Published in Towards Data Science

Low Wei Hong Follow
Oct 27, 2019 · 5 min read ★ · Listen

How To Start Your First Data Science Project

Kick Start Your Data Science Journey

Having some data science projects to showcase during your interview is one of the most important prerequisites to enter the data science field.

You may have asked some questions as below:

> How should I start a data science project?
>
> Which kind of topics should I do?
>
> What kind of dataset should I use?

Previously, when I was thinking of what kind of data science projects should I work on, I strived to come out with some practical yet unique ideas to start my first data science project. There are tons of data science projects online but I did hope that I could do something different.

One thing to note is that the data science project which I am talking about is not a school project, it is about a personal data science project.

You might be wondering, **why not school projects**?

Let's imagine you are the hiring manager and you are looking to hire only one data scientist. Let's say there are only two candidates in your mind. Both of them seem to have a similar background, but you are thinking of only one to hire. Therefore you give both of them a chance to interview and you told them to prepare their previous projects to be presented to you.

Candidate A: This is a school project which I have done in a team of 4. I am in charge of

Candidate B: Here is a problem which I have been facing when I was looking for a rental unit — How do I know whether the rental fee is rational in the area base of my preference on the unit. Therefore, I started to work on this project to clarify my doubts.

So, which candidate you would be more interested to listen to? I think the answer is quite obvious, which is candidate B.

In this article, I am going to share step by step guide on how to start a personal project.

Step 1: Identify a Real-World Problem to Solve

Find your own itch. For example, you have a website selling keychains. You have been tired of gazing through comments, and you would like to automate this process. Therefore, you can build a sentiment analysis model to identify whether your customers are satisfying.

Maybe you are thinking this project is too easy, and willing to take to the next level. You can use build a topic modeling model to know which area you can improve. For instance, one of the topics you found out your customers keep commenting about is the lack of variation on keychain design. Thus, by not looking through the comments, you can know which area you should improve on.

On the other hand, maybe you are owning some stocks. You are following some rules or technical indicators to buy and sell your stocks but you would like to automate it. Here are two ways to automate it, either you write a program to find out the signals or to train a model that could read in the latest news so that you would be able to sell or buy it faster than the appearance of the signals.

You can always still start from Kaggle projects, but if you can identify and solve your problems, it implies that you have the ability to define and solve a problem on your own.

1. https://www.analyticsvidhya.com/blog/2018/05/24-ultimate-data-science-projects-to-boost-your-knowledge-and-skills/
2. http://intellspot.com/data-science-project-ideas/

Step 2: Decide which dataset to work on

You can choose to use open-source datasets such as Kaggle, but you can also choose to gather the data by yourself.

There are a few ways to retrieve data by yourself. The simplest way will be to use the Application Program Interface (API) provided by the target website. If you want to challenge yourself, you can try to scrape websites to obtain your data. Usually, the data will be dirtier and hence you can showcase your data cleaning skill.

If you do not want to crawl data, my advice is to choose a larger dataset, so that you could have more exposure in dealing with large datasets.

On the other hand, you can choose a dataset that is more difficult to be retrieved. For example, when I was trying to carry out analysis on Malaysia Car Market, the public dataset I could get is only PDF scanned images. Therefore, I have to perform OCR (Optical Character Recognition) to extract the data from the tables that reside in PDF.

If you are interested to know more detail on how I manage to extract the data, feel free to visit this link.

Links to access open-source datasets:

1. https://github.com/awesomedata/awesome-public-datasets
2. https://www.kaggle.com/datasets
3. https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-

Step 3 Perform analysis and modeling

Before trying out any models, try your best to deep dive into the data. Look in detail to identify possible patterns or trends to be fed into the machine learning model.

Do note that the collecting, cleaning, and analysis part will consume most of your time. **Don't emphasize too much on the modeling part, instead, you should spend more time on the feature engineering portion**. If you are able to identify a good feature for your model to learn, there is a great chance that your model could learn well from the feature.

Besides, plot useful charts which you think are important to solve the problem. **Don't anyhow draw graphs just for the sake of showing your plotting skill**. Always remember, the skill that most of the people are looking for is the **capability to overcome business obstacles**.

Final Thoughts

For those who are not coming from a data science background, projects could be described as a proxy of your ability in solving a task. Therefore, do choose a topic wisely and work hard on it.

> Work hard in silence, let your success be your noise — Tiffany Trump

Thank you so much for reading till the end, this is my new website on providing web crawling service. If you would like to find someone to crawl the data for you, do not hesitate to approach me via Linkedin or the above-mentioned website.

Low Wei Hong is a Data Scientist at Shopee. His experiences involved more on crawling websites, creating data pipeline and also implementing machine learning models on solving business problems.

He provides crawling services that can provide you with the accurate and cleaned data which you need. You can visit this website to view his portfolio and also to contact him for **crawling services**.

You can connect with him on LinkedIn and Medium.