Somraj Saha
Mar 4, 2020 · 6 min read · Listen

Domain Knowledge — The Second Most Important Skill to Have as a Data Scientist.

You mightn't realize it but if you've ample years of experience in a very specific domain of expertise, you might be eligible to be part of a Data Science team.

Photo by Matthew Feeney on Unsplash

Data Science, dig right into building models, who cares about Domain Knowledge, Right?

Data Science is all in the blaze at the moment! A quick search of the keyword on Google yields not a Wikipedia page of the field but hundreds of tutorials & courses on the first page of the search results.

Although not a bad thing per se, the easily accessible online learning resources helped a lot of self-learners, myself included, to wet our feet into the ocean. Without which, it's difficult to comprehend trying to learn all by ourselves. Surprisingly though, I noticed how rare it is, to come across any mention of "Domain *Knowledge"* in those resources, albeit even briefly.

Perhaps, they are more targeted towards the misguidance of "*learn Data* Science *in a month and top the Kaggle leaderboards!*" which is a major problem the community needs to address urgently. Rather than making people competent to be employable in a Data Science role, the massive online learning platforms are only churning out incompetent "programmers" who think a simple RandomForests Classifier from the Scikit-learn library is a solution to all problems!

Granted, basic Statistics, Mathematics & coding skills are some of the harder skills to pick up, the prospect of a minimum viable subject matter expertise is more often than not, neglected by a Data Scientist. But in contrast, Subject Matter Expertise should be considered king! Especially with extensive experience working in a specific field it could potentially be the most valuable skill in his/her bag.

Domain Knowledge — How Does It Help a Business?

Contrary to what was mentioned earlier, notice how most participants in a Kaggle competition don't have any substantial subject matter expertise. Yet, regardless of the absence, they go ahead to win competitions one after the other with a high score in the leaderboards.

And that's because, fortunately, someone, somewhere, was smart enough to think & ease the process of making predictions. Thus high-level Predictive Analysis libraries like Scikit-Learn do most of the heavy lifting in the backend, yet, the libraries are robust enough to still yield surprisingly good results even with default parameters. With just a couple of lines of code, literally, any Tom, Dick & Harry is capable of training a model on the dataset & submitting it to Kaggle, achieving at least a top 50% score on the leaderboard with minimal effort.

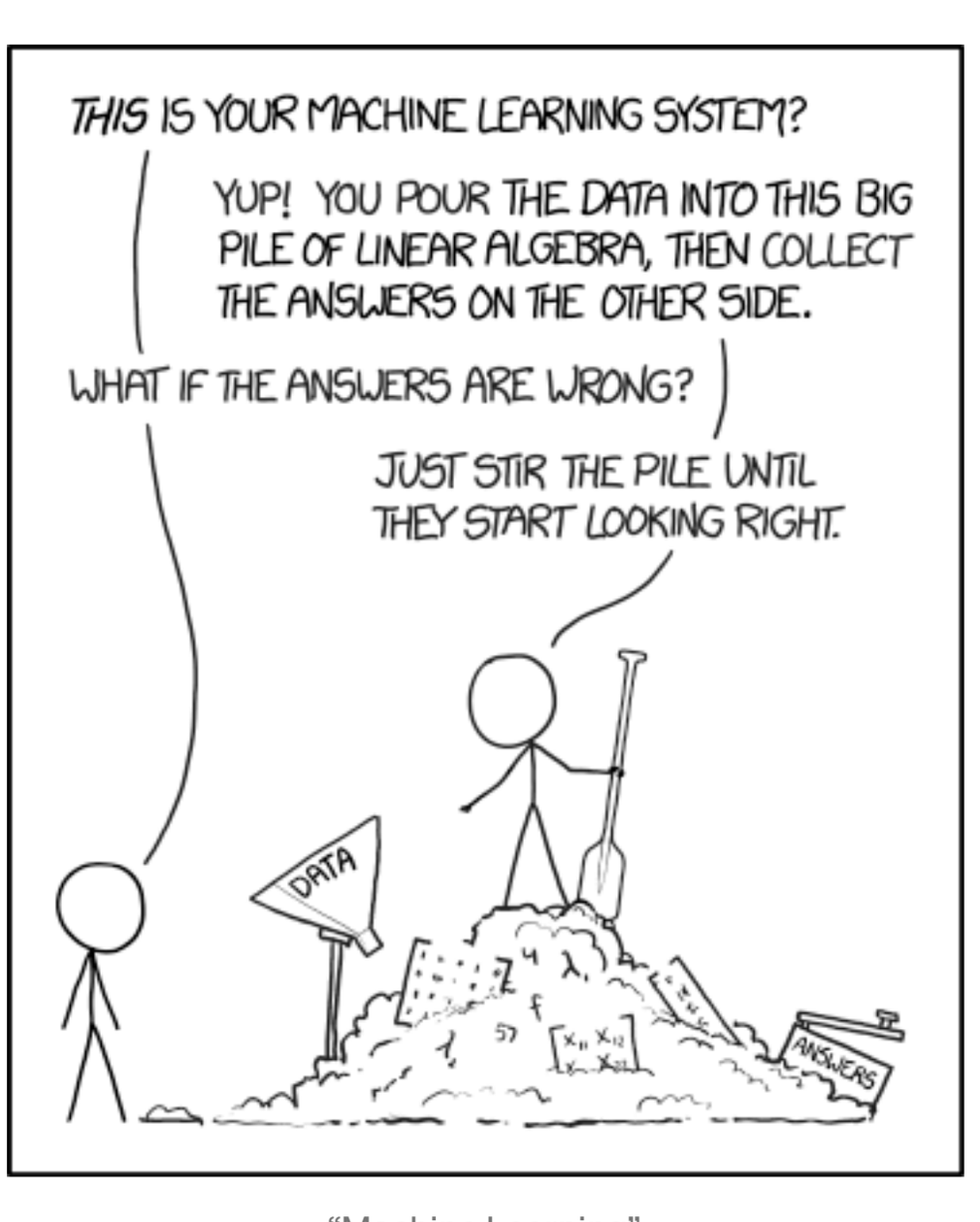
On the flip side, businesses work under major financial & time constraints while trying to sustain their place in the market. Not to forget, they are also in the market to sustainably create a profit margin for themselves. Besides in general, for most businesses, it's just not viable enough to invest in developing an algorithm specific to their domain, in-house. Hence, they hire for the much needed Data Science role, hoping that the new hire would help resolve the problem they were facing. Also if an opportunity arises, to move forward with it or possibly, to capitalize on it.

