Why it is so hard to get a Machine Learning Engineering job

Umberto Michelucci[1,*]

[1] *TOELT LLC, Duebendorf, Switzerland*
[*] *Email: umberto.michelucci@toelt.ai*

October 14, 2020

A brief story of how a real machine learning project looks like and why is so difficult to get a Machine Learning (ML) jobs. The goal of this paper is not to demoralize people that are interested in making the switch to ML, but to give them realistic expectations on what skills they may need. I also discuss briefly why many online courses are only a waste of money and finally I hope I give a good idea about how exciting this field is.

1 The story

A friend messaged me on LinkedIn the other day, saying that she has a friend that is really interested in getting into artificial intelligence. She told me he would like to know if I have some time to talk with him since I work in Artificial Intelligence (AI). After a few messages exchanged on LinkedIn, I found myself looking at this man through the computer video in a zoom video call. He told me that he did not have much background nor in programming nor in math (actually he told me he was bad at math), but he thought that going into AI would be a good career move. He had always worked in Business as project manager and in a role similar to a business engineer. He taught it would be a great job to act as a bridge between the business world and the technical one in AI projects. He wanted to know from me what was the best way of achieving that. Could I give him some tips? I get this kind of request many times every month from the most different people. I mentioned to him that to get into machine learning (call it AI if you wish), he would need programming and a lot of math and that he would probably require at least a couple of years to get a job in a junior position. Still, I could tell this was not the answer he was really expecting. Many people slowly realize that technical jobs in AI are challenging and require a lot of work and a solid background in math and programming. Still, they don't really want to believe it. Apparently, a commonly believed opinion (or maybe a hope) is that it is very easy to get the needed skills with a few online (or in-person) courses. It will be useful to tell a story about a project I am working on to give the reader a vague idea of practitioners' technical challenges. Note that this article has not the goal of convincing people that it is not possible to get into AI. On the contrary. I hope you get an idea of the incredibly challenging and exciting aspects of the job. My goal is only to give the reader interested in getting into AI more realistic expectations. I also want to give people a practical way of judging online (or in-person) courses. One should avoid wasting lots of money in courses that promise to make you an "expert in <insert number of weeks here> weeks"! Or paying internet gurus that promises to teach "everything" in <put a machine learning topic here> as it is also generally a bad idea. This ability to judge those kinds of courses is, in my opinion, incredibly challenging to get for young people that may have not enough experience in the field. If you are not one hundred percent convinced, please ask someone with experience. The Machine Learning community is a friendly one, and people are typically helpful. Contact people on LinkedIn, for example, even myself (you will find me on Linkedin, I am happy to help you if I can) if you are in doubt. Doing machine learning as a job is incredibly rewarding, incredibly exciting, but also very hard. Don't be discouraged, but also have realistic

expectations about the time and effort needed to get the necessary experience. Many online courses and many universities or "academies" promise to make you an expert in a few weeks in machine learning, deep learning, or data science (remember? The sexiest job in the 21st century). Unfortunately, too often, those courses are simply frauds. Imagine advertising a course that would make you an Olympic athlete in 3 weeks. Would you believe that? Of course not. Someone will, naturally, tell me that there are lots of other jobs in AI that do not require such skills. For example, report developers do not need to know how neural networks are working, right? Right, but a report developer is not working in Machine Learning. I am not talking about jobs in Business Intelligence, not in data engineering or software development. I am talking about jobs in Machine Learning. These jobs deal with complex models in advanced fields like computer vision, outlier detection, algorithm research, big data processing and so on. Now you may be curious about the story, so here it is.

Imagine a big company with a big Unix server farm, roughly more than 1000 servers, each logging user activities. Imagine also a user base of approximately 3000 people. Imagine now the following dialog. It did not happen this way, but it could well have.

2   The Request

"Hey, you are the AI guy, right?"

You know things are not good when someone knows you as the AI guy.

"Well, yes. How can I help you?"

"Easy. I need to check all users' login activities on all servers and find out if something strange is happening."

And, if you have some experience, your next question should be:

"What do you mean by strange?"

And if you have any experience, you should know that the answer will be:

"We don't know exactly. Can you use some kind of AI to find automatically any suspicious login activity?"

