Methodology

Here we are going to explain the methodology that can be applied to any data science project which has been outlined by John Rollins, a senior data scientist. The methodology has an objective and it is to answer to 10 questions to make sure the project is well lead.

From problem to approach

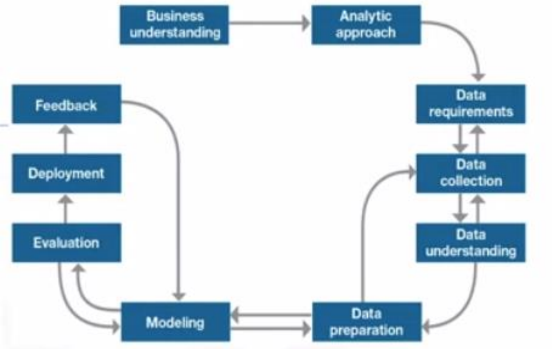
- Forming a concrete business or research problem
- Collecting and analysing data
- Building a model
- Understanding the feedback after model deployment

Working with the data

- Determine the data requirement
- Collecting appropriate data
- Understanding the data
- Preparing the data for modelling

Deriving the answer

- Evaluating and deploying the model
- Getting feedback on it
- Using that feedback constructively so as to improve the model



You have here the schema of the methodology, as you can see It is an iterative method that never ends. We are going to describe the steps here and explain each one of them.

The first step is business understanding, and it is placed at the beginning of the methodology, because getting clarity around the problem to be solved, allows you to determine which data will be used to answer the question.

The next step is the analytics approach and it means identifying what type of patterns will be needed to address the question most effectively. If the question is to determine probabilities of an action, then a predictive model might be used. If the question is to show relationships, a descriptive approach may be required, and if our problem requires counts, then statistical analysis is the best way to solve it.

- **Descriptive Analytics** tells you what happened in the past. Descriptive analytics helps a business understand how it is performing by providing context to help stakeholders interpret information. This can be in the form of data visualizations like graphs, charts, reports, and dashboards.
- **Diagnostic Analytics** helps you understand why something happened in the past.
- **Predictive Analytics** predicts what is most likely to happen in the future. Mathematical process that seeks to predict future events or outcomes by analysing patterns that are likely to forecast future results.

Let's now focus on the data part. We need to know the data requirement, that means which type of data we want, where we collect them and how we are going to work with them to get the result we want.

Then the data collection, in this phase the data requirements are revised, and decisions are made as to whether or not the collection requires more or less data. Once the data ingredients are collected, the data scientist will have a good understanding of what they will be working with. Techniques such as descriptive statistics and visualization can be applied to the data set, to assess the content, quality, and initial insights about the data.

The Data understanding step is here to make sure that the data scientist understands the content the quality and the insights in his dataset. To do so, we apply descriptive statistics on each variable such as mean, median, minimum etc. We also can search for a correlation between different variables. In this step we may need to revisit previous step.

Then the data preparation step is the most consuming phase in the data science project. It usually takes nearly seventy percent of the overall project time. His goal is to process the data in a state where it is easier to work with. In includes data cleaning, combining data from multiple sources, removes duplicates, ensuring everything is properly formatted…

Once all the data part is finished, we go into the modelling step. We use a training set to develop models that are either descriptive or predictive. The modelling process is highly iterative. An example of descriptive model would examine things like : if a person is doing this, then he is likely to prefer that. Whereas a predictive model tries to yield yes/no or stop/go outcomes.

Then, the data scientist evaluates the model's quality and check if it answers the initial question. There are two phases in this step, first ensure that the model is working as intended. The second phase is statistical significance testing, this can be applied to ensure that the data is being properly handled.

Once the data scientist is satisfying, he can deploy the model into the production environment or a comparable test environment.

By collecting results from the model, the organization gets feedback on the model's performance. Analysing the feedback enables the data scientist to refine the model, increasing the accuracy.