST API?

M10. REST API to use the ML service

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = \text{True if the ML service can be used through a REST API; False otherwise.}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>A REST API would facilitate integration with other cloud services.</td>
</tr>
</tbody>
</table>

GQ4. Ensure compatibility with data sources for the ML service

Q9. What types of data sources can the ML service use directly?

M11. Compatibility with a specific type of data source

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = \text{True if the specific type of data source is supported; False otherwise}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>If a data source is supported by the ML service it means the data can be imported from it directly. The term data source can have a very broad meaning. It can mean a SQL database, a NoSQL database, a file. It’s possible to be even more specific because there are different SQL databases, different types of file formats etc.</td>
</tr>
<tr>
<td>Examples</td>
<td>Check if Azure ML supports Excel files.</td>
</tr>
<tr>
<td></td>
<td>Check if Google ML supports MySQL databases.</td>
</tr>
</tbody>
</table>

M12. Compatibility ratio with required data sources

<table>
<thead>
<tr>
<th>Formula</th>
<th>$\text{DS}<em>{\text{required}}$ – set of data sources required by the consumer to be compatible with the ML service. $\text{DS}</em>{\text{compatible}}$ – subset of $\text{DS}_{\text{required}}$ that are compatible with the ML service.</th>
</tr>
</thead>
</table>

45
\[ M = \frac{|DS_{compatible}|}{|DS_{required}|} \times 100\% \]

**UoM**  
\%  

**Comments**  
This metric is useful for someone who intends to use a wide variety of data sources. \(|DS_{required}|\) means cardinality of the set \(DS_{required}\).

**Examples**  
John is a consultant whose customers use various databases such as MySQL, PostgreSQL, Oracle, Neo4J. A specific ML service he intends to use is compatible only with MySQL and Oracle from this set.  
\[ DS_{required} = \{MySQL, PostgreSQL, Oracle, Neo4J\} \]  
\[ DS_{compatible} = \{MySQL, Oracle\} \]  
\[ M = \frac{2}{4} \times 100\% = 50\% \]

**GQ5. Maximize accepted size of the data**

**Q10. What is the maximum data size accepted?**

**M13. Maximum size of training data (in terms of digital storage)**

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{The size of the data that is accepted by the ML service.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UoM</strong></td>
<td>Gigabyte (GB).</td>
</tr>
<tr>
<td><strong>Comments</strong></td>
<td>This metric measures the maximum size of training data accepted by the service in terms of digital storage, irrespective of the type of data source. Ideally, the size of accepted data should be infinite. However, in practice, users of this metric should check for a specific limit they need.</td>
</tr>
<tr>
<td><strong>Examples</strong></td>
<td>Check if the ML service accepts training data with the size of 4GB.</td>
</tr>
</tbody>
</table>

**Q11. What is the maximum number of data records accepted?**

**M14. Maximum number of records in the training data**

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{The number of records in the training data that is accepted by the ML service.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UoM</strong></td>
<td>A non-negative integer value.</td>
</tr>
<tr>
<td><strong>Comments</strong></td>
<td>Ideally, the maximum number of records in the training data accepted by the ML service should be infinite. However, in practice, users of this metric should check for a specific limit they need.</td>
</tr>
<tr>
<td><strong>Examples</strong></td>
<td>Check if the ML service accepts training data containing 100000 records.</td>
</tr>
</tbody>
</table>
Q12. What is the maximum number of data features accepted?

M15. Maximum number of features in the training data

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = \text{The number of features in the training data accepted by the ML service}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A non-negative integer value.</td>
</tr>
<tr>
<td>Comments</td>
<td>Each data record is usually characterized by a fixed set of features. There can be datasets that have a big number of features.</td>
</tr>
<tr>
<td>Examples</td>
<td>Check if the ML service accepts training data that has 1000 features/attributes.</td>
</tr>
</tbody>
</table>

GQ6. Maximize performance of the data import

Q13. How fast is the data import from on premise storage?

M16. Average import speed of data located on premise storage

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = \frac{\text{Size of data located on premise}}{\text{Import time}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>Megabyte/second</td>
</tr>
<tr>
<td>Influencing Factors</td>
<td>Internet speed, read speed of the storage device on which the data is located.</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric can be useful in case large volumes of data must be imported. The word average is used in the name of the metric because the speed can vary in small intervals.</td>
</tr>
<tr>
<td>Examples</td>
<td>Importing 1000 MB of data into Amazon ML tool 2 min. (Data size = 1000MB, \text{ Import time} = 120 \text{ sec}) (M = \frac{1000 \text{ MB}}{120 \text{ sec}} = 8.33 \text{ MB/sec})</td>
</tr>
</tbody>
</table>

Q14. How fast is the data import/load from cloud storage?

M17. Average import speed from a specific cloud storage service

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = \frac{\text{Size of data located on a specific cloud storage}}{\text{Import time}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>megabyte/second</td>
</tr>
<tr>
<td>Influencing factors</td>
<td>Data transfer speed of the cloud storage service</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric is useful to check how fast does the ML service import/load data from a specific cloud storage service of a cloud provider.</td>
</tr>
</tbody>
</table>
**GQ7. Maximize performance of ML model creation/training**

**Q15. How long does it take to create/train the ML model?**

**M18. Model training time**

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{The time duration it takes to train the model.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>second</td>
</tr>
<tr>
<td>Influencing factors</td>
<td>Dataset (training data), algorithm.</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric is useful to see how quick can the ML service train a model. This metric is useful to compare multiple ML services, by using a fixed dataset and algorithm (or at least an algorithm that achieves a similar purpose, for example creating a predictive model).</td>
</tr>
</tbody>
</table>

**Q16. How quick is the ML model creation process?**

**M19. Model training speed**

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \frac{\text{Training data size}}{\text{Model training time}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>megabyte/second</td>
</tr>
<tr>
<td>Reused metrics</td>
<td>M18. Model training time</td>
</tr>
<tr>
<td>Influencing factors</td>
<td>Dataset, algorithm.</td>
</tr>
<tr>
<td>Comments</td>
<td>This purpose of this metric is to show how fast the ML service trains the model considering the data size factor. This metric is useful to compare between multiple ML services, by using a fixed dataset and algorithm (or at least an algorithm that achieves a similar purpose, for example creating a predictive model).</td>
</tr>
</tbody>
</table>

**GQ8. Facilitate comprehensibility/understandability of the ML model**

**Q17. Does the ML service allow to create white-box models?**

**M20. Number of algorithms to create white-box models**

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{The number of distinct algorithms to create white-box models that the ML service offers.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A non-negative integer value.</td>
</tr>
<tr>
<td>Comments</td>
<td>Very often an algorithm can be customized with additional parameters, for example, the depth of decision tree etc. Variations of the same algorithm will not be considered as a different algorithm because this leads to a potentially unlimited number of options.</td>
</tr>
</tbody>
</table>
GQ9. Ensure integrability of the ML model

Q18. What technologies is the ML model compatible with?

M21. Model compatibility with a specific technology

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{True if the created model is compatible with a specific technology, False otherwise.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>The reason for this metric is to show if the created ML model can be used in combination with a specific technology or programming language. To integrate a ML model into a system it must be compatible with the technology that the system was built on. However, the term “compatible” can have a broad meaning.</td>
</tr>
<tr>
<td>Examples</td>
<td>Check if a predictive ML model created using Google ML can be used with C# code.</td>
</tr>
</tbody>
</table>

GQ10. Ensure portability of the ML model

Q19. What environments/platforms support the ML model?

M22. ML model compatibility with a specific environment/platform

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{True if it’s possible to deploy the ML model to a custom environment outside the ML service, False otherwise} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>The reason for this metric is to check if the ML model can be run/executed on a specific environment or platform.</td>
</tr>
<tr>
<td>Examples</td>
<td>Check if the created ML model can be executed on Linux. Check if the created ML model can be executed on Java Virtual Machine (JVM).</td>
</tr>
</tbody>
</table>

M23. Size of the ML model

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{The size of the model.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>Megabyte (MB)</td>
</tr>
<tr>
<td>Influencing factors</td>
<td>Dataset, algorithm.</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric is useful in case the model must be deployed to devices with a limited amount of memory. Use this metric only if the ML service allows to deploy the model to a custom environment.</td>
</tr>
</tbody>
</table>

GQ11. Evaluate the quality of the ML model.

Q20. How good is the ML model?

There are several metrics that can be used for this purpose. A comprehensive list of these metrics will not be provided because they are well described in various literature sources. Examples of such metrics are
accuracy, precision, recall, F-score, ROC curve etc. (Tan, et al., 2006). The idea is that these metrics are also applicable to ML models created with MLaaS.

**Q21. How much space does the created model require?**

The metric M23. Size of the model can answer this question.

**GQ12. Maximize performance of the predictive web service**

**Q22. How fast is the predictive web service?**

**M24. Average response speed for individual prediction requests**

| Formula | \( N \) - number of individual prediction requests sent \\
|---------|-------------------------------------------------------------------
|         | \( T_i \) - the time between sending a request \( i \) to the predictive web service and getting a response \\
|         | \( M = \frac{\sum_i^n T_i}{N} \) \\
| UoM     | millisecond (ms) / request \\
| Influencing factors | Internet connection speed, ML model behind the predictive web service \\
| Examples | Three requests have been made to the predictive web service. The response times for each request are 200 ms, 100 ms and 160 ms. \\
|          | \( T_1 = 200 \text{ms} \), \( T_2 = 100 \text{ms} \), \( T_3 = 160 \text{ms} \) \\
|          | \( M = \frac{200 + 100 + 160 \text{ms}}{3 \text{request}} = 153.3 \text{ms/request} \) \\

**M25. Maximum response time for individual prediction requests**

| Formula | \( N \) - number of individual prediction requests sent \\
|---------|-------------------------------------------------------------------
|         | \( T_i \) - the time between sending a request \( i \) to the predictive web service and getting a response \\
|         | \( M = \text{max}(T_i) \) \\
| UoM     | millisecond (ms) \\
| Influencing factors | Internet connection speed, ML model behind the predictive web service \\
| Examples | Three requests have been made to the predictive web service. The response times for each request are 200 ms, 100 ms and 160 ms. \\
|          | \( T_1 = 200 \text{ms} \), \( T_2 = 100 \text{ms} \), \( T_3 = 160 \text{ms} \) \\
|          | \( M = \text{max}(200 \text{ms}, 100 \text{ms}, 160 \text{ms}) = 200 \text{ms} \) \\

**M26. Response time for a batch request**

| Formula | \( M = \text{The time period between sending the batch request to the predictive web service and receiving the response} \) \\
<p>| UoM     | millisecond (ms) |</p>
<table>
<thead>
<tr>
<th><strong>Influencing factors</strong></th>
<th>Internet connection speed, ML model behind the predictive web service.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comments</strong></td>
<td>A simple metric that shows the total time it takes for ML service to process the batch request and return the results.</td>
</tr>
</tbody>
</table>

**M27. Response speed for a batch request**

<table>
<thead>
<tr>
<th><strong>Formula</strong></th>
<th>[ M = \frac{\text{Response time for the batch request}}{\text{Number of instances in the batch request}} ]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UoM</strong></td>
<td>milisecond /instance</td>
</tr>
<tr>
<td><strong>Reused metrics</strong></td>
<td>M26. Response time for a batch request</td>
</tr>
<tr>
<td><strong>Influencing factors</strong></td>
<td>Internet connection speed, ML model behind the predictive web service</td>
</tr>
<tr>
<td><strong>Comments</strong></td>
<td>This metric attempts to show how quick the ML service produces the results for an individual instance in the batch request. It makes it possible to compare with the results obtained with the metric M24.Average response speed for individual requests.</td>
</tr>
</tbody>
</table>
| **Examples** | The predictive web service returned the results in 2 minutes for a batch request containing 1000 instances.  
  \[ \text{Response time for the batch request} = 2 \text{ min} = 120000 \text{ms} \]  
  \[ \text{Number of instances in the batch request} = 1000 \text{ instances} \]  
  \[ M = \frac{120000 \text{ ms}}{1000 \text{ instances}} = 120 \text{ ms/instance} \] |

**GQ13. Maximize scalability of the predictive web service.**

**Q23. Can the predictive web service handle a workload increase?**

**M28. Maximum number of requests that the predictive web service can handle**

<table>
<thead>
<tr>
<th><strong>Formula</strong></th>
<th>[ M = \text{The number of simultaneous requests to the predictive web service that do not result in the service breaking its SLA or visible degradation of performance.} ]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UoM</strong></td>
<td>A non-negative integer value.</td>
</tr>
<tr>
<td><strong>Comments</strong></td>
<td>This metric shows the ability of the predictive web service to handle several requests. Ideally, the service should be able to handle an unlimited number of requests. However, in practice, users of this metrics should check for a specific limit they need.</td>
</tr>
<tr>
<td><strong>Examples</strong></td>
<td>Check if the predictive web service can handle 200 simultaneous requests.</td>
</tr>
</tbody>
</table>
M29. Maximum number of instances in the batch request that the predictive web service can handle

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{number of instances in the batch request that does not result in the service breaking its SLA or visible degradation of performance} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A non-negative integer value.</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric shows the ability of the predictive web service to handle a batch request containing many instances. Ideally, the service should be able to handle a batch with an unlimited number of instances. However, in practice, users of this metrics should check for a specific limit they need.</td>
</tr>
<tr>
<td>Examples</td>
<td>Check if the predictive web service can handle a batch request with 10000 instances.</td>
</tr>
</tbody>
</table>

GQ14. Minimize cost of ML service usage

Q24. What is the total cost for using the ML service?