Why is Domain Knowledge essential for a Data Scientist?

Interrelated to each other, yet clearly distinguishable, three aspects of Domain Knowledge, a Data Scientist should keep in mind, can be defined in context to the —

1. *The source problem, the business is trying to resolve and/or capitalize on*.
2. *The set of specialized information or expertise held by the business*.
3. *The exact know-how, for domain specific data collection mechanisms*.

On the other hand, a rather unfortunate misconception the general public has about Data Science & ML is, how ML & AI is the mythical Noah's Ark, set on resolving every trivial problem ever faced.



"Machine Learning"

Depicted humorously, the author summed it up on the xkcd comic where Data Scientists are viewed as wizards from Hogwarts with a Magic Wand named "*Machine Learning"* capable of resolving any problem they're facing or want to make some profits from.[1]

But contradictory to popular belief, a Data Scientist needs to prioritize planning ahead with a sustainable & logical business strategy, followed by the implementation. To give an analogy, constructing a Space Shuttle to travel between New York & Tokyo sounds like a fool's errand. Similarly, a Cats & Dogs classifier doesn't have any sustainable & profitable business prospects. Instead, adapting to the business sector & gaining the necessary knowledge of the domain will be more beneficial to the business overall, rather than the technical know-how to build the prediction algorithm right away.

Secondly, and perhaps the most discussed topic in the Data Science community is in context to the information held by the business. This information acts as the Rosetta Stone, helping the analysts find better ways and/or means to perform his/her job. Prior information about the industry & the domain augments the process of making more precise & accurate predictive models based on the available features in the dataset. The other benefit being that, the model would then generalize better into real-world situations.

Besides, emphasis on the importance of Feature Engineering & how doing so can improve the overall accuracy of the model are common & is a topic of discussions across every corner of the community. But performing proper & insightful feature engineering is a skill, only a few experienced ones among the whole bunch is capable of doing properly.

Hence, reminds me, that I came across a rather interesting piece of work by Xavier Martinez published at — Catalonia GDP: Insights & Regression Analysis which is a very detailed & prime example of feature engineering components of a dataset to create newer columns/features for further analysis. He predicted the Catalonia GDP growth rate based on feature engineered GDP components from the dataset & it shows how being extensively versed in a domain can help make very insightful & precise observations. Xavier did exactly that, based on what we Economists call a "*Demand-Driven Growth*".

Lastly, note that while you read through this article, 1.7mb of data is being generated worldwide each second which accounts to 2.5 quintillion bytes of data per day. That's a whole lot of data to harness & process [2]. Comprehending what portion, the how & when to process that chunk of data, is paramount. Not only would it reduce inefficiency in the business operations, but as mentioned earlier, time & finances are the biggest constraints for a business. Being able to trim down to just the bare minimum for the required analysis helps reduce costs & processing time as well.

The Community Should Be More Vocal About Domain Knowledge.

Hence, I assume it is safe to conclude stating the importance of focusing on Domain Knowledge in a Data Science role. Besides, it's the community that should preach about the same only then would a business find a competent employee for a Data Science role in his/her company. But bear in mind, even with all the preaching, Domain Knowledge can be picked up while on the job and isn't much of a difficult thing either but neglecting it, would be utter irresponsibility.

References:

[1] xkcd, Machine Learning(2017)
[2] Domo, Data Never Sleeps 6.0(2020)

. . .

I'm a **Freelance Data Analyst** helping business understand how to harness their data needs. If you've any related questions, feel free to reach out to me on **Twitter**.

Besides, if you enjoy reading such articles of mine, you might even enjoy reading these:

- **Dropping Missing Values? You Probably Shouldn't**.

Or you could even **SUBSCRIBE** to my mailing lists to be instantly notified of any new updates!