Customer Requirements Analysis in Machine Learning Projects

Djordje Rodriguez

2020 Laurea

Laurea University of Applied Sciences

Customer Requirements Analysis in Machine Learning Projects

Djordje Rodriguez
Business Information Technology
Bachelor's Thesis
May, 2020

Laurea University of Applied Sciences
Business Information Technology

Djordje Rodriguez

Abstract

Customer Requirements Analysis in Machine Learning Projects

2020 Pages 30

This Bachelor Thesis explores the various requirements of Machine Learning projects that utilize numerical data in their models. By interviewing various experts in the fields; varying in expertise, age, gender and ethnicity; we discover the key factors that often represent the biggest obstacles for the teams working with such technologies. The interviews were conducted between September and December of 2019 and studied until February of 2020. Using 2 research strategies: Interviews and Literature Reviews from news media articles and published academic work. The analysis was then executed through the recording of the interviews, their transcription and analysis conducted in tandem with the review of literature focused on Machine learning models and numerical data.

This thesis gives a general understanding into the work that is typically required these days to run these Machine Learning models. Uncovering that the problems more often originates from the lack of understanding from those wanting these technologies to be implemented in their systems that from the systems themselves. Highlighting the importance of understanding data and good practices when it comes to storing, structuring and modelling for ML models.

Keywords: Machine Learning, Customer Requirements, Data, Data Readiness, Project Management

Table of Contents

Terms and abbreviations

ML	Machine Learning
NLP	Natural Language Processing
ETL	Extract, Transform, Load
DS	Data Science
BI	Business Intelligence
EDA	Exploratory Data Analytics

1 Introduction

Machine learning (ML) is becoming more and more available in today's world, with companies showing increasing interest in implementing these technologies as part of their offering, whether to predict the likelihood of clients leaving, or process medical scans of patients to look for cancerous cells. We are seeing the use of ML technologies broadening and outperforming experts that have been working in their fields for decades. With the ever-growing amount of data being produced every day, it is now essential for us to utilize these technologies to understand the data and unlock its potential in supporting the decisions we make in our daily lives and businesses.

Avanade Oy is a large Finnish consulting company present in 24 countries which works towards bringing businesses into the digital world, through various types of projects and helping them in finding the best ways to use technology and achieve the best results. In this way, they seek to identify customer requirements in these ML-related projects and use the results to improve their offer and streamline their services.

The goal of this thesis is to explore the customer requirements in ML projects and identify the following:

- Contributing factors for successful ML projects;
- Common pitfalls in which customers find themselves, stifling progress; best practices when considering ML implementation.

With this, Avanade Oy will be able to consult their customers more efficiently, ensuring that all requirements are met before endeavouring on implementing ML tools into their systems and offering.

I hope to convey a coherent description of successful ML projects by digging into the interviews that I conducted and researching the factors that need to be covered for these projects to provide as much value as possible to the company using it.

1.1 Avanade Finland Oy

Avanade Oy is an IT Consultancy company with offices in Helsinki and many other cities across the globe. They specialize in technology and the Microsoft ecosystem; helping companies solve complex issues and projects with the use of new emerging technologies and the Microsoft long list of tools (Avanade, 2019). The work they do often means co-operating with big organizations that have large datasets that they wish to use and from which they wish to generate insights and predictions. The company's interest in this thesis lies in drawing up a list of

requirements for developing and implementing processes to ensure the success and efficiency of ML projects.

1.2 Thesis Background

Avanade Oy wants to compile all the information on how these projects are handled, the obstacles that are most frequently encountered, and the common practices used in these projects to ensure their success. Using the information gathered to further develop their processes and streamline their ML offering for their clients.

To this end, the thesis takes an exploratory approach relying on the interpretations and professional experiences of the interview participants. With the purpose of understanding the work that goes into ML technologies and their implementation. Avanade has asked to focus on ML projects involving numerical data.

1.2.1 Objectives

The main objective of this thesis is to extrapolate the pre-requisites for successful ML projects. Using interviews as the main method of data collection and gathering information on the participants' common practices to in handling and solving these problems and tasks. This requires digging into the participants' experiences and identify the key insights and factors making their projects successful.

1.2.2 Scope

The scope covers projects that have a focus on ML and those have ML as a main component of the system. As requested by Avanade Oy, it focuses mainly on projects involving numerical data. Each interview participant was informed about this beforehand. Participants could however use other examples to describe ideas or insights that they may have. This will allow us to generate more valuable insights on the success factors of these types of projects.

1.3 Methodology

This section covers the methodology used for collecting the data/information on the matter i.e. interviews (and how these were conducted), and literature reviews. Once the data has been gathered, each interview will be transcribed and analysed. All analysis will then be consolidated to identify most frequently occurring topics and their relevant importance.

1.3.1 Snowball Sampling

Snowball sampling is a method by which uses existing participants to recommend and recruit from within their networks. (Statistics How to. 2014)

It was used to find participants. First using my own and the partner companies' network and then engaging with the participants themselves for recommendations on who to invite to the interviews. Participants are also made aware that they have no obligation to recommend anyone.

1.3.2 Interviewing

Interviews were conducted in a semi-structured manner, thus giving the interviewer the ability to dive deeper into the content of the interviewee's answers. The length of the interviews was from 20 to 30 minutes and the full duration was recorded. Each interview was then transcribed and analysed. All analyses were then consolidated in order to identify the most frequently recurring topics and the degree of importance accorded to them.

2 What is Machine Learning?

Shalev-Schwartz and Ben-David (2014) define Machine Learning as "the automatic detection of meaningful patterns in data". They go on to say that it has become a common tool for anyone looking to extract information or insights from big data sets. Where we previously relied on our own ability to analyse and use large datasets for our predictions and business development, ML takes that and does it with more accuracy and efficiency that we have been able to.

