A-Z Machine Learning Workflow

CHECKLIST

1. Analysis of the Problem

- ❑ Define problem
- ❑ Define project goal
- ❑ Frame the problem
- ❑ Define performance metric
- ❑ Look for approaches to similar problems
- ❑ List assumptions made so far

2. Data Retrieval

- ❑ Find data and document where it has been obtained from

3. Exploratory Analysis

- ❑ Check size and type of data
- ❑ **Sample a test set and don't snoop**
- ❑ Create copy ± downsample
- ❑ Study of each variable
  - ❑ Type, NaN, outliers, distribution...
- ❑ Visualize data
- ❑ **Select target attribute** (supervised)
- ❑ Study correlations
- ❑ Identify possible transformations
- ❑ **Document every step**
- ❑ Get more data if necessary

4. Data Wrangling

- ❑ Work on a copy of the dataframe
- ❑ Use functions for data transformation
  - ❑ Pandas or sklearn
- ❑ Clean null values
- ❑ Clean outliers
- ❑ Feature selection
- ❑ Feature engineering
  - ❑ Discretizing, encoding, decomposing, transforming, aggregating, etc.
- ❑ **Feature scaling**

5. Choosing a Model

- ❑ Choose a few different models to train using "standard" parameters
- ❑ Measure and compare their performance
- ❑ Analyze most significant attributes for each algorithm
- ❑ Measure and compare performance
- ❑ Quick round of feature selection and/or engineering
- ❑ Iterate a few times over these steps
- ❑ **Shortlist the top 3-5 most promising models**

6. Fine Tuning

- ❑ Use all training data
- ❑ RandomizedSearchCV or GridSearchCV
- ❑ Ensemble best models
- ❑ **Only when you are happy with a model: measure performance on test set**
- ❑ Do not modify model after measuring performance on test set
  - ❑ Risk to overfit it

7. Presenting

- ❑ Document every step of your process
- ❑ Highlight the big picture
- ❑ Interesting points you noticed along
- ❑ Make sure your key findings are communicated
- ❑ Use beautiful visualizations
- ❑ Use easy to remember statements

Daniel Eiroa
TA | DAPT BCN DEC19

Based on and modified from:

Hands-on ML with Scikit-learn, Keras and TensorFlow, 2nd edition, by Aurélien Géron (O'Reilly). Copyright 2019 Kiwisoft S.A.S, 978-1-492-03264-9.