M30. Total cost for ML service usage

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{Cost}<em>{\text{data import}} + \text{Cost}</em>{\text{model training}} + \text{Cost}_{\text{predictive web service usage}} + \text{Other costs} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>$</td>
</tr>
<tr>
<td>Reused metrics</td>
<td>M32. Cost for data import, M35. Cost for model training, M37. Cost for predictive web service usage</td>
</tr>
<tr>
<td>Comments</td>
<td>This is a simple metric that computes the total cost of using the ML service by summing the cost incurred for some of the individual tasks. For simplicity reasons, it is assumed that the cost for data pre-processing is included in the cost for model training.</td>
</tr>
</tbody>
</table>

Q25. Is there a free tier plan for the ML service?

M31. Has free tier

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{True if the ML service offers a free tier, False otherwise.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>A free tier means that the ML service can be used for free, with some certain limitations. Such a feature of the service can valuable for someone who intends to perform small ML tasks or just try out the service.</td>
</tr>
</tbody>
</table>

GQ15. Minimize cost of data import

Q26. What is the cost for data import?

M32. Cost for data import

| Formula | \( M = \text{The cost to import the data} \) |
Influencing factors

Comments

This is a simple metric that shows the total cost incurred to import the data. This metric is suitable for comparison of several ML services.

Q27. What is the price to import data from on premise storage?

M33. Price for importing data from on premise

Formula

\[ M = \frac{\text{Cost to import the data from on premise}}{\text{Data size}} \]

UoM

\$/Gigabyte

Comments

Importing/uploading/loading the data to be used for model training can cost money. This metric takes into consideration the data size as the influencing factor for how the pricing is calculated. It is possible that a such a value is directly specified in the SLA.

Examples

Importing a database of 2 Gigabytes cost $10.

Cost to import data from on premise = $10,

Data size = 2GB

\[ M = \frac{10}{2 \text{ GB}} = 5 \text{ $/GB} \]

Q28. What is the price to import data from storage services of a cloud provider?

M34. Price for importing data from storage service of a cloud provider

Formula

\[ M = \frac{\text{Cost to import the data from a cloud storage service}}{\text{Data size}} \]

UoM

\$/Gigabyte

Comments

Data transfer between services of different cloud providers is likely to incur a cost. It’s possible that a cost is incurred even if the data is located on the storage services of the same cloud provider. The issue with this metric is that a cost can be incurred both ways, on the ML service and on the storage service.

GQ16. Minimize cost of model training

Q29. What is the cost for model training?

M35. Cost for model training

Formula

\[ M = \text{The cost incurred to train the model} \]

UoM

$ 

Comments

This is a simple metric that shows the cost incurred to train the model.
Q30. What is the price for model training?

M36. Price for model training

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \frac{\text{Cost for model training}}{\text{Model training time}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>$/hour</td>
</tr>
<tr>
<td>Reused metrics</td>
<td>M18. Model training time, M35. Cost for model training</td>
</tr>
<tr>
<td>Influencing factors</td>
<td>Dataset, algorithm.</td>
</tr>
<tr>
<td>Comments</td>
<td>It is advised to use a fixed dataset and algorithm, or at least an algorithm that achieves a similar purpose, for example creating a predictive model.</td>
</tr>
<tr>
<td>Examples</td>
<td>Using Google ML, the cost incurred for model training is 4 $. The time it took to train the model is 30 minutes, which is 0.5 hours.</td>
</tr>
</tbody>
</table>

\[
\text{Cost to train the model} = 4 \$ \\
\text{Time to train the model} = 0.5 \text{ hour} \\
M = \frac{4 \$}{0.5 \text{ hour}} = 8 \$/ \text{hour}
\]

GQ17. Minimize cost of the predictive web service usage

Q31. What is the total cost for the predictive web service usage?

M37. Cost for the predictive web service usage

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{Cost for prediction requests} + \text{Cost to keep the predictive web service running} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>$</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric shows how much does it cost to use the predictive web service created for the ML model. The cost to keep the predictive web service running is the cost incurred during a period of time just for the fact the web service is active. Ideally, this would be 0 $.</td>
</tr>
</tbody>
</table>

Q32. What is the price of a single prediction request?

M38. Price for a prediction request

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \frac{\text{Cost for the request}}{1 \text{ request}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>$/\text{request}</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric shows the pricing for making individual prediction requests to the predictive web service. It is possible for the ML service to directly state such a value in the SLA.</td>
</tr>
</tbody>
</table>
Q33. What is the price of a batch (multiple predictions) request?

**M39. Price for a batch request**

**Formula**

\[
M = \frac{\text{Cost for making the batch request}}{\text{Number of instances in the request}}
\]

**UoM** $/\text{instance}$

**Comments** This metric is concerned with how the pricing is affected in case multiple predictions in the form of a batch are requested.

**Examples** A batch requests that contains 1000 instances has been sent to the predictive web service. The cost incurred for getting predictions for all the instances is 2$.

Cost for making the batch request = 2 $

Number of instances in the request = 1000 instances

\[
M = \frac{2 \text{ $}}{1000 \text{ instance}} = 0.002 \text{ $ / instance}
\]

Q34. What is the price to keep the predictive web service running?

**M40. Price for keeping the predictive web service running**

**Formula**

\[
M = \frac{\text{Cost for predictive web service usage} \text{− Cost for predictions}}{\text{Time period}}
\]

**UoM** $/\text{hour}$

**Comments** The reason for having this metric is that normally a web service consumes resources even while it's not used directly.

**Examples** During 24 hours, the total cost incurred for the predictive web service usage was 10 $. The total cost for predictions incurred during this time interval is 8 $.

Cost for predictive web service usage = 10 $

Cost for predictions = 8 $

Time period = 24 hours

\[
M = \frac{10 \text{ $} \text{−} 8 \text{ $}}{24 \text{ hour}} = 0.083 \text{ $/hour}
\]

GQ18. Reduce maintenance for the ML service

Q35. How much time does the user spend for maintaining the outputs of the ML process such as ML model, predictive web service etc.?

**M41. Maintenance effort**

**Formula**

\[
M = \frac{\text{Time for maintaining ML process outputs}}{\text{Total time using the ML service} \times 100}\%
\]

**UoM** %

**Comments** It is assumed that maintaining the ML process outputs is also done through the ML service. Therefore, the time spent on maintaining the ML process outputs is included in the total time using the ML service. This metric makes a ratio of the two. It
4.2.2 Questions and metrics for functional goals

GF1. Perform predictive analysis

Q36. Does the ML Service allow to perform predictive analysis?

M42. Number of algorithms for predictive analysis

<table>
<thead>
<tr>
<th>Formula</th>
<th>M = The number of distinct algorithms offered by the ML service that can be used for predictive analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A non-negative integer value.</td>
</tr>
</tbody>
</table>

GF2. Perform clustering

Q37. Does the ML Service allow to perform clustering?

M43. Number of algorithms for clustering

<table>
<thead>
<tr>
<th>Formula</th>
<th>M = The number of distinct algorithms offered by the ML service that can be used for clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A non-negative integer value.</td>
</tr>
</tbody>
</table>

GF3. Perform anomaly detection

Q38. Does the ML Service allow to perform anomaly detection?

M44. Number of algorithms for anomaly detection

<table>
<thead>
<tr>
<th>Formula</th>
<th>M = The number of distinct algorithms offered by the ML service that can be used for anomaly detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A non-negative integer value.</td>
</tr>
</tbody>
</table>

GF4. Allow collaboration on ML project

Q39. How many individuals can collaborate on a project?

M45. Maximum number of users working on a project

<table>
<thead>
<tr>
<th>Formula</th>
<th>M = The maximum number of users that can collaborate on a project supported by the ML service</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A non-negative integer value.</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric shows if the ML service allows several users to collaborate on a project (ML workflow). The value for this metric might be directly specific in the SLA.</td>
</tr>
<tr>
<td>Examples</td>
<td>The ML service allows 10 users to work on a project. M = 10 users</td>
</tr>
</tbody>
</table>
GF5. Perform data preprocessing

Q40. Does the ML service have capabilities to check the data quality?

M46. Has data quality tools

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = True$ if the ML service offers tools and functionalities to check for data quality, $False$ otherwise.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>Checking the data quality is an important step in the ML process. It is convenient if the ML service offers tools to check the data quality. In this way, the whole ML workflow is performed in one place.</td>
</tr>
</tbody>
</table>

Q41. Does the ML service have capabilities to perform feature engineering?

M47. Has feature engineering tools

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = True$ if the ML service offers tools and functionalities to perform feature engineering, $False$ otherwise.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>Performing feature engineering might be necessary to achieve better results for the created models.</td>
</tr>
</tbody>
</table>

GF6. Allow data visualizations

Q42. Does the ML service offer functionality to visualize the data?

M48. Has tools to visualize data

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = True$ if the ML Service offers tools to visualize the data, $False$ otherwise.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>Data visualizations can provide rich insights about the data.</td>
</tr>
</tbody>
</table>

GF7. Use ML service through graphical interface

Q43. Does the ML service provide a graphical interface?

The metric M3. Has graphical interface answers this question.

GF8. Use ML service through code

Q44. Can the ML service be used through a programmable API?

M49. Has a programmable API

<table>
<thead>
<tr>
<th>Formula</th>
<th>$M = True$ if the ML service can be used through a programmable API, $False$ otherwise.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value.</td>
</tr>
<tr>
<td>Comments</td>
<td>This metric checks if the ML service can be used through code. Being able to use a programmable API makes it possible to automate the ML workflow.</td>
</tr>
</tbody>
</table>
GF9. Have multiple algorithms for model creation

Q45. What algorithms to create ML models does the ML service offer?

M50. Has a specific algorithm

| Formula | M = True if the ML service offers a specific algorithm, False otherwise. |
| UoM     | A Boolean value. |
| Comments| Different problems require different approaches and algorithms. If someone needs a specific algorithm to use and the ML service doesn’t offer it, then the person might want to look at other services. |
| Examples| Check if Azure ML offers Naïve Bayes. |

GF10. Retrain ML model

Q46. Is it possible to retrain the ML model?

M51. Allows ML model retraining

| Formula | M = True if the ML service allows retraining the ML model, False otherwise. |
| UoM     | A Boolean value. |
| Influencing factors | Algorithm, ML model. |
| Comments | This metric checks if the ML service can retrain the model with additional data. Users of this metric should take into consideration that this can depend on the model itself and what algorithm was used. |

Q47. Is it possible to use data with additional features when retraining the model?