Another way in which Shalev and David (2014) define machine learning is as follows:

"…the ability to turn experience into expertise or to detect meaningful patterns…"

With this, you the understand that machine learning is a system which learns from data to define what the important factors of that data set are and to then know how to predict the likelihood of new data accomplishing the same results, turning that experience (historical data) into expertise (the predictions).

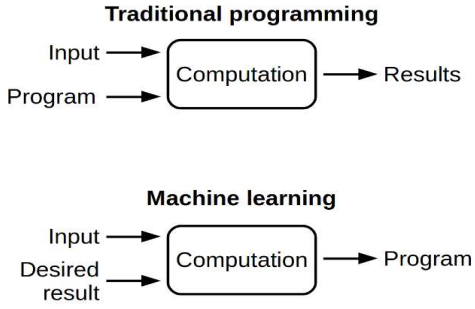

Figure 1 - Traditional Programming vs Machine Learning. Ajanki, A. 2018.

Software development has focused on implementing rules and giving data to generate the results. Machine learning differs from this in its creation/training phase by providing the results and data to generate the rules (Moroney, 2019). What that essentially means is that the ML model uses the data and the answer whose likelihood it is trying to predict, to generate the set of rules that provides optimal result. Once the model has been trained with these, it will then be able to use those rules to predict the probability of X result occurring from the data with which you will subsequently provide it.

The model therefore goes through the data which represents something that has already taken place. The model therefore knows what has made X possible historically and can use that to analyse similar cases. If you consider banks, the way customers use their accounts generally stays routine, making ML models easier to use. If the behaviour of the customers were to change however, the model would eventually get outdated depending on how significant the changes in behaviours are (Roman, V. 2018). Therefore, it is so important to keep training the model and to provide it with new data, which is then representative of the changes taking place in the data, thus allowing the model to adapt itself to those changes and to keep providing updated predictions.

2.1 Case Examples

There are many different applications of ML in systems all around us every day. Most of them go unnoticed since most people would not see them or think to check in their everyday routine. If you consider credit card fraud, search engines optimization, content suggestion in services like Netflix or YouTube, there are many ways in which ML is being used to help people and business get the best results in a variety of ways. Then there are the more obvious and visible ways that people are more aware of; these would include driverless cars, robotics and facial recognition on people's phones for example. These examples are all cases where the data used is either too vast or too varied for humans to process easily. ML in these cases will go through the data and recognize the meaningful patterns within the data that will then allow the system to operate with a certain degree of certainty based on the results given by the model. (Lawrence, 2017)

Pranav (2019) gives us a couple more recent examples where we are seeing ML being used to improve the results of these services. Most of the technologies implemented into smartphones nowadays will have ML running in the background, whether it is to improve the image quality of the photos the phone takes or to give a smoother 'Hey Google' experience. There are several models running in the background that allow these services to understand and know how to deliver their service in the best way possible.

Another interesting example given by Pranav (2019) is that of 'Dynamic Pricing in Travel'. If you have ever used a ride hailing app such as Uber or other, you will know that the price for

the same journey is not always identical and that often it is initially shown as a ranged price. For the app to calculate the best price to give to the customer, it will run all the data it has on the journey being proposed against the demand that the app is currently experiencing. It could be said that it auctions the price of the ride according to the number of people that are looking to travel (supply and demand). While the customers only see themselves going from point A to point B, the systems behind the app are running all the data they must find the optimal way to deliver their service.

ML is also entering the domain of security. Where guards would previously write down number plates of vehicles coming in and out, that can now be done using a camera and a model. Cutting out tasks that are heavily time-consuming allows employees to focus on different and perhaps more important tasks. (Pranav, 2019)

3 Interviews

To recap the details of the interviews:

- Time period: September and December of 2019.
- Location: ½ Online ½ Face to Face.
- Average length: 26:37 minutes.
- All interviews were recorded and transcribed.
- Interviewees came from various countries and worked in companies varying in size and field specialization.

The semi-structured nature of the interview really allowed the participants to answer as they saw fit and gave me the ability to dig deeper when I felt there was a need to. This was particularly true for one interview in which the complexity of the projects described demanded further explanation and clarification.

For the sake of anonymity, each participant was informed that their identify would remain anonymous in order to protect them from anyone unwanted attention which their contribution to this work may bring about. To that end, the recordings, transcriptions and notes shall be destroyed once the thesis has been completed and officially handed in to the university.

3.1 Interview Structure and Questions

Six questions were planned for the interviews. These however, were more guidelines for discussion topics than strict wording and steps for the interviews. In all cases, there were several follow-up questions to the answers given, generating additional insights and/or clarification on the answer where needed.

Before the interviews, I studied many books and listened to podcasts to identify where the focus of questions should be. I established that the questions should cover Data Quality and Quantity, Skills and level of understanding of ML and the systems running these technologies, as well as ML capabilities and limits. These were sometimes covered under different names, for example, data quality being referred to as data readiness. These served as a starting point which proved quite efficient when having these interviews with the participants.

3.2 Interviewee 1

This interview began by discussing clients of the participants who often have large quantities of data available in some form, but with little knowledge of what it contained and how it could be used. When working on these types of heavy data related projects this made the goal for the client much more open-ended as they simply did not know what the possibilities were. First conducting a project with the main objective of discovery through which the team can identify what was possible to do with the given data and what was not.

This required the team to spend a great amount of time on cleaning and wrangling the data, as this data has simply been collected and not necessarily structured in most cases. Without this step, they would not be able to run any business intelligence (BI) or visualization tools on the data. The data cleaning also allowed the interviewee to gain enough familiarity with the data to inform the client about what was being represented if this had not already been established. From that point, it should be possible to identify the potential ways and purposes for which the data could be used. Further helping in identifying whether to use ML tools or not. While such work can be exciting, it did not make the project efficient. The chances of finding a way to use the data that will be invaluable to the customers without their input and understanding of the business that it represents are slim, therefore the goal of such projects is simply to inform them on what exists. This has value, but less so than providing a solution based on that data.

