Data Science Pipeline

Data Science | CCDATSCL

OSEMI Pipeline

O — Obtaining our data
S — Scrubbing / Cleaning our data
E — Exploring / Visualizing our data will allow us to find patterns and trends
M — Modeling our data will give us our predictive power as a wizard
I — Interpreting our data

Research/Business Question

So before we even begin the **OSEMN** pipeline, the most crucial and important step that we must take into consideration is understanding what **problem** we're trying to solve.
Before we even begin doing anything with "Data Science", we must first take into consideration what **problem** we're trying to solve.

- How can we translate **data** into **dollars**?
- What **impact** do I want to make with this data?
- What **business value** does our model bring to the table?
- What will **save** us lots of **money**?
- What can be done to make our **business** run more **efficiently**?

Knowing this **fundamental concept** will bring you far and lead you to greater steps in being successful towards being a **Data Scientist**
No matter how well your model predicts, no matter how much data you acquire, and no matter how OSEMN your pipeline is, **your solution or insight will only be as good as the problem you set for yourself.**

Obtain your Data

As a rule of thumb, there are some things you must take into consideration when obtaining your data.
You must identify all of your **available datasets** (from the internet or internal/external databases)
You must extract the data into a **usable format** (.csv, json, xml, etc..)

Skills Required

- Database Management: MySQL, MongoDB
- Querying Relational Databases
- Retrieving Unstructured Data: text, videos, audio files, documents
- Distributed Storage: Hadoop, Apache Spark/Flink

Scrubbing / Cleaning Your Data

This phase of the pipeline should require the most time and effort.
The results and output of your machine learning model is only as good as what you put into it. Basically, garbage in garbage out.

Examine the data

- Understand every feature you're working with, identify errors, missing values, and corrupt records

Clean the data

- Throw away, replace, and/or fill missing values/errors

Exploring (Exploratory Data Analysis)

During the exploration phase, we try to **understand** what patterns and values our data has.
This is where we use different types of **visualizations** and **statistical testings** to back up our findings.
This is where we will be able to derive **hidden meanings** behind our data through various graphs and analysis.

Objective

- Find patterns in your data through visualizations and charts
- Extract features by using statistics to identify and test significant variables

Skills Required

- Numpy, Matplotlib, Pandas, Scipy (for Python)
- GGplot2, Dplyr (for R)
- Data Visualization
- Inferential statistics

Modeling (Machine Learning)

Think of a machine learning model as tools in your **toolbox**.

You will have access to many algorithms and use them to **accomplish different business goals**. The better features you use the better your predictive power will be

After cleaning your data and finding what features are most important, using your model as a predictive tool will only enhance your business **decision making**.

Predictive Analytics is emerging as a game-changer. Instead of looking backward to analyze "what happened?" Predictive analytics help executives answer "What's next?" and "What should we do about it?" (Forbes Magazine, April 1, 2010)

One great example can be seen in Walmart's supply chain. Walmart was able to predict that they would sell out all of their Strawberry Pop-tarts during the hurricane season in one of their store location. Through data mining, their historical data showed that the most popular item sold before the event of a hurricane was Pop-tarts.

Objective:

- In-depth Analytics: create predictive models/algorithms
- Evaluate and refine the model

Skills Required

- MachineLearning: Supervised/Unsupervised algorithms
- Evaluation methods
- Machine Learning Libraries: Python (Sci-kit Learn, Tensorflow, Pytorch)
- Linear algebra & Multivariate Calculus

Interpreting (Data Storytelling)

The most important step in the pipeline is to understand and learn how to explain your findings through communication.

It's about connecting with people, persuading them, and helping them.

The art of understanding your audience and connecting with them is one of the best part of data storytelling.

Emotion plays a big role in data storytelling.

People **are not going to magically understand** your findings.

The best way to make an impact is **telling your story through emotion**. We as humans are naturally influenced by emotions.

If you can **tap into your audiences' emotions**, then you are in control.

When you're presenting your data, keep in mind the power of psychology.

The art of understanding your audience and connecting with them is one of the best part of data storytelling.

Objective

- Identify business insights: return back to business problem
- Visualize your findings accordingly: keep it simple and priority driven
- Tell a clear and actionable story: effectively communicate to non-technical audience

Skills Required

- Business Domain Knowledge
- Data Visualization Tools: Tablaeu, D3.JS, Matplotlib, GGplot, Seaborn
- Communication: Presenting/Speaking & Reporting/Writing