M52. Allows ML model retraining with additional features

| Formula | M = True if it’s possible to add additional features to the data when retraining the model without losing the model solution, False otherwise |
| UoM     | A Boolean value. |
| Comments| This metric was suggested by one interviewed stakeholder. He implies that it is likely that the number of features in the data they collect might increase in the future. This means that the new data used to retrain the model will have additional features. |

GF11. Evaluate ML model

Q48. What model evaluation methods does the ML service offer?

M53. Has a specific model evaluation method

| Formula | M = True if the ML service offers a specific model evaluation method, False otherwise |
| UoM     | A Boolean value. |
There are various model evaluation methods described in the literature such as holdout method, cross-validation, bootstrap etc.

Check if the ML service allows doing cross-validation.

GF12. Create predictive web service

Q49. Does the ML service allow to create a predictive web service for predictive ML models?

M54. Allows creating a predictive web service

<table>
<thead>
<tr>
<th>Formula</th>
<th>( M = \text{True if the ML service allows creating a predictive web service, False otherwise} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UoM</td>
<td>A Boolean value</td>
</tr>
</tbody>
</table>

4.2.3 Analysis and systemization of metrics

Two types of metrics are distinguished: quantitative metrics and Boolean metrics. The quantitative metrics, as the name implies, can be used to measure quantitatively some property or aspect of the ML cloud service. The Boolean metrics can be used to check if the ML service has a specific feature or functionality.

There are subjective metrics. For example, metrics \( M1 \) Usability rating of a user and \( M2 \) Average usability rating of a group of users depend entirely on the persons selected to provide the rating. The metric \( M41 \) Reduce maintenance effort for the ML service depends on the person who is using the service. The results obtained by using these metrics are subjective. Subjective metrics can be useful when it is not possible to measure objectively some aspect of the service.

There are metrics which have influencing factors. Examples of such metrics are \( M16, M17, M19, M23, M24, M25, M26, M36 \) etc. Influencing factors can affect the measurement result obtained by using the metric. The most common influencing factors are the dataset (training data) and the algorithm used to create the model. The list of influencing factors is not exhaustive. There might be additional factors which the researcher did not take into consideration when defining the metrics. Identifying all influencing factors would require additional research work.

There are multiple metrics which are concerned with the financial aspects of the MLaaS. Some of these metrics are related to cost while others are related to price. There is a difference between the two. The cost metrics measure the amount of money that was billed but they don’t measure how it was billed. The price metrics measure how the costs are calculated. Price metrics help to understand the amount of money that was paid or will be paid for individual resources or units of the MLaaS.

There are several metrics which have the word “maximum” in their name. The purpose of these metrics is to measure the limits of the MLaaS. Examples of such metrics are \( M13, M14, M15, M28, M29 \). The word “maximum” is a bit confusing
because, in practice, it might be difficult to measure the “maximum” and that is why it is recommended that the users of the metric just check for a specific limit they need. Just like in the example provided for the metric M28: “Check if the predictive web service can handle 200 simultaneous requests”.

The list of metrics is quite big and might be difficult to comprehend. Therefore, it is necessary to organize the metrics so that it is easier for a potential user of the metrics to find the metric that he needs. There are multiple ways to organize the set of metrics based on different criteria. In this case, the metrics were organized according to the aspect of the ML cloud services that they cover. Each aspect will form a category that will contain several metrics. The metrics have been divided into the following categories: ML capabilities, cost, availability, scalability, performance, integrability, usability. Some categories have subcategories to organize the metrics even further. The metrics are compared with similar metrics identified in the literature and previous research. Further analysis is provided for some of the metrics.

**ML capabilities**

This category contains metrics that measure or evaluate capabilities of the ML cloud services which are strictly related to ML.

a) **Algorithms**

- M50. Has a specific algorithm
- M42. Number of algorithms for predictive analysis
- M43. Number of algorithms for clustering
- M44. Number of algorithms for anomaly detection
- M20. Number of algorithms to create white box models

Nketah (2016) uses similar metrics to all the metrics listed in this subcategory to evaluate qualitatively ML cloud services. The main difference is that he uses Boolean metrics. For example, he does not count the number of algorithms for clustering but checks if such algorithms are offered by the service or not.

The metrics M42, 43 and 44 are related to some of the data mining tasks previously mentioned in the theoretical framework.

b) **Data preprocessing**

- M46. Has data quality tools
- M47. Has feature engineering tools
- M48. Has tools to visualize the data

Nketah (2016) uses a metric similar to M48 to check what visualizations tools are offered by the ML service.

The metrics only check if the service offers such tools, but they don’t evaluate the tools. For example, it is possible that two ML cloud services both offer data quality tools but these metrics do not allow to make a comparison between the tools. Additional metrics or criteria would be required for this.
c) Other

- M53. Has a specific model evaluation method
- M51. Allows model retraining
- M52. Allows model retraining with additional features
- M54. Allows creating a predictive web service

Nketah (2016) uses a metric like M53 to check what model evaluation options do the ML services offer.

Cost

This category contains metrics that measure the financial aspects of ML cloud services.

a) Data import

- M32. Cost for data import
- M33. Price for importing data from on premise
- M34. Price for importing data from storage service of a cloud provider

The metrics M33 and M34 are very similar. Having two different metrics for data import makes it possible to compare the pricing to import from on premise vs from cloud storage services. This is relevant because the interviewed revealed that the companies are still using both on premise and cloud storage.

In the comments of the metric M34, it is said that the cost can be incurred both ways, on the ML service and on the storage service. However, the metric doesn’t explicitly state what to use in the formula: the cost incurred only by the ML service or the total cost on both sides.

b) Model training

- M35. Cost for model training
- M36. Price of model training

c) Predictive web service

- M37. Cost for predictive web service usage
- M38. Price for a prediction request
- M39. Price for a batch request
- M40. Price for keeping the predictive web service running

Zheng Li et al. (2012) mention a metric called Cost over a Fixed Time. The metric incorporates the time factor into consideration when computing cost. It is similar to the metric M40.

The metrics M38 and M39 allow comparing price effectiveness of individual prediction requests vs batch requests. The metrics have different units of measure: $/request and $/instance. A prediction request is made for one instance while a batch request is made for multiple instances. For metric M38
it’s possible to consider that $/request is equivalent to $/instance which allows comparing with the results from M39.

d) Other
- M31. Has free tier
- M30. Total cost for ML service usage

Nketah (2016) checks if the ML services have a free tier. Therefore, he is using a metric similar to M31.

Zheng Li et al. (2012) mention Total Cost as a general metric applicable to cloud services. The metric is similar to M30. Also, they mention a metric called Component Resource Cost, which is basically the cost for different “components” that make up the service. This is similar to M32, M35, and M37 from the subcategories Data import, Model training and Predictive web service.

The metric M30 sums up all the costs incurred for the usage of the MLaaS. It adds up the cost for data import, model training and predictive web service usage and other costs. Other costs include additional costs which might be different from service to service.

**Availability**
This category contains metrics that measure the availability of the ML cloud services and their individual functionalities.

- M5. ML service outage frequency
- M7. ML service maximum outage duration
- M6. ML service availability
- M8. Availability of a specific service functionality during outages

Garg et al. (2011) describe a metric called Availability, which is the same as M6. Their metric is applicable to cloud services in general.

Becker et al. (2015) describe a metric called Number of SLO Violations. This metric is defined as the number of violating elasticity requirements in a fixed time unit. This metric covers a different attribute than availability, however, its formula has a similar structure to the formula of M5.

Applying the availability metrics in practice requires monitoring the service and its individual functionalities. Doing this process manually might prove to be difficult, therefore some form of automated monitoring would be better.

**Scalability**
This category contains metrics that measure the scalability of the ML services.

a) Volume of training data
- M13. Maximum size of training data
- M14. Maximum number of records in the training data
• M15. Maximum number of features in the training data

Nketah (2016) uses a very similar metric to M13. He checks the maximum size of data supported by different ML services.

As mentioned in the theoretical framework, data is the main input in the ML process. Therefore, it’s important to check the ability of the ML cloud service to handle increased amounts of data. Finding the actual maximum for any of these metrics might be difficult, so in practice, the users should just check for a specific value.

b) Predictive web service scalability

• M28. Maximum number of requests that the predictive web service can handle
• M29. Maximum number of instances in the batch requests that the predictive web service can handle

Becker (2015) define a metric called Scalability Range, which shows the maximum range of requests that the service can handle. It is like the metric M28. However, it is applicable for cloud services in general.

Performance

This category contains metrics that measure the performance of different tasks that can be done with the ML service.

a) Data import

• M16. Average import speed of data located on premise storage
• M17. Average import speed from a specific cloud storage service

b) Model training

• M18. Model training time
• M19. Model training speed

Dorard (2015) uses a metric called Time for Training to compare between multiple ML services. As the name implies, the metric measures the time it takes to train a ML model using the same dataset. It’s similar to M18. Nketah (2016) also uses this metric to compare between several ML services.

c) Predictive web service

• M24. Average response speed for individual prediction requests
• M25. Maximum response time for individual prediction requests
• M26. Response time for a batch request
• M27. Response speed for a batch request

Dorard (2015) uses a metric called Time for Predictions, that sums up the time it took to make 5000 predictions. The metric was used to compare between multiple ML cloud services. Nketah (2016) uses a similar metric.
Garg et al. (2011) describe a metric called Average Response Time, which has an almost identical formula to M24. They also describe a metric called Maximum Response Time which is related to M25, the difference being that it doesn’t measure anything, it just takes the value specified by the cloud service provider. Both Average Response Time and Maximum Response Time are applicable to cloud services in general.

M24 and M27 have different units of measure: ms/request and ms/instance respectively. However, it’s possible to compare the results obtained from these two metrics by considering the unit of measure for M24 as ms/instance. It is possible to do it because an individual prediction request is for one instance.

**Integrability**

This category contains metrics that measure integrability of the ML service and its outputs.

a) Data sources and services

- M9. Integration with storage services of a specific cloud provider
- M11. Compatibility with a specific type of data source
- M12. Compatibility ratio with required data sources

Nketah (2016) uses a combination of M9 and M11 to see what kind of data sources the ML services support.

Garg et al. (2011) describe a metric called Interoperability. This metric characterizes the “ability of the service to interact with services offered by the same provider or other providers”. It’s defined as a ratio between Number of platforms offered by the provider and Number of platforms required by users for interoperability. The formula of this metric has similar logic to the metric M12.

b) Interfacing options

- M10. REST API to use the ML service
- M49. Has a programmable API

c) ML model

- M21. Model compatibility with a specific technology
- M22. Model compatibility with a specific environment platform
- M23. Size of the ML model

Nketah (2016) checks if the ML model can be downloaded and transferred. This is related to the metric M22 and M23 because the “transfer” most likely means deploying the model to a custom environment platform.

M23 is the only quantitative metric for evaluating the model that was defined.

**Usability**

This category contains metrics that measure how easy it is to use the ML service.
• M3. Has graphical interface
• M1. Usability rating of a user
• M2. Average usability rating of a group of users

Nketah (2016) in his qualitative analysis of several ML services, check if they have a graphical interface. Therefore, he uses something like M3.

Garg et al. (2011) have a different approach to create metrics for usability. They split usability into four components which are operability, learnability, installability, and understandability. These individual components can be quantified using metrics that measure the average time it took previous users to operate, learn, install and understand respectively. Even though the metrics measure time, they are subjective because it depends on the users themselves. It’s implied that usability is subjective even if it’s measured using quantitative metrics.

Other
This category contains metrics that didn’t fit in any category previously mentioned.

• M4. Has documentation
• M41. Maintenance effort
• M45. Maximum number of users working on a project

4.2.4 Evaluation of the metrics using questionnaires
The created metrics were evaluated using a questionnaire addressed to the previously interviewed stakeholders. The aim of this questionnaire is to evaluate the set of metrics based on two criteria: relevance and ease of use.

