datamover.ai

The Machine Learning workflow

0️⃣a) Understand the Problem

1. Look at the big picture and study design;
2. Define business objective;
3. Check existing solutions/workarounds (if any)

0️⃣b) Define Analytical needs

1. Frame the problem statement mathematically (supervised/unsupervised, online/offline, regression/classification etc.);
2. Select performance measure [F1-score, AUC, RMSE, MAE, etc.]
    1. Is the performance measure aligned with the business objective?
    2. What minimum performance would be needed to reach the business objective?
    3. What are similar problems? Can we reuse experience or tools?
3. How would we solve the problem manually?
4. List assumptions coming from research questions made so far.
5. Verify assumptions (if possible).

1️⃣Data Preparation

1. Fetch dataset;
2. Check dataset size and ensure your workspace has enough storage if you are dealing with big datasets;
3. Check the data type (time series, sample, geographical, etc.) and make sure they are what they should be.
4. If necessary, convert the data to a format that is easy to manipulate (without changing the data itself, e.g. .csv, .json).

5. For training of ML models, sample a hold-out set, put it aside, and **never look at it** ⚠️.
    - typical train/test splits are 60/40, 70/30, and 80/20;
    - it is convenient to store train and test data separately;
    - **Note:** often *test set* and *hold-out* are used interchangeably.

Note: automate scripts as much as possible for future data API calls.