5

THINGS TO KEEP IN MIND WHILE COLLECTING DATA FOR YOUR ML PROJECT

“Data is the new oil.”

- Data collection is the process of gathering and measuring information from countless different sources.
- Data is at the core of nearly every business decision made.
- Predictive models are only as good as the data from which they are built, so good data collection practices are crucial to developing high-performing models.
- The data need to be error-free (garbage in, garbage out) and contain relevant information for the task at hand.

1. Identify what you want

- t’s a good exercise to start brainstorming a list of potential features that may be useful for a given ML task, disregarding feasibility or cost.
- Once you have a list of potential features that could help with the prediction task you can prioritize by availability (do those data exist?), accessibility (do you have the rights and consent needed?) and cost (how costly is it to collect those data?).

2. Data access

- First thing is to search on free and publicly available resource like Kaggle, Reddit, Google dataset search engine, UCI Machine Learning Repository etc... (You can find more about this on my previous post on dataset search engines)
- Second way is to extracting data from web by using either target website API (If it provides) or by using any scraping tools like (Scrapy).
- If you want to use IOT data then you can simply collect your data from the sensors.

3. Data size

- In most cases, more data helps to build better model.
- In some cases, more data will not help (Diminishing points)
- If little or no data available, transfer learning may help you.
- Acquiring labelled data may costs money sometimes. But this depends on you how much money and time you willing to spend.

4. Data quality

- Check your data is not biased, always try to collect data from all perspectives. Check this amazing article on how bad is biased data (Link)
- Check whether your scrapper is collecting good data or not (If it is collecting noise then stop it immediately and rectify it).
- In addition, if there are concerns about the quality of your training data, this is a red flag that you should raise in your PRD. This potentially reduces the quality of your ML model and increases the execution risk.

link -
https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

5. privacy and security

- **How securely you will store the collected data?**
- **Do you have permission to collect/use the data?**
- **How you are protecting/training the personal data of your users?**