The first criterion is the relevance of the metric. This criterion shows how relevant is this metric to the stakeholder or to his company. Basically, how important is the metric for him. For this criterion, the stakeholder must provide a rating value between 1 and 5, with 1 being that the metric is Not relevant to him and 5 being that the metric is Relevant to him. This criterion is used for both Quantitative metrics and Boolean metrics.

The second criterion is the ease of use. This criterion shows how easy the stakeholder thinks it is to apply the metric in practice. For this criterion, he must provide a rating value between 1 and 5, with 1 being that the metric is Very easy to use and 5 being that the metric is Very hard to use. This criterion is present only for quantitative metrics.

The stakeholder can mark the metric as not clear if he doesn’t understand it. This information is especially useful for showing which metrics are not clear or ambiguous and require additional refinement in the future.

To make the rating scores for Relevance and Ease of Use easier to understand, they are presented as ordinal values according to Table 6. Good metrics are those which have VH for both Relevance and Ease of Use.
### Table 6. Mapping of rating scores to ordinal values

<table>
<thead>
<tr>
<th>Ordinal value</th>
<th>Very Low (VL)</th>
<th>Low (L)</th>
<th>Medium (M)</th>
<th>High (H)</th>
<th>Very High (VH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance score</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Ease of Use score</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Unfortunately, it was not possible to get the questionnaire results from company A. B1 answered the questionnaire. Interviewed people from company C didn’t provide individual answers but instead provided a common answer. The questionnaire results are presented in Table 7. Some additional statistics for the results are presented in Table 8.

### Table 7. Questionnaire results

<table>
<thead>
<tr>
<th>ID</th>
<th>Metric</th>
<th>Relevance</th>
<th>Ease of use</th>
<th>Not clear</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B1</td>
<td>C</td>
<td>B1</td>
</tr>
<tr>
<td>M1</td>
<td>Usability rating of a user</td>
<td>VH</td>
<td>VL</td>
<td>H</td>
</tr>
<tr>
<td>M2</td>
<td>Average usability rating of a group of users</td>
<td>VH</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>M3</td>
<td>Has graphical interface</td>
<td>H</td>
<td>M</td>
<td>~</td>
</tr>
<tr>
<td>M4</td>
<td>Has documentation</td>
<td>H</td>
<td>VH</td>
<td>~</td>
</tr>
<tr>
<td>M5</td>
<td>ML service outage frequency</td>
<td>H</td>
<td>VH</td>
<td>M</td>
</tr>
<tr>
<td>M6</td>
<td>ML service availability</td>
<td>VH</td>
<td>VH</td>
<td>L</td>
</tr>
<tr>
<td>M7</td>
<td>ML service maximum outage duration</td>
<td>M</td>
<td>VH</td>
<td>M</td>
</tr>
<tr>
<td>M8</td>
<td>Availability of a specific service functionality during outages</td>
<td>M</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>M9</td>
<td>Integration with storage services of a specific cloud provider</td>
<td>VH</td>
<td>M</td>
<td>~</td>
</tr>
<tr>
<td>M10</td>
<td>REST API to use the ML service</td>
<td>VH</td>
<td>H</td>
<td>~</td>
</tr>
<tr>
<td>M11</td>
<td>Compatibility with a specific type of data source</td>
<td>H</td>
<td>M</td>
<td>~</td>
</tr>
<tr>
<td>M12</td>
<td>Compatibility ratio with required data sources</td>
<td>M</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>M13</td>
<td>Maximum size of training data</td>
<td>L</td>
<td>VL</td>
<td>H</td>
</tr>
<tr>
<td>M14</td>
<td>Maximum number of records in the training data</td>
<td>M</td>
<td>M</td>
<td>VH</td>
</tr>
<tr>
<td>M15</td>
<td>Maximum number of features in the training data</td>
<td>L</td>
<td>M</td>
<td>VH</td>
</tr>
<tr>
<td>M16</td>
<td>Average import speed of data located on premise storage</td>
<td>M</td>
<td>VL</td>
<td>H</td>
</tr>
<tr>
<td>M17</td>
<td>Average import speed from a specific cloud storage service</td>
<td>M</td>
<td>VL</td>
<td>H</td>
</tr>
<tr>
<td>M18</td>
<td>Model training time</td>
<td>VL</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>M19</td>
<td>Model training speed</td>
<td>VL</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>M20</td>
<td>Number of algorithm to create white box models</td>
<td>H</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>M21</td>
<td>Model compatibility with a specific technology</td>
<td>M</td>
<td>H</td>
<td>~</td>
</tr>
<tr>
<td>M22</td>
<td>ML model compatibility with a specific environment/platform</td>
<td>H</td>
<td>H</td>
<td>~</td>
</tr>
<tr>
<td>M23</td>
<td>Size of the ML model</td>
<td>L</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>M24</td>
<td>Average response speed for individual prediction requests</td>
<td>VH</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>M25</td>
<td>Maximum response time for individual prediction requests</td>
<td>VH</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>M26</td>
<td>Response time for a batch request</td>
<td>H</td>
<td>VL</td>
<td>M</td>
</tr>
</tbody>
</table>
4.2.5 Analysis of the questionnaire results
The first thing to notice by looking at the column “Not clear” inside Table 7 is that all the metrics were understood by the questioned people. There is no metric marked as not clear.

Table 8 shows that B1 gave high Relevance for 38 metrics out of 54, which is more than half of them. Only 5 metrics were given low Relevance. This means
that B1 found most of the metrics to be relevant. Looking in Table 7 at the metrics which got a high relevance score from B1 it seems that the most relevant metrics for him are related to aspects of the MLaaS such as ML capabilities, cost, predictive web service performance, availability and integration. The least relevant metrics for B1 are related to performance of model training, performance of data import and size of training data.

Company C gave high Relevance to 25 metrics. They gave low Relevance to 19 metrics which is much higher compared to B1 who gave low score to only 5 metrics. The most relevant metrics for C are related to availability, cost, ML capabilities, integration. The least relevant metrics for them are related to performance.

Metrics that got very high (VH) Relevance from both B1 and C are M6, M30, M37, M38, M40, M48, M49, M51. Among them are both Boolean and quantitative metrics. They are related to aspects such as cost, availability and ML capabilities.

The Ease of Use criteria was only considered for quantitative metrics. Table 8 shows that B1 gave low Ease of Use score to 12 metrics, while C gave low score to 11 metrics. The most difficult to use metrics for B1 are related to cost. The most difficult to use metrics for C are related to performance and usability.

The best metrics are those metrics which are relevant and easy to use. This equates to having a high Relevance and Ease of Use. The results in Table 8 for Relevance and Ease of Use show that overall the metrics are relevant to the stakeholders and are not very hard to use. However, given the fact that the number of respondents is small, it is not possible to generalize from these results. Therefore, the validity of the results is limited to the companies B and C.
5 Discussion and conclusions

5.1 Discussion of method

The objective of this research is to create a set of metrics that can be used to evaluate ML cloud services. To create the metrics, the research method Design and Creation was used. Design and Creation method is suitable when new artifacts must be created. In this case, the metrics are the artifacts. Design and Creation method also requires that a form of evaluation is applied on the created artifacts.

GQM paradigm was used as development method. The GQM paradigm presents a structured approach for the metric creation process. Applying GQM in practice was easy. Goals were collected from stakeholders. After that, questions were asked about the goals. Finally, metrics were created to answer the questions.

The interviews proved to be an effective way to collect goals from various stakeholders. Doing semi-structured interviews helped generate additional discussion with the stakeholders from which more goals were identified. However, the stakeholders would not always directly express goals. In such cases, the researcher had to make assumptions based on what the stakeholders said. An additional issue were the functional goals, which were hard to express using the 1 purpose – 3 coordinate structure of the GQM.

The goals acted as starting points in the metric creation process. Questions were asked for the goals based on the researcher’s own knowledge about the domain. Next, metrics that provide answers to the questions were created. The metrics are also based on the researcher’s knowledge about ML and cloud computing. This means that the metrics creation process is somewhat subjective even if it’s based on the goals of the stakeholders.

Because the metrics were developed based on the goals of the stakeholders, it was reasonable to ask the stakeholder’s opinion about the metrics. Questionnaires were used to evaluate the metrics by two simple criteria: relevance and ease of use. The evaluation was done by the same stakeholders who expressed the goals based on which the metrics were developed.

The GQM is an iterative process. It is recommended to involve the stakeholders in other steps of the metric creation process besides goal definition. It can be necessary to perform several iterations until an acceptable set of metrics is reached. However, for this research, it was not possible to do it due to limited availability of the stakeholders and the time limits for the research. Basically, only one iteration was performed. This is what could have been done better. Maybe it would have been easier to have a close collaboration with only one company so that the stakeholders are involved during the whole metric creation process. This would also allow performing several iterations of the GQM.

The companies contacted for interviews are from the near vicinity of Jönköping area because the researcher resided in Jönköping while doing the research. Interviews were arranged with 3 companies, which the researcher considered was enough for the purpose of the research. Most of the interviewed stakeholders are just starting with ML. Therefore, their experience is limited. This probably
influences the goals they expressed. It is hard to generalize the results of the research because the metrics were developed mainly based on the goals these stakeholders expressed and the researcher’s knowledge. This means that the research has low external validity.

The reliability of the research is ensured by the fact that the interviewed stakeholders are users or potential user of MLaaS. These stakeholders will be applying the metrics and making decisions. Therefore, it is reasonable to develop the metrics based on what is important for them.

5.2 Discussion of findings

The purpose of the research was to systematically define a set of metrics that can be used to evaluate ML cloud services. Therefore, the metrics are the main findings of the research. Other findings besides the metrics are the goals based on which the metrics were created and the evaluation of the metrics.

5.2.1 What goals related to the ML Cloud Services the stakeholders have?

By analyzing the interview data, several goals were identified. Two main types of goals were distinguished: quality goals and functional goals.

The quality goals are concerned with a quality attribute of the MLaaS or of some of its individual processes and products. This type of goals fit very well with GQM structure and it’s possible to express them with the 1 purpose - 3 coordinates (issue, object, and viewpoint) structure. An example of such goal can be GQ2 Maximize (purpose) availability (issue) of the ML service (object). Also, it is easy to notice that most of the defined goals have the purpose to minimize or maximize a certain quality attribute. The quality goals seem to be more related to the cloud aspects of the MLaaS. The goals common to all interviewed stakeholders are related to cost, usability, availability, integrability and performance of the ML cloud services.

The functional goals are concerned with a specific functionality of the service. They are similar to a functional requirement. They don’t fit well into the 1 purpose - 3 coordinates structure because they are not concerned with a specific quality attribute. An example of such a goal can be: GF4 Allow collaboration on ML project. These goals are expressed as simple sentences. Several stakeholders emphasized goals that revolve around predictive analysis such as GF1 Perform predictive analysis and GF12 Create predictive web service. The functional goals seem to be more related to the ML aspects of MLaaS.

5.2.2 What metrics can be used to determine the effectiveness in meeting the goals?

Based on the goals, several metrics were created. For some goals, it was not possible to come up with quantitative metrics. Instead, Boolean metrics were defined for such goals. The purpose of a Boolean metric is to check if the ML service has a specific feature or functionality. Even though these metrics are not quantitative they will not be discarded because they can make up a checklist used to evaluate a ML cloud service.
For the goal GQ11. Evaluate the quality of the model, it is stated that metrics that are traditionally used to evaluate ML models are also applicable to evaluate ML models created with the help of MLaaS. A comprehensive list of these metrics is not provided because they are well described in various literature sources.

An interesting idea is that many metrics that are used typically for cloud services are also applicable for ML cloud services. Metrics that are normally used to evaluate ML models are also applicable to ML models created with the help of ML cloud services. Nevertheless, the researcher defined some metrics which are specific for MLaaS.