Data Documentation and organization was the following main topic mentioned. On this occasion, the interviewee stated that when working with larger organizations, it would often be an issue that the owners who have access to the data, frequently lacked the skills to exploit it. That complicated knowing the contents of these datasets and gaining access to them not only for the business but also for the interviewee who would come into the project and not

have the information of how to access the data or what it contained. The participant emphasized that simple backend skills would already remedy these types of issues. In this case, data owners would be able to understand and give access to the data much more easily, saving countless hours of work required to remedy the issue. Documentation on the database would also be immensely valuable, as it would save the numerous hours of work that are put into familiarising oneself with the data. It is important to have some examples and clear, easily accessible documentation. Examples in python or other coding languages being used could help understand the structure and content of the dataset. Taking APIs (Application Programming Interface) as an example, data owners would often provide poorly formatted PowerPoint slides in which the person seeking the data would have to dig through each slide to access the data. This could be made easier by having the data readily available on a webpage, or by clearly indicating who to talk to get access to the data. There is always a data exploration required in any of these projects but having this documentation would facilitate that phase tremendously and allow the team to proceed with the project itself.

Even though these databases have been collecting data and been updated over time, there is an issue of the use or non-use of this. With data gaining so much importance now, it becomes a real problem for clients where they must figure out accessibility and availability. According to the participant, the reason why this is so important now compared to before, was simply due to analytics. Analytics tools have developed a lot recently. ML for example, has become much cheaper and simpler to make. Employees in these larger organizations have heard of it but do not know how to go forward with it, making them cautious when having to work with consultants or experts from outside their own company.

Data quality is also important as poor quality data makes it extremely difficult to transform data into a readable state. While time consuming, this is nonetheless essential, as generating analytics on poor quality data can have no positive outcome. As previously mentioned, this is a major part of the participant's work. Thus, they are not only responsible for visualization and analytics, but also for cleaning up the data for their clients and providing them with the added value of being able to use the data easily in any projects that they might take on later.

Awkwardly enough, when asked about data trustworthiness, the participant said clearly that they must put a blind trust in the data most times. For example, if the client were to have data generated by thousands of sensors, unless these outputs a signalled a clear 'ERROR' then it is hard to find the anomalous values inside the data. Having anomalous data however is an aspect that occurs even in high quality data. However, in most of the cases that they have worked on, such data usually came from a start-up which focused on that one type of sensor. Depending on the project, the interviewee also said that they would re-measure the data from the sensors and that if this did not provide the same values, there was plainly an issue there.

The interview concluded with the importance of mixing these various technologies together as they were most useful and efficient when put together. ML is great on its own but becomes much more powerful and efficient when combined with cloud technologies for example.

3.3 Interviewee 2

The first topic brought up be the interviewee related to skills and knowledge, in particular that having statistical and mathematical knowledge when working with these data science related projects is one of the basic requirements to develop a ML model.

At the start of these projects, customers tend to ask for ML or AI (Artificial Intelligence) almost as a standalone product. This does not only fail to take account of the differences between ML and AI in the way that they are used and what they describe, but also shows for a lack of understanding in what these systems are and do. The term AI is used to describe a multitude of technologies under which you can find ML, Natural Language Processing (NLP), and machine vision for example.

What the interviewee noticed in their work, was that clients would often request proof of concepts but do so without having a clear business statement of what the goal for them was. As mentioned by the previous participant, the lack of objectives provides them with freedom to play with the data but does not make the process more efficient and doesn't help in achieving a more beneficial outcome for the client. What is more, these proofs of concepts are often left untouched after completion. This is problematic because these models are no longer as efficient or beneficial when left untrained for longer periods of time, as the behaviour represented in the data often changes over the course of that time. As a result, even if the proof of concept turned out to be very valuable initially, it gradually lost that as the customer needed time to get to it. This is often due to the lack of understanding of what the actual benefits of using such technologies are.

In their experience, the best result did not come from generating actionable outputs with these ML models, but rather generating outputs supporting and easing the decision-making in the company. These systems are much more in the bigger picture and not simply for singular purposes only valid at a specific point in time.

The participant stated that they would often ask their clients what purpose they seek in wanting to use ML, establishing the goal through their answer, and seeing in the process the clients' understanding of ML. For example, should the client want to improve their customer retention, that means that as a consultant working on the project, they have to understand why that retention is important to the client and what kind of benefits they are attempting to achieve by holding on to that client. This will not only help them achieve a better result for the client, but also understand what measurements and features to use in the datasets.

When clients are asked about data and the requirements of the datasets, they sometimes reply that they have issues gaining access to the necessary data. This is more often the case with bigger clients as they tend to have multiple teams or departments who each have their own data silos which can only be accessed by them and share no commonalities with other teams' data and silos. This makes it difficult to have enough data, and even harder to convince those data owners to share the needed data as they may only be willing to provide access to a portion of it, thus stifling the exploration process by which they could determine the data's potential.

Data quality is essential in using these datasets for predictions. This involves various factors ranging from data quantity to trustworthiness and annotation for example. More data will result in more examples for the model to learn from, while the trustworthiness will then ensure that results correspond to the reality of the information it represents in real life. The importance of documenting and annotating the data is in allowing anyone working with that data to understand it without needing to investigate of the meaning of each feature. Documentation allows us to know where the data comes from and how it relates to the overall project and in knowing if the dataset is representative of the wider set of data as well.