3   The data

Knowing that I would not get much more information from him, I thought to start with the data. Remember how the story begins? 1000 servers and 3000 users. How can we get all the log files? This imaginary company uses a tool to get all the log files together. So we were able to get a month's worth of log files downloaded using this tool. That meant to learn the tool's query language (something eerily similar to SQL, but less clear) because, you guessed it, nobody was able to do it for us. Note that at this point, we did not even start with machine learning. We realized immediately that a single login event was split in a variable number of entries in the log files. First things first, take all the log files and merge all those entries in one record. Sounds easy? Think again. We are talking here about roughly 300 millions of events in one single month. Of course, the entries are not always the same. Sometimes they are 10, sometimes they are less or more. Another thing to learn: how linux log events. Having understood that, we ended up with a list of 300 million records, each representing one event. And we did not even start looking at models. By the way that needed some serious text processing of the log files in case you know what I am talking about.

4   Some Machine Learning

Something was clear from the beginning: we were talking about outlier detection. That means identifying cases outside the norm, whatever the norm is (you remember the guy answering "We don't know exactly" to the question "what is strange?"). You may start noticing a trend here: there are many questions in machine learning problems, but not always clear answers. We wanted to find login events that are not standards. A pity we could not define "standard" in any measurable way. Let's talk about some algorithms now. One common choice for this kinds of problem is Principal Component Analysis (PCA). If you don't know what I am talking about, don't worry, it is not relevant to this article. Enough for you to know that the algorithm needs to process all the input data in one shot, and 300 million records don't fit any standard computer memory. Batch processing? Not really an option (for many reasons that go beyond the scope of this article). Ok, first obstacle. The solution? We turned to auto-encoders and reconstruction errors (again, if you don't know what these means, don't worry). What is relevant here is that those are particular kinds of neural networks. That means we needed to be able to write code for neural networks and understand how auto-encoders work. After preparing the data for the algorithms, the dataset size was roughly 1.4 Tb. This presents its difficulties, as you can imagine. Not only we needed to be able to write code for neural networks, but the code was supposed to be so efficient that we could deal with 1.4 Tb of input data. Let's summarise what we needed so far: understand how Linux logs events, how to re-structure the files, how to deal with 1.4 Tb of data in a reasonable time, how neural networks works, and how to write code in libraries like TensorFlow or pytorch (libraries explicitly developed for neural networks) and how to write efficient code

in Python (if you don't like to code you are in trouble). You may imagine that all this was done by a team of many people with different skills. If you think this, you probably don't have much experience. If you have been around the block a few times, you know how many people had to do all this: one. But wait, we are not finished. After all this, we need to validate the models, find the best parameters for the model, judge its stability, make sure that the results make sense, and a few more tasks that may be less interesting to the reader but not less important.

5   And Finally Those Misterious Soft Skills

All right. Still with me? Now you will need all the soft skills you have (you have those right?). How do we present the results? Can you explain how an auto-encoder works to someone who does not even know how to write the word "mathematics" or "algorithm"? Oh I forgot. And you have to do it in 10 minutes, since the managers don't have so much time. They probably have much more important things to discuss. And since the guy before you in the meeting took longer than expected, try to do it in 5 minutes. Easy right?

6   Is That All?

I will throw in here a few other tools and skills that you will need for such projects. If you don't know what I am talking about is fine. Just imagine having to learn about all those in a few weeks (or even days) during this super special online course that will make you the ultimate "deep learning guru." Here are a few: how code versioning and git work, how to do proper hyperparameter tuning, how to write performant code that can work with Tb of data (believe me is not easy), how to read research papers to investigate what kind of auto-encoders is working in such special conditions and a few others that are not worth mentioning, although important in practice.

7   TL;DR (Too Long; Don't Read)

If you have read all the article up to here, congratulations. If you have not, it is also ok. I am not offended. The main message that I have here is that getting into machine learning is really hard, but extremely rewarding. I think that the biggest obstacle is not your background, how much mathematics or programming you know but how much you are willing to work and study. Take your time and ask for help from more experienced people. I think in AI, mentoring is one of the most (if not THE most) meaningful way of getting the necessary experience fast and reliably. If you have the possibility of doing internships, work with more experienced

researchers and practitioners do that. Ask for suggestions, go to meet-ups, and talk to people and ask for help. And please, before wasting money on courses or internet gurus, ask someone on LinkedIn or that you know personally, that has the experience. And the most important thing is you should go into AI if you have fun doing it. It means a lot of work and requires a lot of energy and time. Do it if you love it. Believe me, you will have lots of fun. In case you need help please feel free to contact me, I am happy to help you in any way I can. You can find me on **LinkedIn**. Feel free to get in touch. Would love to be able to help you getting in AI.