Lifecycle Of Data Science

THE OSEMN FRAMEWORK

Obtain

Obtain data from various sources.

- Receive data in applications like Microsoft Excel.
- Gather by connecting Web APIs to Web servers to extract data – Ex: Facebook, Google.
- Obtain data directly from text files, CSV (Comma Separated Values) or TSV (Tab Separated Values)
- Obtain data to scrape from the websites using web scraping tools such as Beautiful Soup.
- Use MSSQL, MongoDB Python, R, PostgreSQL to read and process the data using Data Science programs.
- Apache Hadoop, Spark or Flink tools can be used for bigger data sets.

Explore

Examine the data.

- Understanding the business equation and transforming to data science questions is vital.
- The first step is to inspect the data and its properties.
- The next step is to compute descriptive statistics to extract features and test significant variables.
- The term "Feature," which is used in Machine Learning or Modeling, refers to data features that help us identify characteristics that represent the data. For example, "Name," "Age," and "Gender" are typical features of an employee's dataset.
- Utilize data visualization to help identify significant patterns and trends in data – Ex: Tableau, bar charts, pie charts.
- Python → Numpy, Matplotlib, pandas or SciPy (Data Exploration)
- R → ggplot2, Dplyr (Data Exploration)

ADDITIONAL TIPS:

- ☑ Be curious. This can help you develop "spidey senses" to spot patterns and trends.
- ☑ Know your audience. Understand their background and lingo so you can present the data in a way that makes sense to them.

Scrub

Clean and Filter Data:

- Scrubbing is the process for 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.

STEPS OFTEN INCLUDE:

- ☑ Converting the data from one format to another and consolidating everything into one standardized format across all data.
- ☑ Scrubbing data also includes the task of extracting and replacing values.
- ☑ Data must be split, merged, and extracted into columns.
- ☑ Scrubbing tools include Python, R, OpenRefine, SAS Enterprise Miner.

Model

"Where the magic happens."

- Generating a model is easy. Generating a useful model is hard.
- The first step in modeling data is to reduce the scope of your data set and select the relevant data elements that contribute to the prediction of results.
- Modeling tasks include the "Train" model for classification, forecasting using linear regressions, and grouping data using clustering algorithms like k-means or hierarchical clustering.
- Use regression and predictions for forecasting future values, and classification to identify, and clustering to group values.

Interpret

"Interpreting the N in OSEMN"

- Present data with actionable insights.