ML modes output comes as a percentage probability that tells you the likelihood of the desired outcome happening. Having a goal for the model is crucial at this stage, which is where the proof of concept generally shows and teaches the clients about the possibilities and limitations of these models. Another important factor is that once established, the model needs to be kept up to date by continuing to teach it (training it continuously). With regard to the participants, a common and efficient strategy saw multiple models being built and competing against each other. The most efficient of these would be deployed in the client's services. When using this strategy, it is important to use the same data for all the models. A good example of this would be credit card fraud. When considering that people purchase in a certain manner at present and how that is subject to change, what might have been considered as fraud previously may have been retired by those committing the fraud and they would have come up with new ways to be fraudulent. Or more simply, a customer has a specific pattern of purchases then for no visible reason in the data, that changes.

Making a ML model is not a project with a defined endpoint, more a project which, once up and running, will need monitoring and tweaking to stay relevant and to keep producing good results. These models, if left alone, will quickly become outdated if the data being used is not up to date. This is highly representative of the fact that the client then needs to have a data strategy in place, which not only ensures that the data itself is being stored and structured efficiently, but also that the models processing the datasets are still performing as they should be and are replaced when outperformed by newer models being developed.

When considering which features to use in the data, it is important to ensure that the individuals represented in the data are unidentifiable. The participant says that often enough, this type of data is not essential for the models anyways. A feature like a phone number or the social security number would not help the model in determining or predicting behaviour as these don't impact the goal of the model. There are features that could be useful but are not necessarily essential, such as profession, height or weight. While these factors could be helpful, they often would not be significant enough to be considered. Knowing how to be careful in combining data sets is also very important as doing so can allow for individuals to become identifiable.

While ML projects may seem easy to understand from a business standpoint, once technical aspects come up, these experts don't always understand why such large amounts of data are required nor why they should give access to more data, which could have much more valuable features. These clients' representatives were not informed about the data, leading to complications when needing information on how the data is collected for example. Knowing where the data comes from, how it is being stored, and what is in the data itself are crucial to these projects. It gives the client a clearer picture of what they have in their data, which facilitates the early stages of these projects tremendously.

The last topic discussed was 'Data Silos'. As each team or departments in companies, collect and process data in their own systems, aka 'Silos'. This is an issue as it increases the workload in the early stages of these projects. Formulating a data strategy resolves the issue by setting the purpose and objectives that the data will serve in the future. Possession of it will also help in collecting data early, which could prove invaluable for any BI or ML work.

3.4 Interviewee 3

The interview began by discussing the main concerns to keep in mind when talking about ML technologies, and in particular that it is much more important to talk about the big picture than the details. ML alone is only valuable when it is integrated into larger projects or systems at which point it then becomes extremely valuable and interesting to employ such technologies. In addition, within ML itself, there are different levels that need to be considered. These include pre-processing of data, going through all the maths for building the model itself and using different types of learning for training the models.

First and foremost, the user needs to know where the data comes from. Understanding how that data is collected and what it represents is essential to move forward efficiently in these projects. When cleaning the data, it is necessary to ensure that the handler goes through the process of familiarizing themselves with the structure and ensuring that only useful data is left, as not all features are always useful or necessary. For example, when using ML in hiring staff, as described in 'Weapons of Math-Destruction' written by Cathy O'Neil (2016), there

are a lot of features which need to be taken into consideration by the models. These include names, gender, origins for example which shouldn't be relevant to the hiring process, or at least shouldn't be important enough to remain in the training data given to the models. While the sample needs to be representative, it is also necessary to make sure that the user is not introducing bias in the data. There are many issues that can be solved or given insight into with ML. In the field of healthcare, we are seeing ML models being able to identify certain types of cancer with more accuracy than the experts themselves.

The interviewee emphasizes the importance of not introducing bias into the data, acknowledging the fact that it is easier said than done. Ensuring that you have a team with diverse thinking that is working on the data so that they can support each other in spotting the potential sub-conscious bias that could be present in the data, is one way of ensuring this. In considering the open data angle further, access to the data happens through some sort of API or other. This means that when developing a model on this data, the developer has to essentially have full faith that the data is trustworthy and representative. He/she needs to understand in the process that ML is a part of the stack, part of the solution and not the whole solution.

Gaining that basic understand of what ML technologies do and are capable of is essential in developing robust models that can be trusted for their results and the data it uses.

3.5 Interviewee 4

The projects on which this interviewee works are specific to a problem type that they know their customers have. When they engage with these customers, they already know with a high degree of certainty that the problem they are called on to solve as a company, is indeed real. Once they have discussed and established its existence, they proceed with the project.

The company often begins with a proof of concept in order for the client to visualize the type of benefits that the project will bring them and to demonstrate how it will solve the issue at hand. ML is but a part of the solution and not the solution itself in this case. The company sells a solution aimed at a specific problem. In short, this means that the client is not required to have understanding and knowledge on ML. This facilitates the process tremendously for the interviewee's company as they then can discuss the more technical requirements directly with the CTOs (Chief Technology Officer) and engineers of those companies once the project begins. These customers also tend to be more technological inclined and are often familiar with good practices in data collection and storage. The solution itself requires that these parameters are covered before implementing the solution into their systems. When these are not covered, the goals are not easily and efficiently determined, thus making it difficult to know that the work will be valuable.

This facilitates communication, as it is not necessary to explain the technical aspects of how their proposed solution would work to the business leads who approve the projects. If asked, they are happy explain, but these customers often are just getting familiar with these technologies and might not be able to grasp some of the concepts used in the solution. It is nevertheless important to the participants' team to manage expectations. They do this by preparing a proof of concept as previously mentioned. This helps the customer to visualize and understand the potential of the solution offered. It also allows the customer to test the proof of concept and see if it can be used in the ways that they need it to. The proof of concept is needed because of the nature of these types of projects, which applies to ML in general.