It was possible to identify in the research literature metrics that cover similar aspects such as time to train the model, time to make predictions, accepted size of the dataset etc. Also, other researchers have used Boolean metrics to evaluate cloud services.

Evaluation of the metrics was done using questionnaires addressed to the same stakeholders that expressed the goals. No metric was marked as “not clear” which means that they understood the metrics’ definition. Also, the questionnaire results seem to suggest that overall the created metrics are relevant for these stakeholders and they don’t think it’s very hard to use them. However, given the fact that the number of respondents is small, it is not possible to generalize from these results.

5.3 Conclusions
ML is becoming more accessible than ever due to the ML cloud services. ML as a service promises to be a form of service provisioning that brings to the power of ML to the masses. Given that there are several providers now and is likely that other will appear in the future, the potential users need a way to compare and evaluate such services. In this research, a set of metrics was developed that can be used for this purpose. The metrics were developed based on the goals expressed by professionals using or interested in using these kinds of services. The metrics can be used to measure various aspects of the ML cloud services both from the ML perspective and from the cloud computing perspective. Most the metrics are quantitative. There is no priority given for the metrics, so it’s up to the user to decide which of these metrics to use and how to prioritize the results from different metrics.

During developing the metrics, the researcher realized that many aspects of the ML cloud services are hard to quantify, therefore it is hard to develop quantitative metrics for them. This was especially true when developing quantitative metrics for some of the functional goals expressed by the stakeholders. It might not be possible to perform a quantitative analysis of some aspects of the services. However, a qualitative analysis can be done by using Boolean metrics.

GQM paradigm was used to create the metrics. As previously stated, questions refine goals, metrics refine questions and the ability to collect data refines the metrics. GQM is an iterative process. In this research, one iteration was performed followed by an evaluation of the metrics. Performing more iterations would have allowed refining the metrics even further. It can be a good idea to
involve the stakeholders in the whole metric creation process because they would be able to provide valuable input throughout the whole process.

In future research, additional work can be done to refine the metrics even further. The created metrics have not been applied in practice yet. It would be interesting to test them out on several ML cloud services to measure the services, to see how easy it is to apply the metrics, and what kind of data can be collected using the metrics. The set of metrics has no priority for individual metrics. Research work can be done in this direction to create a framework that would prioritize the metrics and maybe define a set of conditions for how to combine results from several metrics.
6 References


Low, Y. et al., 2012. Distributed GraphLab: a framework for machine learning and data mining in the cloud. s.l., s.n., pp. 716-727.


74


7 Appendices

7.1 Set A - Interview questions for people with limited experience with ML/data mining or which are new to the field

(General)
1. ML can be used for predictions, anomaly detection, clustering etc. What would you like to use it for?

(Data storage and retrieval)
2. How large is the volume of data you are dealing with?
3. What types of data do you have (database records, images, audio, video etc.)?
4. How do you store and later retrieve the data for analysis?

(Data understanding and preparation)
5. What tools or techniques, including statistics, do you use to analyze and understand the data?
6. Data often comes with many issues such as missing values, duplicate values etc. How do you solve data quality issues?
7. Do you have labeled data?

(Model creation)
8. The data is usually fed into a ML algorithm that creates a model, which can be used for example to make predictions. Would you prefer to create such a model through a user interface or using a programming library?

(Model validation)
9. Would you like to evaluate and compare the produced models?
10. Many of the produced ML models are like a black-box. Is the comprehensibility of the model important for you?

(Deployment)
11. How would you like to apply the created models to your business decision processes?
12. What technological compatibility requirements do you have for the produced model so that you can integrate it with your systems?

(Cloud services)
13. Are you using cloud services to deal with the data?
14. What criteria do you normally use to evaluate cloud services?
15. MLaaS are a specific type of SaaS. What are your expectations from such a cloud service?
16. Many of the MLaaS allow creating a web service that offers predictions. What requirements would you have for such a web service?

7.2 Set B - Interview questions for people who are experienced with ML or data mining

(General)
1. ML can be used for predictions, clustering, anomaly detection etc. What do you use it for in your projects?
2. Do you have any prior experience with MLaaS?

(Data storage and retrieval)
3. How large is the volume of data you are dealing with?
4. What types of data do you have (database records, images, audio, video etc.)?

5. How do you store and later retrieve the data for analysis?
(Data understanding and preparation)

6. What tools or techniques, including statistics, do you use to understand the data or detect issues in the data?

7. How do you solve data quality issues?

8. Do you have labeled data?

9. What kind of transformations do you perform on the data (filtering, normalizations, feature engineering etc.)?
(Model creation)

10. What ML algorithms/techniques do you use?

11. How do you create ML models?

12. How do you store the created models?

13. What software tools or programming libraries are you using for modeling?

14. What else do you think is important in the process of model creation?
(Model validation)

15. How do you evaluate the produced models?

16. How do you compare different models?

17. Is the comprehensibility of the model important?

(Deployment)

18. How do you apply the created models to your business decision processes?

19. What technological compatibility requirements do you have for the produced model so that you can integrate it with your systems?

20. How do you deal with model maintenance and monitoring?
(Cloud services)

21. Are you using any kind of cloud service during your projects? (for example storage services or computational services etc.)

22. What criteria do you normally use to evaluate cloud services?

23. MLaaS are a specific type of SaaS. What are your expectations from such a cloud service?

24. Many of the MLaaS allow deploying predictive models as web services. What requirements do you have for such a web service?
Augustin: OK, we can start. Before I go to the questions, I just want to give a brief introduction to this MLaaS. Basically, in the past, it required a lot of knowledge with ML and developing systems to actually be able to create ML solutions. So, in order to simplify this, the cloud providers started offering these ML cloud services that allow to easily build and deploy ML solutions, so that everyone has access to the power of ML whether they want to use it for data mining, data analysis or something else, it’s up to them. The goal is to make it easy for professionals to use this ML services. However, because at the moment there is not standard for them, it might be a bit difficult for people to select the service. Like what should they look for when they select the service? Should they just go, for example, with Google because they already use google cloud services and are familiar with them? So, for this there is a need of metrics, some kind of metrics, based on which they can measure and compare these different cloud services. That’s actually the purpose of my thesis, to define some metrics, but in order to define these metrics, I would first need to talk to professionals who actually know what they want or have an idea what they could use this for so that the metrics would actually help them achieve those goals. I found a lot of literature saying that there is no point in creating metrics that nobody is going to use, so, that’s why it is first good to talk to people who actually use or want to use this in order to learn what is important for them and then proceed to the metric definition.

Augustin: OK. Now I think we can go to the questions. So, the first question. What is your professional role?

A1: My professional role is as a senior consultant, which is a great title since nobody knows at all what I am doing. I have been here for 10 years and basically what I do is anything and everything with regards to data. The databases, the data storage, the data analysis, the data transformation, usage of data is pretty much what I do. I’ve started out as DBA and realized that having this huge amount of data was interesting, and data in itself was more interesting than just toying with the database. So, that’s pretty much how I got started in the whole business.

Augustin: Ok. What experience do you have with data analysis, ML, data science and what kind of projects have you worked on?

A1: Those are fairly different questions. When it comes to ML, I have very little actual hands on experience. I’ve spent quite some time reading up on it and I’ve listened to a lot of the very smarter people than I am in the field. One of my favorite persons to listen to is Dayan Sarica [I don’t know if the spelling is correct]. He is from Croatia or Slovenia I think. He is pretty much mister ML when it comes to Azure ML in the world. So very good guy, very interesting to listen to. When it comes to data mining and data analysis, that is something I’ve been doing for some time. That’s more of a visualization through Power BI and stuff like that. I dipped my toes into statistics and me and math don’t mix very well, so, I usually get a headache when I have to do serious statistics and that’s when I call in my friend who is doing a PhD in mathematics and applied mathematics. She usually just laughs at me but that
is just part of the fun. So, when it comes to ML and data mining my actual experience is fairly limited, but as I said, I’ve looked at it, I’ve put together a few ideas for internal use here at the company and it something we are talking to our customers.

**Augustin:** Actually the next question is about this. Because this ML can be useful for many things such as generating predictions, detecting anomalies, doing some clustering and then studying the clusters. What would you like to use it for? What kind of problems do you think you could solve with ML?

**A1:** Very many different problems. Since we are a consultancy company, my customers range from retail, to production, to financial institutions, the whole nine yards. Meaning we have very different sets of problems and I think ML can be applied to each and every set of these problems, but ML also is subjective to the same issues that you have with data analysis in large, as in if your dataset is crap, then ML will give you crap, only automatic crap. So, that’s also part of the fun I think.

**Augustin:** Yes, as you say, the data is at the center and is very important. Now, how large is the volume of data you are dealing with?

**A1:** Oh, it’s anything from a couple of GB to hundreds of terabytes. It depends on which client I am working with at the moment. And to actually answer your question, now that I am reading it, I would probably use this for predictions. That’s the main goal, for instance in production and manufacturing. Anomaly detection, sure that could come in very handy with banks, for transactions and stuff like that, clustering also very interesting in retail. You’d be surprised what [Company name hidden] does.

**Augustin:** Oh, really?

**A1:** Yes. We found some interesting clusters with regards to [Company Name Hidden] in retail shopping. That was interesting.

**Augustin:** I didn’t know that [Company name hidden] do data analysis.

**A1:** Neither did we.

**Augustin:** What types of data do you have? Is it just database records, or you also have like images, audio video?

**A1:** 90% is basic relational data. Big data... Let’s put it this way, people are still struggling with the concept of Big data. You perhaps know the term data lake?

**Augustin:** Data lake? No.

**A1:** It’s pretty much a storage space where you just poor in each and every kind of data you can ever find in a data lake. In essence, it is a marsh, it’s more like a swamp. It’s a holy mess. So, it’s very hard to get insights out of this Big data and especially since Hadoop is pretty much, depending on who you are talking to, it’s a bit of a failed project, so to speak. It’s very difficult to actually get insights from Hadoop, so in 5 years’ time, I don’t think there will be a Hadoop. The idea is going to be with Big Data but Hadoop is probably gonna be switched out for something easier to work with.
Augustin: It’s interesting about Hadoop. Didn’t think it’s not so useful anymore.

A1: Well, everybody is talking about Hadoop. But when you start to actually dig through how many people have managed to do something with Hadoop that is not a university with a whole staff of mathematicians and statisticians. Normal people, generally won’t be able to get something out of Hadoop, since Hadoop requires you to know very well what you are looking for. It’s not very useful just for fact-finding, fishing expeditions.

Augustin: I see. How do you store and later retrieve the data you do analysis on?

A1: Mostly on stone tables. No. (it was a joke). Most of my clients are still on premises, meaning just spinning disks and SSDs.

Augustin: So they have their own storage?

A1: Yes. More of my clients are also looking at the MLaaS, or even better analysis tools as a service such as Data Lake in Azure, or SQL Server Analysis services in Azure, meaning they store more and more analytical data load in Azure.

Augustin: Meaning in the cloud?

A1: Yes. I work 98% with Azure. Azure is the cloud. So it’s the same thing for me.

Augustin: So the data is stored there and there you do the analysis also there with the tools they provide?

A1: Most often. If they store their data in the cloud, they are gonna use the cloud offerings for analysis. If they store it on premises, then they probably not gonna use the cloud.

Augustin: I see. Ok and now regarding the data understanding and preparation. What tools, techniques, including statistics do you use to analyze and understand the data?

A1: This is where my knowledge is not enough. I am not qualified to do the statistical analysis. If I need the specific information I am going to use statistician or a math mathematician. We have none at this company at this moment, unfortunately. So, if I need to do that kind of analysis I will team up with some specialists.

Augustin: But do you use some kind of software tools that would just give some kind of brief details like statistics?

A1: What I would use and this comes to your next question where it comes to missing values and duplicate values, I use the Master Data management tools in SQL server since I work 99% with SQL server. There we have data quality tools and data management tools. It’s not a statistical analysis in that sense but it gives me tools to ensure that my data quality is up to something.

