The Machine Learning Workflow

🗓️ Last update 16 giu 2023

Table of content

0a) Understand the problem

0b) Define Analytical Needs

1 Data Preparation

2 Exploratory Data Analysis (EDA)

3 Statistical Inference & Feature Engineering

4 Cluster Analysis

5 Create a Baseline Model

6 Train a Machine Learning Model

7 Fairness and Explainability Analysis

8 Deployment and Monitoring

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).

1️⃣ Data Preparation

1. Fetch dataset;

   💁‍♂️ Check this article to learn how to load dataset OS agnostic.

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;
   - 📝 often, *test set* and *hold-out* are used interchangeably.
6. ⬇️ Store train and test locally
   - Store both datasets in data folder in csv format;
   - Save train and test set as data_train.csv and data_set.csv, respectively.
   - In both datasets, retain the column names and discard the index if it is not informative.

💁‍♂️ Automate scripts as much as possible for future data analysis.

2️⃣ Exploratory Data Analysis (EDA)

1. Load the train set and sample the dataset to a manageable size if necessary;
2. For supervised learning tasks, identify the target attribute(s);
3. Study each attribute and its characteristics, namely:
    a. Name
    b. For tabular data, define the data type of each variable, namely:
        i. Nominal: Named categories, e.g., gender : ['Female', 'Male']
        ii. Ordinal: Categories with an implied order, e.g. quality : [Low, Medium, High]
        iii. Discrete: Only particular numbers, e.g., age: {1, 2, ..., 59, 60}
        iv. Continuous: Any numerical value, e.g. weight: {38.9, ..., 45.5}

    📝 Nominal and ordinal data types are considered **categorical** (qualitative) features, whereas discrete and continuous data types are considered **numerical** (quantitative) features.

    c. Percentage of missing values, namely np.NaN
        i. missingno can be a useful tool for visualisation;
        ii. Ensure missing values are not encoded in specific ways, e.g. -1, "?".
        iii. Inspect rows with missing values to assess if a specific pattern exists.
    d. Check if there are any duplicates and inspect them;
    e. Noisiness and type of noise, e.g. stochastic, rounding errors, etc. (might require business knowledge);
    f. The frequency of each group within each categorical variable and the type of distribution for numerical variables (refer to this link for common types of distributions). It is recommended to visualise each variable by using:
        i. a countplot for categorical variables;
        ii. a histplot for numerical variables;
    g. Examine possible outliers in numerical variables and check whether they make sense (might require business knowledge). For details on identifying outliers, refer to this link.

4. Annotate all information from EDA, such as:
    a. the type of data;
    b. if there are missing values and how to deal with them;
    c. summary statistics of both numerical and categorical variables;
    d. the type of distribution;
    e. identify the promising transformations you may want to apply (e.g. log-transformation for highly skewed distribution or cluster facets to mitigate group imbalance);
    f. identify additional data sources that would be useful;
    g. anything else that is noteworthy for model training.