We also spoke about the importance of data quality, which was described as an aspect which made the whole process much easier in later stages of the project development. When the customers have not done any data audits or such, it means that there first needs to be wrangling work done with the data before being able to tell what additional work there needs to be for the project itself. I.e. If we were to think about this with a clean, ready-to-go dataset, it would be necessary to dig through the data straightaway, to see what is and is not there, thus allowing implementation of the changes required for the successful delivery of the project. This is not to say that a clean data set will always make for a smooth ride. In fact, there are project types where you would still struggle to find the feature that would bring it all together to provide that key insight for pushing the business forward.

The interviewee continued to discuss the various requirements of data readiness, referencing the paper "Data Readiness Levels" written by Neil D. Lawrence. According to which, it is important to have all the data available on a virtual private cloud, making it harder for external IPs to access their systems. This is particularly important when working with ML technologies, which consume vast amounts of data very rapidly, referring, in this instance, to public clouds provided by various companies. According to the participant, all the servers need to write the data back to the same private cloud. Paying close attention to the connections going in and out of these silos, when not done securely, is a serious liability in the system. When it comes to storage, in the interviewee's experience, it really is a factor that can be case-specific depending on the industry that you are working in. However, in their experience, having the data available through blob or json worked well for them.

3.6 Interviewee 5

This interviewee works on much larger, industrial type projects, often involving ML. The work they do focuses on the earlier stages of these projects, which requires setting up the infrastructure required for ML technologies to function efficiently.

The team first goes through a list of aspects that need to be identified before beginning work. They start with the customer needs and intentions to identify what type of solution is required, moving on from there to consider whether it is possible to implement ML, Deep learning or another type of solution. They then look at what the customer's vision is and what they do they want to achieve in the next 3 to 5 years. If there is no such vision, the team will help them create one. This will assist in identifying the needs of the project at hand. Once these two factors have been covered, they will generally be able to tell what type of solution is best suited for the project.

ML projects are generally those that require predictions in some form. When that is the case, they begin by collecting the data and seeing whether the customer is already using their data for any predictions or not. Often enough, there is not a real need for ML as much as advanced BI, meaning that what these customers are actually looking for, instead of prediction, is understanding of past data. This aspect needs to be covered before thinking about using the data for predictions. If no ML is required, a BI platform is built for them to gain the insights from the historical data. This way, the team working on the project will also be able to tell if there is enough data for any ML work.

Having sufficient data is one of the most crucial aspects for ML to be able to function. This goes together with good data collection and storing practices, so the data can be easily accessed. This is often done in some sort of cloud environment. Once that is in place, there needs to be a clear case example of a use case, one for which the data is easily gathered and understandable for the customer. The result of these projects is often that customers gain the understanding they need to be able to know what they actually do with their data and in which direction they can take it to help their businesses.

A big problem that arises in larger companies arises when the data is scattered across multiple teams that each have their own silos for it. This is problematic as it then becomes a mission of getting all data owners to agree to collect their data in a single place. It is important for the company to have some sort of data vision on the highest level, which also covers data collection to remedy this. When they really do not understand what they want to do, that the initial use case becomes very important. Demonstrating that "using this data, you can predict this" or "to predict this, you would need data on this". The difficulty arises from how advanced some of these algorithms are and how complicated they get. If the company does not have a data vision, instead of being able to work on the ML model, the interviewee first has to clean the data and get it to a state in which it can be processed and analysed. What helps in those cases is having someone inside the company that understands the requirements of both the company and the project itself. That person will engage and often be more efficient than having the external teams coming in to tell these companies how to run their business and teams.

Once the data is available in sufficient quantities, then they run visualization on it in order to obtain an initial impression of what is present in the data and what it represents. ML needs a clear objective that the model is trying to predict. Running BI on the data gives us that essential understanding of it and helps in assessing any work which may still be required before using the data for training ML models. The participant referred to exploratory data analytics (EDA) as a common practice and using the customers favourite examples. These are the important factors to cover before being able to move on to building the ML model.

The participant also said that it is often a matter of getting the people responsible on the client side to understand how ML technologies work. In other words that it is not possible to simply say that you will achieve X with ML but that it is rather a calculation made using the data that will give a percentage probability of the desired outcome set for the model. The results will not be clear in the same way than is the case with, for example, a mobile app project. In this case. It is possible to establish from the outset what it will look like, what features it will have, how people will be able to interact with it and so on. ML is more about finding out the likelihood of something happening. It is only possible to measure in terms of how high or low and then only when the data has been analysed. Most customers are unsure about allocating resources for these projects which could potentially not be valuable enough to justify the investment. There must therefore be an initial processing of the data before being able to know what the possibilities are and even then, there are cases in which it might not be possible to gain this information before building the model and seeing initial results. It is very helpful to have such a resource inside the company and most companies have such a person inside the company that can communicate these factors inside the company. Often these will be Product Owners, CIOs, or data/information/technology officers.

3.7 Interviewee 6

This participant has been working on BI and ML technologies for a long time and his first reflection was regarding the price point of these technologies. A decade ago, it was much more expensive, whereas now, it is possible to test and run them relatively cheaply. Not only that, but these are easily available in any region using the various cloud platforms available. For these reasons, companies are starting to show more interest in these types of experiments progressing from data analytics and intelligence to data engineering and science, as the participant describes.

The first requirement for these projects is to understand the problem that the client is trying to solve, followed by the data quality level. Low quality makes it harder to understand the problem. It can sometimes take the data scientist several months before understanding very large datasets, as some of these data sets can have many hundreds of features to go through. Once the data scientist has done that, they can run their calculations to find out the correlations and contributing factors.