Augustin: Because as you mentioned earlier, the quality of data is very important, if you put crap into the ML it will produce crap. Ok, then I think we have answered the question about the quality issues. Now, do you have labeled data? Basically, labeled data is necessary for predictions. Just to give an example for labeled data. For example, you have images and you also have labels for what is represented on those images, for example, a tree, a car or
something else. So, the label is very important for predictive analysis. Do you deal with labeled data or most of the data is unlabeled?

A1: Most of the data is unlabeled. But I have a few clients that do a lot of image processing and image processing without labeled data is very, very difficult. So, they are in the automotive industry. They have cameras both looking at the driver and looking out the window, then trying to do interesting stuff. But there we have a whole lot of labeled data, yes.

Augustin: Now, the way ML works, you take the data, you feed into a ML algorithm and it gives you a model. This model might be a predictive model that offers predictions, it might be some clusters that you have to look at and analyze. Now, would you like to be able to create such models through a user interface or maybe through a programming API or library, or it doesn’t matter? What would be the best for you do you think at the moment?

A1: Once again it depends entirely on the dataset, the client, and the requested outcome. Since graphical user interfaces are exceptionally well suited for prototyping. Just in talking to a customer, the client and discussing and at the same time creating a model. That's fine. But when it comes to do something that it needs to be run quickly, efficiently, and continuously, I would say “code all the way”. Definitely “code all the way”. I think you’d be hard pressed to find some professional that would say that they prefer the UI. Since you have no control over what exactly the code looks like.

Augustin: I understand. Ok. You can produce actually several models using the same data. Would you need some way to be able to compare these models and evaluate them how good they are or which one is best, for example, the accuracy of the predictive models? Would that be important for you?

A1: I’d say in a general sense yes. Very much so. And since we have a few models available to use, we would be probably fairly stupid not to evaluate them against each other. Then again, different kind of datasets are going to get different kind of results for the different models. But oh, yeah, definitely comparability is important.

Augustin: Many of the produced models, ML Models, are like black-boxes. They have an input and output; you know what to give in and they produce the results. But it’s not really possible to understand it’s inner structure or why does it give this result. Would it be important to have models that actually allow you to see what are the rules that produce the results? My question is clear?

A1: It is. But my answer is not so clear. My answer is once again: it depends on the dataset. Some things I’d be happy to leave to black-box, it doesn’t matter just as long as it produces results. Some things I would not be happy to leave to black-box. My friend, the mathematician she is working on a PhD with functional MRI. MRI is scanning things. They are creating a mathematical model for diabetes pretty much. They cannot accept a closed black-box model since they need to know why things are happening in a specific order. And applying this to my clients, some of my clients would be happy to just something in, something out, everybody is happy. But depending on what are they doing, they probably need to see what’s inside the box as well.
Augustin: Ok, I understand.

A1: There are no clear-cut answers.

Augustin: Once you create this model you probably want to apply it for business, for what’s needed.

How do you like to apply the created models to the business process? Basically, would you like them to be integrated into some system or just, for example, a decision tree that you can look at and produce the answer? How would you like to use them later?

A1: That was an interesting one.

Augustin: I hope it was clear.

A1: Yeah. I am just trying to grasp the ramifications. Once again, it depends on the dataset. I just this morning read that ML algorithms are between 7 and 13 % better at detecting cardiac issues than cardiologists. And in that case, it’s been integrated not so much as a decision tree but as a support system. Whenever cardiologist gets a set of tests, the system might recommend just at these 2 and check them out as well. So, in that case, it was integrated into a support system. Most likely I would say into a system rather than in a decision tree.

Augustin: Or rather than doing it manually, checking the model parameters and taking the decision.

A1: Then again, if are doing prototyping, just validating the whole model at all you might be fine doing manually since it might take too much time putting it into the system.

Augustin: It’s interesting when sometimes the ML models actually give you some rules based on which they work. So, this could be useful for example, for recommendation systems or something like that. So, to check what are actually the rules based on which we do the recommendation. In case this model would be integrated to a system, what technological compatibility requirements would you have so that you can actually integrate it to a system? Do you prefer a specific technology or?

A1: No. I have clients that still use basic, I have clients that still use Pascal, I have clients that still use Assembler, so, anything goes. What I like with the cloud offerings is the huge array of integrated APIs. I can use REST, I can use JavaScript, I can use pretty much anything I want to just connect things and it works. So, that’s one of the key things of the analytical software in the cloud.

Augustin: Yes. About the cloud services, are you using specific cloud services to deal with the data, not just for storage but maybe to do some kind of computations when you process the data?

A1: Personally I work with 99% with Azure, Microsoft’s cloud. And in there we have the Stream Analytics which is basically just handling huge amounts of streaming data. We have the ML learning toolkit; we have the Data Lake and HDInsight which is basically Hadoop running on Windows. So, we have each and every kind of the tool that we need inside the Azure cloud. And this kind of answers your original questions, what kind of metrics do you have. And my opinion is that your question is maybe not the right question to ask. It doesn’t matter what metrics you find on the ML part, it’s more of do you need other parts of this cloud. Since if I
have my data in Azure storage, if I am using Azure Data Lake, I will not be going to Google to
do the ML since I already have the ML part in Azure and if I have everything on Google Drive
I will probably not be using Azure ML to crunch the data. So, it's the same discussion we are
usually having when it comes to technology, it doesn’t matter what it says on the box, as
long as it’s integrated with other things that you need. So, what kind of business need to do
you have and how can this specific ML component help you with that.

Augustin: Yes, I think it’s clear. I understand that they sometimes make these services attached to
other services they offer and it’s hard to just to do something separately in a different cloud.

A1: Especially since the pricing is often in-bound and out transfers and considering the amount
of data you need for ML it’s gonna be very pricey.

Augustin: Yes. This moves to the next question. What criteria do you normally use to evaluate cloud
services?

A1: Integration and what my customers already have in place. Some of my clients turn out to be
fairly entrenched into Amazon. I say good choice and eject. I leave this to someone else who
actually knows Amazon. When it comes to ‘We don’t have any cloud stuff’. I’m here, let’s go
Azure. So, personally, I prefer Azure but when it comes to actual performance wise and
toolsets, Amazon is right up there. I haven’t looked at Google very much, but Amazon and
Azure are pretty much on par.

Augustin: Ok, I understand. So, it’s more like you are using a cloud service that you already have
knowledge about, how it works and are familiar with. This MLaaS it’s a form of SaaS because
most of the time it’s offered either as an API or actual interface that you can use. What
expectations do you have from such a cloud service? Like what would you like it to offer? It’s
a very general question.

A1: It is a very general question. I would like to have a simple, fairly low starting threshold, it
should be easy for me to start working with which incidentally Azure ML is. There I want it to
be integrated so I can add data sources, very different data sources, disparate data sources.
Then I’m happy it actually produces some results.

Augustin: That’s clear, I think. Now, many of the MLaaS allow to create a web service directly on their
infrastructure that could offer predictions. You just call their REST API and get the
predictions. What requirements would you have for such a web service?

A1: I can see that would be very useful for prototyping. Then again If you are going to use this
ML algorithm that created, you probably don’t need that website to do that. You probably
take it and integrate it into the process, the system or whatever. But as a general, just
checking the predictions it’s great. So, I don’t have very much requirements on them.

Augustin: You mentioned an important point, taking the model as a software entity and integrate it,
for example, into your own systems. From what I know some of this cloud services, they
allow to just create this web service and you can call it through a REST API but it’s not
possible to download it.
A web service would be enough in most cases and in specific cases it would probably be
good to be able to extract the whole thing, so to speak, and integrate it into a system. But
most of the time, once again if we are in the cloud, it doesn’t matter if it’s a part of the ML
toolkit or if it’s a part of the data lake or whatever. Everything is interconnected and
everything is using rest APIs so it’s probably ok.

Augustin: We are actually done with the questions. I just wanted to maybe do a quick demo, an
example.

A1: Sure. This is always interesting. I’ve seen one before.

[3-4 min wait to set up the equipment]

Augustin: A quick demo with Azure. Want to show you an experiment where we use some car data.
We have different attributes for a car and its price. We would like to create a predictive
model that would be able to predict the price of a new card based on those attributes.

[3-4 min wait because of some difficulties to connect to the web site]

Augustin: I hope their services are not down now.

A1: This is one of the interesting things with Cloud Services. They work just fine until they don’t.
And then you realize just how you just put yourself into a corner when you use cloud
services.

Augustin: Because you rely completely on them

A1: Yes. And I teach a lot of Azure course. And twice now when I taught the Azure
infrastructure, Azure has actually gone down. The whole thing has gone down and in that
case, it was Outlook.com that fell over. Meaning that none of my students were able to login
and it doesn’t matter if the cloud services working just fine. If you can’t login, you can’t use
it, so it’s kind of an issue that you usually don’t have on premises.

Augustin: So it’s a tradeoff?

A1: Yes, it is.

[Demo of a predictive experiment in Azure ML]
7.4 ERI - Interview with B1 (2017-04-25)

Augustin: Before we proceed to the questions, just to give a brief introduction to this MLaaS. Basically, ML is very useful because it allows to get some insights into the data that is hard to get using traditional data analysis techniques. However, it’s not so easy to use ML. It requires a lot of theoretical knowledge and technical knowledge, so creating and deploying ML solutions might be a bit hard. But in recent years some of the cloud providers started offering MLaaS which should make it easy to build ML solutions and deploy them. So, they basically made ML accessible to a wider audience. But now I think we can go to the questions. First, we have some introduction questions. So, what is your professional role?

B1: I work as Business Intelligence Consultant. My main focus is in data visualization and this ML area where we are just starting I would say. I’ve started a course one and a half year ago. It takes more time than you want to. As you say it is a bit more complicated, there is a lot of theory to learn. So, we are working on it and we are starting our first project in 1 hour. So right after this meeting.

Augustin: Ok, that’s great. Basically, you provided a bit of answer for the next question, which is: what experience do you have with ML, data mining?

B1: Yeah, exactly. We come from data analysis part. I guess we have been working with that in like 20 years in this company. Now we are getting some small data mining problems we have solved for some customers, but not very much. And now we are getting into ML and data mining. They go hand in hand I would say. I have a personal project as well where I work with ML. I have been working for this project like year, where I choose Swedish hockey league and I use ML on that to predict the next winner in the next game.

Augustin: That’s cool.

B1: Yes, it is.

Augustin: Actually it’s also related a bit to the next question. This ML can be used for several things such as predictions, clustering, anomaly detection and others. What do you use it for or what would you like to use it for?

B1: For us, as a company, I would say predictions and clustering. Clustering customers, customer segmentation, I guess it would be a pretty huge deal in the next coming years. And that knowledge we would use for predictions, what customer would like to buy what, so we can get segmentations.

Augustin: So, would you like to study those clusters? Maybe get some insights about them and then you also use them for predictions later?

B1: Yeah.

Augustin: Now, do you have any prior experience with MLaaS or ML cloud services (it’s the same thing)? Have you used such tools?

B1: Yes, Azure ML. We are partner with Microsoft. We at the campus use Microsoft Products primarily. We use R as well but that’s not a cloud service.
Augustin: A question. Do you prefer Azure ML because maybe you have other services from them so that it’s all together integrated, easier to use? What is the particular reason?

B1: Yes, I would say it’s like that. We have customers who use Azure Databases and stuff like that. Then of course you can use it with other services as well but we get everything in one place, we get the bills from the same service.

Augustin: So, everything is easier?

B1: Yes. It gets easier.

Augustin: OK, I understand. Now, at the center of ML there is data. Data is very important. Without data, ML, well, it doesn’t work. So, how large is the volume of data you are dealing with?

B1: I would say we don’t have like super much data. We don’t use real time data on our customers today. We work as a business intelligence (BI) consultancy. I work as a BI consultant in a BI team. Our customers use company data so it’s like hundred gigabytes maybe. I would say, it’s not small but it’s not huge. It’s controllable.

Augustin: Still, it’s not small.

