Osemn framework

1. Obtain: obtain the data that we need from available data sources.

Types:

Structured-data -> mysql

Unstructured data -> mongodb

Big data -> spark

Flat files -> csv, tsv from  [Kaggle](https://www.kaggle.com/)

API

Python/r packages with datasets

1. Scrub data: clear/preprocess the data. organizing and tidying up the data, removing what is no longer needed, replacing what is missing and standardizing the format across all the data collected.

-60–70% of our time just on gathering and cleaning the data.

garbage in, garbage out philosophy: if the data is unfiltered and irrelevant, the results of the analysis will not mean anything.

-consolidate data

-extracting and replacing values

-split, merge and extract columns

Tools used: scripting tools like Python or R to help you to scrub the data.

1. 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. **Make sense of it**. figure out the business question and transform them into a data science question.

-**inspect the data and its properties**. Different data types like numerical data, categorical data, ordinal and nominal data etc. require different treatments.

-the next step is to compute descriptive statistics to extract features and test significant variables. Testing significant variables often is done with correlation.

-utilize data visualization 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.

1. Model data:

-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.

-reduce the dimensionality of your data set.  select the relevant ones that contribute to the prediction of results.

-perform classification to differentiating the emails you received as “Inbox” and “Spam” using logistic regressions. use modelling to group data to understand the logic behind those clusters(k-means or hierarchical clustering)

-After the modelling process, you will need to be able to calculate evaluation scores such as precision, recall and F1 score for classification. For regressions, you need to be familiar with R² 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.

-**The predictive power of a model lies in its ability to generalise**

1. Interpret data

-The predictive power of a model lies in its ability to generalise. How do we explain a model depends on its ability to generalise unseen future data.

-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

-you will need to visualise 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.

-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](https://seaborn.pydata.org/), [Tableau](https://www.tableau.com/), [d3js](https://d3js.org/)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.