This is one of the most crucial elements to identify as the start of these projects which will have the biggest impact on the success of it at the end. This can be quite complex and advanced in certain projects where the data scientist must jump in and figure out the specifics of the projects that they are working on. Should the project be based in a scientific field, or propose a complex solution with which they are not fully familiar , the data scientist not only has to go through the data to get familiar with, but they also have to gain knowledge of the field with which the project is concerned. The participant said that it sometimes takes weeks for the team to go through the academic literature and find the measures/features that would be necessary to get the right prediction, which will help solving the complex problem being addressed through this project. Having a specialist or expert from the client company in those cases is probably the most useful thing for them, as such experts can go through the data sets with the team to explain what it represents in practice, giving understandable examples along the way. The participants said that there had been cases where they didn't get that kind of support, and it took them months to educate themselves on the science before being able to start work on the project. In this situation, an expert would have saved not only the team's time, but also a tremendous amount of resources spent on the project. Having that industry knowledge and expertise helps getting the right starting point and avoids starting with one that changes entirely after months of research.

There are more and more tools related to data science that are available across the various cloud platforms and data analytics software. The entry point is now easier for companies, which are now more able to get familiar with these technologies and try them out easily. It provides the possibility for these companies to get started and test it for themselves and to see the potential results. This does not mean that it would be possible to get the full benefit from ML but is a start. According to the participant, only trained and experienced data scientists can really use ML to its full potential, which still prove very challenging.

When it comes to data quality, they said that they would do a lot of filtering of the data, which can only really be done when there is sufficient data. If there are errors in the data that are fixable, they filter it, otherwise they discard it. The projects where this happens, requires them to look for exceptional cases, with which they can segment the data with more ease. When this is the case, they run a considerable amount of data pre-processing before running it through the models, because of the highly technical nature of the project.

According to the participant, the focus should be on having a methodical approach to the whole process of building the ML model. In other words, it is only by following the data science process that best results can be obtained in the models. This process consists of establishing a hypothesis, a baseline, taking measurements and defining the goal of the project. It is also necessary to be strict with how data is measured, errors, accuracy and position of the results, metrics defining how good or bad the results are. Having that expertise in the team

working on the solution will save 80% of the time that the team will need to spend dealing with these issues. The participant emphasized the importance of getting familiar with the problem as soon as possible in these projects.

4 Discussion

Many topics were covered in the interviews. Through their expertise, it was possible to identify recurring issues which require considerable amounts of time and resources to deal with. This highlighted the importance of being aware of these requirements going forward as all these various technologies become more accessible and utilized. Data usage is still being defined for many companies who are only now beginning to process all the data in their possession.

The success of ML projects is not only about delivery upon completion, it is also about the follow up and maintenance which these models will need as they operate. If the model gave the promised results but was then left without unused and not kept up to date with new data, then it quickly becomes useless. Some of the comments made by the participant stated that clients need to be ready to implement and maintain the models, and not take six months to make the decision to do so.

4.1 Clear Objective

Most of the participants stated that the importance of having a clear goal for the project was a crucial factor to its success. Having the client company clearly defining the business objective that they want to achieve is paramount in ensuring the success of such a project.

One of the examples given was by the second participant who spoke generally about how they tend to start these projects. A lot of the time, clients come with little to no idea of what they want to achieve with their data, largely due to their lack of knowledge of what the data contains and how it is compiled. The open-ended nature of these projects means that a great amount of time is spent just going through the data, deciphering what is possible in the first place; assessing the data quality and seeing if there is enough to train a model on it prior to modelling it to figure out what it represents. Having that clear objective allows everyone on the project to focus his or her efforts towards that one objective. If a company wishes to optimise its sales strategy or improve employee retention at the company, this would already help greatly in finding all the most relevant features related to the end-goal.

Using the example from the fifth interview, where the project involved looking into exceptional cases such as faults in a production line. The objective here might seem clear, but that is not always the case and it was not for the projects in which the participant worked on. Let us consider car factories, they might want to identify the cause of an issue they have that

makes the production line halt every time it occurs. ML in this case has a pretty clear objective, making it much easier for the team to know what the objective of their model is going to be.

This is an issue that we have seen bigger companies being well aware of in the technology field. According to Zawadzki (2018), who discusses Google's way of goal setting and states that Objectives and Key Results (OKRs) are what allows these larger tech companies to get ahead of the curve and have bigger impact in their markets. He describes OKRs as "… the method to transparently align and prioritize resources towards a common goal". It does so by setting a clear objective to the project and supporting it with several key results that the end result will be measured against. The objective representing the 'what' is being done and the key results the 'how' it is achieved.

4.2 Proof of concept

There were various reasons mentioned in the interviews on why proofs of concepts are so beneficial for both the service provider and the client. Taking that initial step before committing to fully develop a ML model is crucial and will likely yield more valuable results once the client is familiar with and can better express what they wish to achieve.

For the client, there are various aspects being covered that they might not know about or believe themselves to have covered but have not done correctly. The purpose of the proof of concept (POC) for them is to get familiar with the possibilities of ML. Once they understand it, they will be better equipped to understand their own needs and how these technologies can help. As the first interview pointed out (echoed by other participants), these projects tend very often to be open-ended, due to customers not being sure what they have in their data, nor about how it's being collected and stored.

For the service provider, many aspects are covered through this. In fourth interview, the participant said that almost all their projects involved them conducting a POC before anything else. It first allows them to identify what needs to be done. It allows them to also plant seeds and prepare the data for the needs that they can already foresee should the project proceed past the POC. From there, they can assess the structure of the system that the model is intended for, seeing whether or not there are major flaws in the silo(s) that could compromise its integrity and trustworthiness. It helps to see the data spread and if it is easily accessible in one place. Once these points have been dealt with, they will then be able to say what they are able to do and develop the POC for the customer. This allows them to demonstrate the possibilities of ML and inform on the work that needs to be done at a foundational level before venturing into ML models. Through this, they will also be able to explain with much more clarity the implications of having a ML model, as often customers seem not to be aware about matters such as technical debt and training the model to keep it up to date for example.

