Data Science East africa

Steps of a Data Science Project Lifecycle

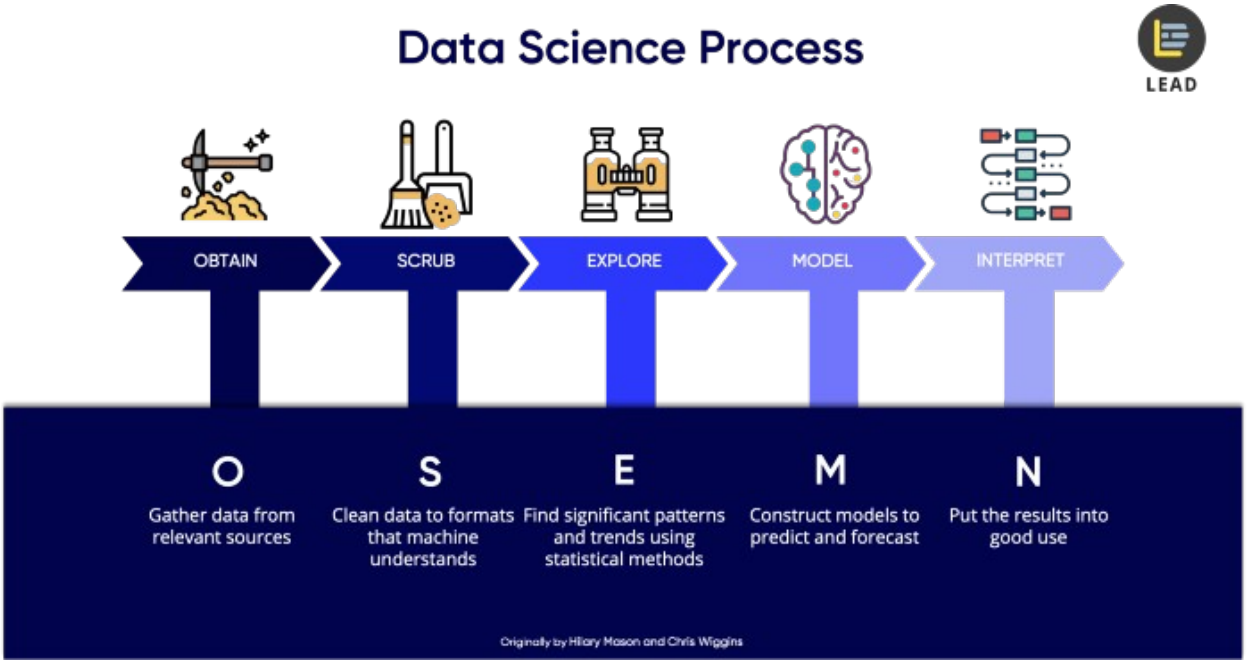
Day 19/20

Often, when we talk about data science projects, nobody seems to be able to come up with a solid explanation of how the entire process goes. From gathering the data, all the way up to the analysis and presenting the results.

In this post, I break down the data science framework, taking you through each step of the project lifecycle, while discussing what the key skills and requirements are.

The OSEMN framework

Here we will walk through this process using OSEMN framework, which covers every step of the data science project lifecycle from end to end.



1. Obtain Data

The very first step of a data science project is straightforward. We obtain the data that we need from available data sources.

In this step, you will need to query databases, using technical skills like MySQL to process the data. You may also receive data in file formats like Microsoft Excel. If you are using Python or R, they have specific packages that can read data from these data sources directly into your data science programs.

The different type of databases you may encounter are like PostgreSQL, Oracle, or even non-relational databases (NoSQL) like MongoDB. Another way to obtain data is to scrape from the websites using web scraping tools such as Beautiful Soup.

Another popular option to gather data is connecting to Web APIs. Websites such as Facebook and Twitter allows users to connect to their web servers and access their data. All you need to do is to use their Web API to crawl their data.

And of course, the most traditional way of obtaining data is directly from files, such as downloading it from Kaggle or existing corporate data which are stored in CSV (Comma Separated Value) or TSV (Tab Separated Values) format. These files are flat text files. You will need to use special Parser format, as a regular programming language like Python does not natively understand it.

Skills required

To perform the tasks above, you will need certain technical skills. For example, for Database management, you will need to know how to use MySQL, PostgreSQL or MongoDB (if you are using a non-structured set of data).

If you are looking to work on projects on a much bigger data sets, or big data, then you need to learn how to access using distributed storage like Apache Hadoop, Spark or Flink.