B1: So, we don’t have much of noise I would say. We don’t collect everything. We get much clean data. So, we have data in nice format to start. We clean it before we get it. It’s good for some reasons, but you know with ML, maybe we have thrown away some data that could be useful for the future.

Augustin: That’s true.

B1: We’ll see.

Augustin: Next question. What types of data do you have? Is it only database data, or do you have also multimedia data such as images, audio, video and stuff like that.

B1: I would say yes, database. We don’t have any customers who use the others.

Augustin: Good. And how do you store the data and later retrieve it for analysis?

B1: Almost, every customer I would say they use SQL databases. We are Microsoft partner, we use theirs. We are starting to look into the cloud service like Data Lake and stuff like that. Now when the data volumes get bigger and bigger and as I said before, these days maybe we throw away a bit of the data, the noise, but when we have it in Data Lake it’s easier to store it.

Augustin: What kind of tools or techniques including statistics do you use to understand the data or maybe detect issues with the data?

B1: We use Microsoft toolkit. I use Excel a lot more I would say. Excel, and the Power BI, their self-service BI tool. It’s actually very good for these reasons.

Augustin: Is it also integrated in the cloud?

B1: Yes, exactly. And R for some small data analysis. I use personally.
Augustin: R scripts that would do something additionally with the data?

B1: Yes, I use it just for simple box plot and stuff like that to see things that should be there.

Augustin: I understand. How do you solve the data quality issues?

B1: We are a consultant firm. We work close to our customers. Mostly, it is their job to clean the data for us. It’s quite nice this idea. Data doesn’t look good and then it’s like they either pay us to fix it or they fix it themselves. And often, they fix it themselves so we get the data in a better quality.

Augustin: I see. So, it’s actually the customers who do the transforming of the data and provide it to you at the end?

B1: Yes, we say how we want the data in most cases. But most often I would say that because of the price and also for us when we are done with our project, our solution we give it to the customer. So if there is any problem it should be on the customer side and not our side so our product is good. But this is mostly Business Intelligence projects I would say.

Augustin: Now, for predictions you usually need to have labeled data to train the model and then you would be able to predict new instances. Do you have labeled data or is the data unlabeled?

B1: I would say that we have labels. In most cases the data is pretty clean. Yes, I would say we have labels for it.

Augustin: Good. Because I hear from most people that labeled data you don’t get very often. Most of the time there is just some messy lump of data and you have to make sense of it. I guess that’s good. And, what kind of transformations do you perform on the data like doing some filtering on the columns or doing normalizations, or deriving new feature from other features? Do you do stuff like that with the data?

B1: We start our first project in an hour so I can’t say really how we will do it. But of course, we will do some filtering and normalizations. We do in the BI work the ETL (extract-transform-load) so we transform the data how we want it. So, it’s some feature engineering. Of course, we will need it to get the result we want.

Augustin: Now that we are done with the data we can move to the model creation part. Basically, do you have any preference for any ML algorithms or what ML algorithms techniques do you use?

B1: I would say that we would never use the neural net, because of the...

Augustin: It’s like a black-box?

B1: Yes, exactly. And we as a consultant firm can’t give our customers a black box and say trust the result. It would be easier to use a decision tree. And there we can say follow the steps and that’s why you get the results you get. For my hockey predictions, I use Bayes. And it works neatly I would say. And normal regression is preferable as well. So, in the beginning of this we need to use the simpler algorithms to be able to speak with the customers in a better way.
Augustin: Until they build the trust in your ML solutions?

B1: Yes. I would say that it has to be that way. The neural net is nice to give something and then get a nice result but then you don’t know.

Augustin: Yes, that’s the disadvantage of it. It’s not really comprehensible for us most of the time.

B1: Exactly.

Augustin: Ok. How do you create ML models?

B1: We use Azure ML. I think it’s pretty neat to use it even when we do R stuff. It’s nice to do it in Azure ML first, it’s way faster just to drag and drop and use some properties there and then you get it how you want it and then you can build it in R much faster, at least for me.

Augustin: Ok, so you are using it like sort of prototyping? To create a quick prototype and then you would implement in code with R or something.

B1: Yeah. Exactly. We can use R in the SQL database. Like in our ETL where we do store procedures every day. We can use it there, so it’s much more deployable I would say for our area there and use Azure ML to get the data from on-premise out in the cloud and then down again. It’s pretty nice.

Augustin: How do you store the created models? The model itself is an entity which can be used basically. So, do you store it somehow, do you just, for example, if it is a decision tree, do you just print on a paper and then you have it?

B1: I can’t answer that. We haven’t done that yet. But when you do it in Azure you get a web service.

Augustin: It’s convenient. So, what are other software tools or maybe programming libraries are you using for modeling?

B1: Like libraries in R studio?

Augustin: Or maybe Python libraries, R libraries.

B1: I would say R libraries, DeepLer, ggplot2 to get some data visualization. Just the normal ones I would say.

Augustin: This question is a bit general, but nevertheless I will ask it. What else do you think is important in the process of model creation?

B1: As I say, we as a consultant firm, we have to think of a way the customer would think about it. We have to be speaking the same language. We have to keep it somewhere understandable. It’s more important than the accuracy in some way. To build trust, as you said before, for the solution. And then we can twitch the accuracy higher and higher and the understanding down, down. We have to start somewhere. I would prefer it the other way but...
Augustin: That’s how it works in business. I understand. Ok. Once you produce the model you need to assess its quality in a way. How do you evaluate the produced models? Do you have some kind of criteria based on which you assess the quality of the produced model?

B1: Yes. You have to get to the domain users. We can get some really nice accuracy, precision, recall numbers. But we have to talk to our customers, the domain users to get like: “do you think it’s good?”, “do you think it’s understandable?” or “are we not thinking right here?”. Cause we can create a lot of features, get some nice result, but maybe it’s not what the customer wants. It must be somewhere in the range where he expects. In the business we work, we are not in the medical treatment or stuff like that, we are on the edge of exploring, this is more like will this customer buy this.

Augustin: Ok, I understand. So, keep close a collaboration with the customer. And he can see how everything is going and he can give you feedback.

B1: Yes. Nice resume.

Augustin: This is actually very similar. Do you create several models and compare between each other?

B1: Yeah, I would compare them in Azure ML. It’s quite easy to compare them there. I use both the normal evaluation but also cross-validation to validate the models, to get a more wider angle.

Augustin: Yes, because in cross-validation they use all the data.

B1: Yes, exactly. In some cases, it is really good because you don’t have a lot of data. And cross-validation, it makes more sense.

Augustin: Next question. You’ve actually answered this one. Is the comprehensibility of the model important? It’s required that models should be understandable?

B1: Yes.

Augustin: Then we can move to deployment. Once the models are created they should be deployed in business use. How do you apply the created models to your business decision processes?

B1: I can only speak for the project we will begin working with and there we haven’t really decided yet. But we have an application where it will be used. And there we have two options. We either go with upgrading the SQL database to use R or we gonna use Azure ML as a web service from the application. I think it will be Azure ML cause it’s not a lot of..., they’re not gonna have a million of requests per day or something like that. It’s gonna be something like 50 or 100 per day. So, it’s not gonna be that expensive to use Azure ML. It’s one thing to use it for IoT and stuff, where ...

Augustin: Real-time data, continuous data...

B1: Yeah, exactly. There it can be really expensive I would say. Maybe go with R there. In our case, we can go with that at least in the beginning, until our customers understand how good this stuff is when it makes sense to pay a lot, cause you will get a lot from it.

Augustin: They finally will see the magic.
Augustin: Ok, let's move to the next question. What compatibility requirements do you have for the produced models so you can integrate it with your systems?

B1: I would say that it would be, as I mentioned before, either Azure ML as a service so we can get it easy or integrate in our daily. For most of customers we use stored procedures in SQL database once or twice per day. And it's really nice to use R in that. We get that aggregation of data in the morning. Then we have neat results. So it must be able to integrate it. We sit here, type, type code and oh now, we found something cool, we call our customers and “we have found something good”, “oh you work with my data now?”. The customers pay for our hourly work, so they want to see results when we work. We can’t just dig in and play around with the data too much.

Augustin: I understand. Some constraints. Once the model is deployed you would probably want to maintain and monitor how it is performing. How would you deal with model maintenance and monitoring?

B1: That’s hard to say. From customer to customer. Is it 50000 requests per day or is it 5 requests per day? It is a big difference. But of course, we have to monitor and maintain it. I can’t say today how we will do this part. We have to think about that.

Augustin: Ok. Well moving on to the cloud services part. Are you using any kind of cloud services during your projects, for example, storage services or maybe computational services?

B1: We use all of Azure parts. It’s pretty neat just to drag and drop. It’s quite easy these days.

Augustin: Yeah, they made it easy. From what I understood, you are using Azure a lot. I’m not sure if you considered at some point using other cloud services. But my questions is: What criteria do you normally use to evaluate cloud service?

B1: That part we have to think about. We have to look into Tensor Flow to see how we could use that. But I would say we would mostly stick to Microsoft and Azure because of the partnership we have with them. Some pretty good prices on the services we use there. But if get a business problem we can’t solve with Microsoft tools, we would use something else.

Augustin: Yes, that’s reasonable. Two more questions on the other side of the paper. This MLaaS is a specific type of SaaS. You are doing through a web browser, you have a nice interface, maybe you also get an API you can use. So what are your expectations from such a cloud service? What would you like it to offer?

B1: You summarized it pretty good I would say. It’s a web service. It has to be fast. I would say we can’t sit here and just wait for the magic. Where is the magic? I’m gonna get a coffee. Where is the magic again? You have to get pretty fast results. And I think that the cloud services work really good today on that part. Good enough at least.

Augustin: Many of these MLaaS allow you to deploy predictive models as web services. What requirements to do you have for such a web service?
I don’t understand the difference between this one and the one before. Expectations or requirements?

Oh, maybe I should be more clear. For the previous one, it was about, for example, the whole tool. What are your expectations from the whole tool you are using? But this one, it’s only for the web service that returns your predictions. Maybe you have some performance requirements.

OK. I would say my last answer was more for this one. Expectations from such a cloud service … I would say easy to work with it. I said I use it for prototyping. It is really neat to just drag and drop and it works well. I’ve seen other web services somewhere and it’s more like coding or more self-service. Maybe you can do more with them, but that’s not what I want. I want it to be easy. I try my model and it’s just a prototype, so I don’t want mess with a lot of code in that stage. The first result, and then you tune, tune, tune and maybe the code is better but in the beginning, I would say easy to work with.

Ok. That’s clear now. Basically, we’ve reached the end of the questions. Through this discussion, I try to see how are you doing ML now or what would you like to with it and what is important. Based on this I want to develop some kind of metrics, that people could use to compare different cloud services. Because at the moment, I heard many people saying that they would go with what they are already using. All these cloud providers they also offer this MLaaS besides their other services so that they clients don’t have to go to another cloud provider. This was actually the gray area where I thought it would be interesting to do some research. Look at some metrics or create some metrics that people could use to maybe compare. Because now I think there are only a couple of metrics like the price per request, price per batch request. So, you are fairly limited on how you can compare the different cloud services.

It’s quite hard I would guess. As you said in the begging, we stick to what we are already using it. We are quite bad at evaluating other options. Amazon is not really big in Sweden yet. Microsoft has a quite big market here in Sweden.

Well. I think that’s all.
7.5 SUM - Interview with C1, C2 and C3 (2017-05-12)

Augustin: So, before I start with the questions I want to do a brief introduction to what I am doing for my thesis. Developing ML solutions required a lot of theoretical knowledge and technical knowledge so it’s not that easy to actually create a good ML solution and unfortunately not everyone has competence in this so people struggle to do this. However, in recent years many of the cloud providers started offering ML as a service. Basically, services through which people can easily create ML solutions. The thing is at the moment there are several cloud providers out there that do this and it’s very likely that in the future more will appear. So the question for the users is “How should I choose the right MLaaS for me?”. For this they need some kind of metrics so that they can measure various aspects of this MLaaS so that they can make an informed choice and choose what’s right for them. Basically, that’s the purpose of my thesis, to try to define or create some metrics that they can use to evaluate this MLaaS. But in order to define these metrics I want to talk to people who use ML or are interested in using it to see what matters to them, what is important so that the metrics are actually useful to someone. I have a set of questions; they are about ML in general. So, I guess we can start. The first question: what is your professional role?