The POC will give both parties the chance to gain the information that they each need before committing to a larger scale project, mapping out how it would proceed. Depending on the awareness and preparation of the customer, the POC may be done more quickly or slowly, indicating whether the customers initial request is possible to fulfil at this stage or not.

4.3 Data Accessibility

Data accessibility was an important factor that was mentioned many times. The principal reason many companies have not been preparing for the requirements of these new technologies and the data is that they are only available with difficulty. The most interesting aspect of this is the synergy of technologies and how cloud computing allows for ease of access. In an ideal way for technologies such as ML or machine vision.

It's a complicated and long process to talk to all the data owners to convince them to combine their data with that of the rest of the company, chiefly for the change it represents for the team's workflow. They may also have reservations on implementing such technologies into their systems. However, once the data has been consolidated into one easily accessible location and stored in a proper format, then the job of the data scientists and engineers working to develop the model becomes much less complicated.

This is a concept that is explored by Wu, C. Buyya, R. and Ramanohanarao, K. (2016) in their paper in giving the numerous reasons why that is the case. The main reason advanced is that cloud computing is affordable and can offer ML on large-data sets in processing and scalability. This is helped by the various services available in these cloud's marketplaces that offer services which take away much of the hard work of setting up the correct infrastructure and ensures that the data is stored in the correct manner.

4.4 Data Quality & Documentation

Understanding data quality is crucial to preparing for any type of AI technologies. ML is one which is prevalent now, but this may not continue to be the case as time passes. One thing that many companies and media seem to agree upon however, is that data is now the most valuable resource that a company can have (Economist, 2017).

Having that preparation and data readiness is therefore not only useful at the present time but will prove very valuable for a great many other reasons. There are already several ways that data is being used to generate insights other than by ML and these are much simpler to achieve when the data is ready for such analysis work. Many businesses already use Excel as an Analytics tool for their data, then some others which are adopting the use of BI tools (described as "Excel on steroids"). Preparing data will already allow companies to perform better analysis with existing tools that are used all around. It is then the documentation that will allow others looking at the data from an external point of view, to understand its context

and meaning. With that, they can then have a much greater understanding of what is being represented and what the implication of such data is what might then otherwise seem impossible from the numbers alone, becomes a lot more understandable and logical. The documentation will also help those trying to use the data to know the labels applied and how to structure their backend calls when they need to.

4.4.1 Data Readiness Levels by Neil D. Lawrence

Neil D. Lawrence wrote a paper called “Data Readiness Levels” in April of 2017. In it he describes the three main bands or levels by which data can be qualified according to new criteria. Each of these have sub-levels which for the purposes of this thesis, we will only focus on the three main bands. He divides it into bands C, B and A in that ascending order.

4.4.1.1 Band C

Band C being focused on the accessibility of the data set, it addresses the issues related to what data is being recorded, which format it is being recorded into, the privacy or legal constraints of that data, and the topology of the data (present across multiple devices). At the highest level of this band, the data is easily available in a format which can be uploaded to analysis software, being sufficiently readable for machine processing. This band would include any cleaning phase or wrangling of the data to get it to a readable state.

4.4.1.2 Band B

This band deals more with the trustworthiness of the data, ensuring that it is a good representation of what it measures and that everything that is required for the analysis is present in the data set. EDA (Exploratory Data Analytics) is a good way to investigate this band and assess the stage which it is at, running the data into Visualization software and seeing what can be rendered from it. Through this, the decision makers will also be able to grasp the capabilities and limitations of their data set and understand how it can be used to support their operations.

In this phase, data owners should also verify the integrity of the data by:

- Checking that all errors and missing values were handled correctly.
- Checking for any possible anomalous values.
- Ensuring that there is no bias present or if there is, that it is removed.
- Proper documentation of the data and the process through which it has gone through. Mapping the various changes that have been made on the data from collection to its current state.

4.4.1.3 Band A

Once all these criteria have been answered, it is necessary to focus on the context within which the data is being used. This is done by defining the purpose of the data set and assessing how useful it is in that respect. Typically, the purpose could be related to the quality assurance of a product, or the best suited ad for a person.

If this has been done and the data is considered to good enough then it can be deployed for use. For this to happen, it can sometimes take a great amount of annotation by an expert who has a complete understanding of the data and what it represents. During this step, it may be that the user realizes that new data needs to be collected for better results. This can be a long process depending on the issue to be addressed. Hence the importance of starting to collect data early and in a usable format. At this stage, it really becomes statistical work in analysing the value of the existing data while using logic as far as possible to figure out what other data might be valuable to the process.

4.4.1.4 Understanding Data

Lawrence goes through a few examples of how these principles can be applied and how they function in practice. He also reminds us that each of these levels mean certain tasks need to be conducted on the data to ensure it fits all these criteria. Tasks which require different skillsets and maybe different knowledge of the data collection processes that are in place.

As a company, having that understanding of the whole process that the data goes through, becomes critical. They need to understand exactly how and where the data is collected as well as all the annotations or changes that might have been made on it. With that knowledge and preparation, it becomes much easier and efficient to develop any potential model on the data.

4.5 Data Strategy

Most of the participants stated that they wished their clients would have some sort of data strategy in place before getting into ML projects. Two factors that they often have to struggle with before being able to work on the ML models and predictions are documentation and accessibility.

Having a data strategy helps that. It covers not only these two factors, but also the ones mentioned in the Data Readiness Levels and lays the foundations for more advanced technologies, preparing the data for ML models and other technologies which would be data intensive. It also gives the company a macro understanding of the data that they have and how it can be leveraged in their decision-making. Such information is crucial for the successful development of any machine learning models in their systems.