2. Scrub Data.

After obtaining data, the next immediate thing to do is scrubbing data. This process is for us to "clean" and to filter the data. Remember the "*garbage in, garbage out*" philosophy, if the data is unfiltered and irrelevant, the results of the analysis will not mean anything.

In some situations, we will also need to filter the lines if you are handling locked files. Locked files refer to web locked files where you get to understand data such as the demographics of the users, time of entrance into your websites etc.

On top of that, scrubbing data also includes the task of extracting and replacing values. If you realise there are missing data sets or they could appear to be non-values, this is the time to replace them accordingly.

Lastly, you will also need to split, merge and extract columns. For example, for the place of origin, you may have both "City" and "State". Depending on your requirements, you might need to either merge or split these data.

Think of this process as organizing and tidying up the data, removing what is no longer needed, replacing what is missing and standardising the format across all the data collected.

Skills Required

You will need scripting tools like Python or R to help you to scrub the data. Otherwise, you may use an open-sourced tool like OpenRefine or purchase enterprise software like SAS Enterprise Miner to help you ease through this process.

For handling bigger data sets require you are required to have skills in Hadoop, Map Reduce or Spark. These tools can help you scrub the data by scripting.

3. Explore Data.

Once your data is ready to be used, and right before you jump into AI and Machine Learning, you will have to examine the data.

Usually, in a corporate or business environment, your boss will just throw you a set of data and it is up to you to make sense of it. So it will be up to you to help them figure out the business question and transform them into a data science question.

To achieve that, we will need to explore the data. First of all, you will need to inspect the data and its properties. Different data types like numerical data, categorical data, ordinal and nominal data etc. require different treatments.

Then, the next step is to compute descriptive statistics to extract features and test significant variables. Testing significant variables often is done with correlation. For example, exploring the risk of someone getting high blood pressure in relations to their height and weight. Do note that some variables are correlated, but they do not always imply causation.

The term “Feature” used in Machine Learning or Modelling, is the data features that help us to identify the characteristics that represent the

data. For example, “Name”, “Age”, “Gender” are typical features of members or employees dataset.

Lastly, we will utilise data visualisation to help us to identify significant patterns and trends in our data. We can gain a better picture through simple charts like line charts or bar charts to help us to understand the importance of the data.

Skills Required.

If you are using Python, you will need to know how to use Numpy, Matplotlib, Pandas or Scipy; if you are using R, you will need to use GGplot2 or the data exploration swiss knife Dplyr. On top of that, you need to have knowledge and skills in inferential statistics and data visualization.

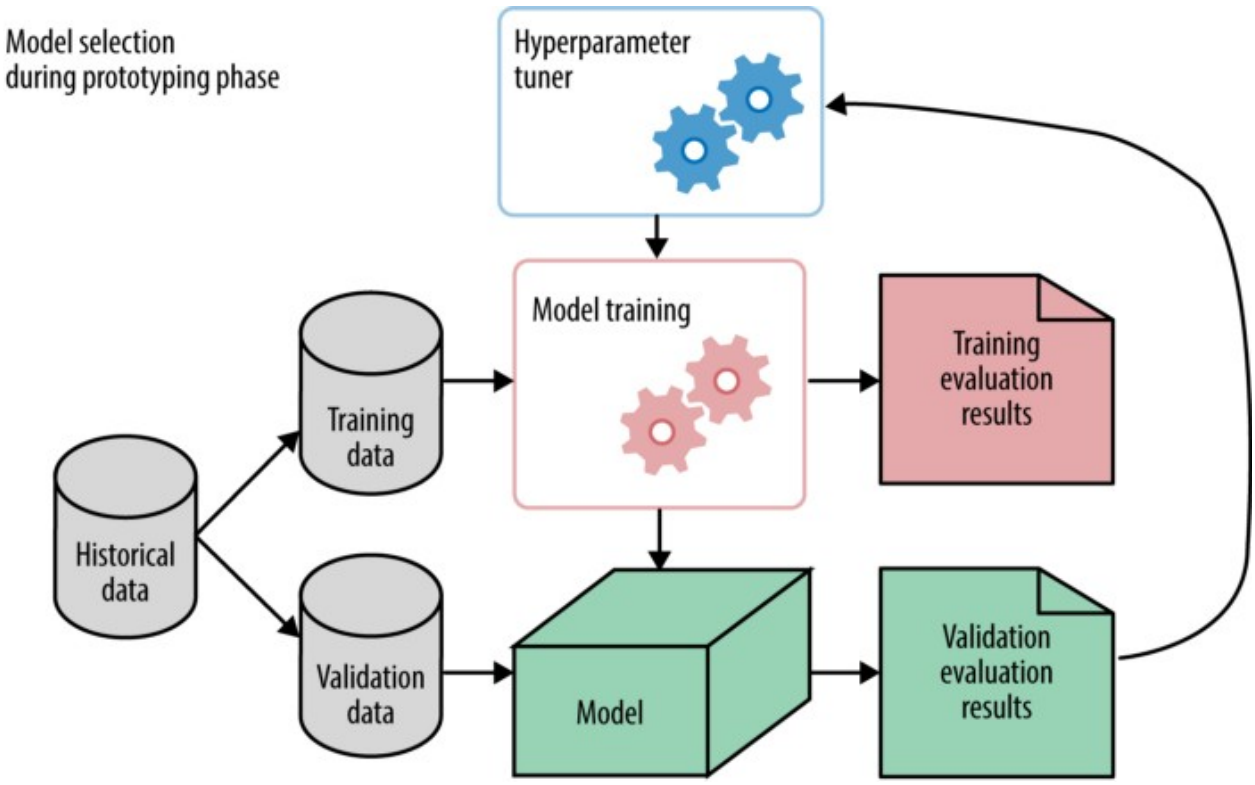
As much as you do not need a Masters or Ph.D. to do data science, these technical skills are crucial in order to conduct an experimental design, so you are able to reproduce the results.