C3: I work here as an embedded software developer. So I work close to the metal, the hardware, small sensors side were we actually gather the data.

C1: He is involved in all.

Augustin: Are you doing the job of a ML engineer?

C1: No, he is focused on that, but he is part of the design of the architecture of the system.

C3: Since we are a small company we are always part of everything. That’s my main focus in embedded software.

C2: In my case, my focus is the opposite. I focus on the high-level programming, all of the interfaces to the user, database and server stuff. My role is also about doing some programming in hardware itself.

Augustin: OK then. The next question is: What experience do you have with ML, data mining and stuff like this?

C3: Some small university courses I’ve taken in AI. I’ve also studied open courses and general interest in reading programming and how AI works

C2: Same here. I’ve took some university courses. It’s very on the surface but...

Augustin: But you understand like the business use?

C2: Yeah, of course.

Augustin: Good. Next question. ML can be used for predictions, clustering, anomaly detection and some other things. What do you use it for in your projects or what would you like to use it for?

C3: It’s the anomaly detection that would be the first focus.
C1: Yes. And then we want to move to predictions.

Augustin: Predictive maintenance?

C1: Yes. Those are the main uses. But the first step is anomaly detection.

Augustin: Ok. Next question: Do you have any prior experience with this MLaaS?

C2: We had a meeting with IBM. It’s just really brief.

C1: They are interested to work with us so we have been in contact with them. And yeah perhaps Microsoft as well but... I mean we are not evaluating it, we are just getting information about them. We have other options but those are the main.

Augustin: To discuss a bit about data, because data is at the center of ML. How large is the volume of data you are dealing with?

C3: Do you have any idea?

C1: Sometimes it’s sound. We take the raw data, that’s the worst case.

C3: Usually in installations we use it like ...

Augustin: So continuous data? A lot of data coming and the volume can get quite high?

C1: Yes and we are already compressing the data (FFTT). But we want at some point the sensor itself to take some basic decision to minimize the data transfer.

Augustin: Put some intelligence into the sensor?

C1: Yes. Just to decide what to send and when.

C3: It could be that we train the model in the cloud and we send down a simple version.

Augustin: That could be interesting. Ok, the next question: What types of data do you have? Is it only database records or do you also have multimedia data such as audio, images, video?

C3: The focus is on vibration data, that can be a sound. But it’s not like we analyze the sound.

C1: We can consider it audio.

Augustin: Probably, probably.

C3: But it’s not like you would play it.

Augustin: How do you store and later retrieve the data for analysis?

C2: We are using SQL database. We have the raw data in data blocks that are fixed amount of size. Then we join them in the part that we have in the server. We get the set amount of records and that’s how we store the data. We do some optimizations to get a more concise view. We join some session data and do like that.

Augustin: A question. Do you store the data on-premise or in the cloud?

C2: It’s on premise.
But we are actually moving to the cloud. It’s just that the decision needs to be taken carefully. Even though we talked about that, it’s flexible in general, the cloud service.

I understand. What tools, techniques, including statistics do you use to understand the data or maybe detect some issues with it?

That’s the anomaly detection I guess and that is not a final thing, so right not it’s manual.

We have some tools that check the data, how it looks. The anomaly detection, I should say it’s ongoing.

Ok. Sometimes there are issues with the data like missing data. Do you have such issues and how would you deal with them?

In general, we shouldn’t have those issues, but of course sometimes we will have it. Our sensors, they always log the data to the memory themselves and once they get that information they upload it. So, in general, there shouldn’t be any missing data. But of course, there can be a malfunctioning sensor. In a more broader sense, we will have that issue but it would be more from a malfunctioning sensor not just because the internet is flaky.

And the sensor itself doesn’t remove the data.

So it has quite a big memory. It can be without internet for a few weeks.

Also, the data format protects the data from loss. If something is missing treat the data accordingly.

But I mean if you look at it in a broader sense since we do battery driven, of course battery can die, then you have to change the battery. So, that can big a big gap. You cannot protect yourself fully.

Ok. The next question: Do you have labeled data? Basically, labeled data is needed for predictions, anomaly detection.

No, in general, we will not have labeled data. We will just have how the machine looks right now and the system needs to figure out.

But for example, can a person who is using the machine report that the machine is broken. And then you would check, ok so when is that period when it was broken and maybe you could label that period with...

That we can do. But in general we do not enforce our customer to do it. But that greatly helps the learning process.

I see. What kind of transformations do you perform on the data like filtering, normalizations, combining features?

For instance, it’s not exactly like this but some filtering, so instead of sending raw data we look at the frequencies and that filters a lot of noise because then you can look at the frequencies. It’s more stable data. That’s some kind of filtering.
C1: Normalization, that we don’t have but that’s part of the anomaly detection.

Augustin: Let’s move then to the model creation where you actually create this model that would give you predictions or anomaly detection. What ML algorithms or techniques do you use at the moment?

C3: In this anomaly detection, one algorithm candidate that we could start with is this kNN.

Augustin: And how do you create the ML models? You write your own code or do you use a library?

C3: No, no. We don’t write it our self. In this test phase I think it’s python that is used to create this and then there are other libraries. But when we move to the cloud, if we move to Azure then it needs to be Azure’s libraries. Or if IBM then IBM libraries.

Augustin: Ok. How do you store the created models?

C3: Since it is mainly test phase we don’t to store it at the moment.

Augustin: But it would be useful to somehow store that model so that you can deploy it?

C3: Yes. If we are gonna send it down there we need to have a format for it.

C1: But we are far from that.

Augustin: I see. You answered a bit from the next question. What software tools or programming libraries are you using for modeling?

C3: Right not its Python and later it would be whatever cloud service.

Augustin: What else do you think is important in the process of model creation?

C2: So, for example now that we are using the clustering techniques, getting the number of clusters, that is an important thing that we wish. How to create the model itself.

C1: And how to evaluate it. Interact with the users. It can be directed to what you want.

C2: So, for example, different machines have different states. One machine can have 5 states that are correct and anything that’s out of this can be wrong. And one machine can be simple and have 3 working states. And everything that is outside is a failure state. Getting that and the number of clusters directly, so if we put a sensor in a machine that we don’t know getting that is one the most important parts.

Augustin: I understand. The next section is about model validation. Seeing how good the actual model is. How do you evaluate the produced models?

C1: With customer feedback. We are talking about a real product, so we will run pilots on this. So more like this and not in a formal way.

Augustin: Do you compare different model between each other? Did you try to create different models?

C1: I don’t remember.
Augustin: Some of the models are like a black-box. You can’t understand what’s inside. You give it an input and you get something as output but you cannot understand the inner workings. Is this comprehensibility important for you? Would you prefer to have a “white-box” model so that you can see the rules?

C1: I think it’s simpler for us. I mean it depends on the application.

C3: I don’t see it as great… I mean it’s good if the model can explain why it says what it says but it’s not exactly a necessary feature.

Augustin: So as long as it works good it’s perfectly fine? Doesn’t matter if it’s a black-box?

C3: Yes.

C1: Cause understand it, it means for a company money to invest, and you expect some reliability. Some reliable model.

C2: Also, white-boxing and modeling is always like giving a little bias. So, the model might not be that smart. How do you say...

Augustin: Maybe not that accurate. You will be able to see the rules but it will not be that good at predicting or detecting the anomaly.

C3: Of course, with the “white-box” you can understand when it fails before it fails. So, you understand the rules.

C3: If we select a model and it would be hard to understand like these neural networks models, they are hard to understand but for me that would be fine.

C2: If the output is reliable it is fine.

Augustin: The next part is about deploying the model. How do you apply the created models to your business decision process?

C3: This would be a bit of guessing because we are developing now the system. We haven’t deployed yet anything like that.

C2: Once we have it, our business model will revolve around anomaly detection.

Augustin: As I understand, it will be integrated into a system?

C3: Yes.

Augustin: Ok. What technical requirements do you have for the produced model so that you could integrate it with your systems? Do you need a specific language?

C2: If we can insert it in the cloud there can be some ways like we can create our model or a REST interface that we can query. It would be perfectly fine for us.

Augustin: You are not doing this yet, but did you think about how would do model maintenance and monitoring?

C3: That’s way out of mind.
Augustin: I made this question and I asked some other company around, but apparently, everyone is just starting with this at the moment.

C1: It is more advanced in the academics that in the real life. In practice, I don’t think there are many companies.

C3: In a few years, there will be.

C1: Yeah, there are many companies who started with this. We have met several but they are going in the same direction. So, I guess in a couple of years we will totally different.

C1: We are having many resources now. So, we can have something this year.

Augustin: Now some questions about cloud services. Are you using any cloud services during your projects for example storage services or computational services?

C3: Not yet.

C1: We are just planning.

Augustin: Because you’ll be moving to the cloud services, do you have any criteria that you would use to evaluate these cloud services?

C3: Price.

C2: Price.

C1: Probably price. That’s the main...

C3: Feature wise it should be easy to use and develop. IBM and Microsoft, they have a lot of features, they have really nice cloud services with a lot of features that we need and probably a lot of features that we don’t care.

C2: We would like to port our code to the cloud. Doesn’t require so much to keep it running.

Augustin: So you don’t have to deal so much with maintenance.

C3: Scalability. If you want to start small you don’t need to pay upfront. You pay as you use.

C1: Everything goes to the price. Maybe in 1 year all cloud providers will be the same or roughly the same

Augustin: I’ve actually heard from many people saying that they prefer to use the same cloud that they have experience with. So even when someone offers the same things, they would still stick to what they are used to.

C1: Once we are integrated to that I don’t think we want to change.

C3: If we select Azure and then suddenly Bluemix dropped the prices, it could probably cost more to move it.

C1: Depend on the cost of course but you need to take a decision and that’s why it takes so long to take that decision.
Augustin: And I have two more questions on the other side. This MLaaS is like a type of SaaS. What expectations do you have from such a cloud service?

C3: My expectation is that it will work. It should be easy; it should be always there. When we have it right now with our server we have to deal with all the maintenance. We have other companies that are helping us with that, but anyways, it’s very direct. If the hardware breaks we need to buy a new hardware even if another guy helps us with that. In that sense, we expect not to have any hardware issues. We would never see that we have hardware behind the cloud. That’s one of my expectations.

C2: My expectations would be that we don’t need to run a virtual premise to be all the time doing nothing. You should be working like a stream processing. So, it only runs when we need it to do some calculations and that would of course lower cost which is one of the main points. We need anomaly detection and having some aggregation for the data that you could show monthly. But usually what ML offers in general.

Augustin: You mentioned one point that it should be available, “always there”. Because if the service is unavailable obviously, you can’t use so you always need it to be available.

C3: I mean from our sensor point of view it’s fine if something goes down, but the customer might not like it. The sensor itself don’t care if it’s down for a little bit because it will store it while it’s unavailable.

Augustin: That’s reasonable I think. The last question. Many of the MLaaS allow you to deploy predictive models as web services directly on their infrastructure. So, what requirements do you have for such a web service?

C2: We will need just a REST API so we can integrate it with ours.

Augustin: Do you expect to make a lot of concurrent requests to it. For example, to get the predictions. Do you expect that the amount of requests will be high because you mentioned that you are doing real-time data so the volume is quite high?

C3: The volume can be high, and in the end, you will maybe not update the model on every new data point that runs through the model.

Augustin: So it’s not like getting “there is an anomaly right now”. Not something like that. Like every week we get some kind of report?

C3: No, it needs to be real time. The model update could be done every day. But to detect anomaly need to be direct. Once the data comes in it should be checked against the model. So, that needs a response directly.

Augustin: Quite quick?

C1: Yes.

Augustin: Ok so we are at the end of the questions.