In 2017, the Harvard Business Review wrote an article on data strategy in which they stated that 80% of analysis time was spent on discovery and data preparation. This does not seem to have changed significantly according to the interviews conducted in this thesis preparation. Most participants state that they would spend a large amount of time on exploring and preparing the data for modelling and model training. (DalleMule, L. & Davenport, T. H, 2017)

Having a data strategy allows a company to put their data to work, meaning that they will use it to support the decision-making taking place in the various projects that they might be working on. If done correctly, this covers all the criteria covered by the data readiness levels. This is reflected by the HBR article as well, which states:

> "Having a CDO and a data-management function is a start, but neither can be fully effective in the absence of a coherent strategy for organizing, governing, analysing, and deploying an organization's information assets." (DalleMule, L. & Davenport, T. H, 2017)

Once the strategy is in place, all the foundation elements for these more advanced technologies should be in it and simplify the entire process for the team tasked with the ML project.

5 Conclusion

In conclusion, through the interviews, we were able to highlight the various topics that were deemed as instrumental to successful machine learning projects. The issues were seemingly more related to the data and the infrastructure in which it is stored than with the machine learning models themselves. Many companies seem to be getting ready now for the technologies that they wish to put in place. Unfortunately, this does not detract from the time required to implement the changes into these systems for the successful implementation of ML tools. Once these requirements are met, then it becomes an issue of understanding how ML tools support the decision-making or how actionable the results of these are, which depending on the context might be more useful either way.

The focus areas can be loosely put into three categories: Being Data Ready – Understanding ML Technologies – Clear Business goal for the ML Model. The importance of having what is referred to as ready data or "clean data" is paramount to the successful execution of the ML models in operation. Should this data not be prepared accordingly, it will keep the model from being able to provide any reliable results. Understanding the data and its representation is then the follow up to that as the results will only be useful if they provide a clear goal for the business, that way provide business value through the analytics that it runs. Enhancing the leadership team's ability to execute and understand the changes occurring in their company.

Companies looking into potentially implementing ML technologies into their offering in their

future should already consider gathering data and doing so in an efficient manner that will avoid any additional work further down the road when they are in a position where they can implement such technologies. Doing the work to familiarise themselves with ML now will not only them money but also time and other important resources that they will be able to allocate towards outputting valuable results from the models.

References

Printed sources

Shalev-Schwartz, S. & Ben-David, B. 2014. Understanding Machine Learning From Theory to Algorithm. Cambridge University Press.

Lawrence, N. D. 6 April 2017. Data Readiness Levels. Amazon Research Cambridge and University of Sheffield.

O'Neil, C. 2016. Weapons Math-Destruction. Crown Books.

Wu, C. Buyya, R. & Ramanohanarao, K. 13 January 2016. Big Data Analytics = Machine Learning + Cloud Computing.

Electronic sources

Pranav, D. Popular Machine Learning Applications and Use Cases in our Daily Life. Analytics Vihdya. 15 July 2019. Accessed 26 February 2020. https://www.analyticsvidhya.com/blog/2019/07/ultimate-list-popular-machine-learning-use-cases/

Shen, S. 2019. How to Create a Successful Data Strategy. TowardsDataScience. 7 November 2019. Accessed 25 February 2020. https://towardsdatascience.com/how-to-create-a-successful-data-strategy-1293bacf463c

Marr, B. How To Create A Data Strategy: 7 Things Every Business Must Include. Forbes. 19 March 2019. Accessed 25 February 2020. https://www.forbes.com/sites/bernardmarr/2019/03/11/how-to-create-a-data-strategy-7-things-every-business-must-include/#7a7cb04c218b

Ajanki, A. Difference between machine learning and software engineering. Futurice.com. 11 November 2018. Accessed 25 February 2020. https://www.futurice.com/blog/differences-between-machine-learning-and-software-engineering/

CHI Software. Supervised vs. Unsupervised Machine Learning. CHI Software. 20 May 2019. Accessed 29 November 2019. https://medium.com/@chisoftware/supervised-vs-unsupervised-machine-learning-7f26118d5ee6

Brownlee, J. 14 Different Types of Learning in Machine Learning. Machine learning Mastery. 11 November 2019. Accessed 29 November 2019. https://machinelearningmastery.com/types-of-learning-in-machine-learning/

Zawadzki, J. The Power of Goal-Setting in Data Science. TowardsDataScience. 11 October 2018. Accessed 19 December 2019. https://towardsdatascience.com/the-power-of-goal-setting-for-your-data-science-project-9338bf475abd?gi=b1d91a178d23

Marr, B. 27 incredible Examples of AI And Machine Learning In Practice. Forbes. 20 April 2018. Accessed 19 December 2019. https://www.forbes.com/sites/bernard-marr/2018/04/30/27-incredible-examples-of-ai-and-machine-learning-in-practice/#4769a5ed7502

Moroney, L. Machine Learning Zero to Hero (Google I/O'19). TensorFlow Youtube Chanel. 9 May 2019. Accessed 20 December 2019. https://www.youtube.com/watch?v=VwVg9jCtqaU&t=1461s

Economist. The world's most valuable resource is no longer oil, but data. The Economist. 2017. Accessed 15 December 2019. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

DalleMule, L. & Davenport, T. H. What's your Data Strategy? Harvard Business Review. May-June 2017. Accessed 28 November 2019. https://hbr.org/2017/05/whats-your-data-strategy

Marr, B. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. 21 May 2018. Accessed 26 November 2019. https://www.forbes.com/sites/bernard-marr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#6747f7c960ba

Avanade Official Website. 2020. Accessed 10 October 2019. https://www.avanade.com/en

StatisticsHowTo. 9 December 2014. Statistics How To. Snowball Sampling: Definition, Advantages and Disadvantages. Accessed 1 October 2019. https://www.statisticshowto.datasciencecentral.com/snowball-sampling/

Figures