Additional Tips:

- Be curious. This can help you develop your spidey senses to spot weird patterns and trends.
- Focus on your audience, and understand their background and lingo. So that you are able to present the data in a way that makes sense to them.

4. Model Data.

This is the stage where most people consider interesting. As many people call it “where the magic happens”.



Once again, before reaching this stage, bear in mind that the scrubbing and exploring stage are equally crucial to building useful models. So take your time on those stages instead of jumping right to this process.

One of the first things you need to do in modeling data is to reduce the dimensionality of your data set. Not all your features or values are essential to predicting your model. What you need to do is to select the relevant ones that contribute to the prediction of results.

There are a few tasks we can perform in modeling. We can also train models to perform classification to differentiating the emails you received as “Inbox” and “Spam” using logistic regressions. We can also forecast values using linear regressions. We can also use modeling to group data to understand the logic behind those clusters. For example, we group our e-commerce customers to understand their behavior on your website. This requires us to identify groups of data points with clustering algorithms like k-means or hierarchical clustering.

Skills Required.

In Machine Learning, the skills you will need is both supervised and unsupervised algorithms. For libraries, if you are using Python, you will need to know how to use Sci-kit Learn; and if you are using R, you will need to use CARET.

After the modeling process, you will need to be able to calculate evaluation scores such as precision, recall and F_1 score for classification. For regressions, you need to be familiar with R^2 to measure goodness-of-fit, and using error scores like MAE (Mean Average Error), or RMSE (Root Mean Square Error) to measure the distance between the predicted and observed data points.

5. Interpreting Data

We are at the final and most crucial step of a data science project, interpreting models and data. The predictive power of a model lies in its ability to generalize. How do we explain a model depends on its ability to generalize unseen future data.

Interpreting data refers to the presentation of your data to a non–technical layman. We deliver the results in to answer the business questions we asked when we first started the project, together with the actionable insights that we found through the data science process.

Actionable insight is a key outcome that we show how data science can bring about predictive analytics and later on prescriptive analytics. In which, we learn how to repeat a positive result, or prevent a negative outcome.

On top of that, you will need to visualize your findings accordingly, keeping it driven by your business questions. It is essential to present your findings in such a way that is useful to the organisation, or else it would be pointless to your stakeholders.

In this process, technical skills only are not sufficient. One essential skill you need is to be able to tell a clear and actionable story. If your presentation does not trigger actions in your audience, it means that your communication was not efficient. Remember that you will be presenting to an audience with no technical background, so the way you communicate the message is key.

Skills Required.

In this process, the key skills to have is beyond technical skills. You will need strong business domain knowledge to present your findings in a way that can answer the business questions you set out to answer, and translate them into actionable steps.

Apart from tools needed for data visualization like Matplotlib, ggplot, Seaborn, Tableau, d3js etc., you will need soft skills like presentation and communication skills, paired with a flair for reporting and writing skills will definitely help you in this stage of the project lifecycle.

Recap

If it is a brand new project, we usually spend about 60–70% of our time just on gathering and cleaning the data. Since it is a framework, you may use it as a guideline with your favorite tools.

The true north is always that business questions we defined, before even started the data science project. Always remember that solid business questions, clean and well-distributed data always beat fancy models.

Regars,

Data Science East Africa Team