Project Workflow

DATA 301

Generic workflow

1. What do you want to do? predict? Cluster? Something else?
2. Get data proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if its too big? Or too small?
3. Preprocess the Data you will spend a lot of time here, clean NaN's, normalize fields, ensure strings are consistent (VA,Va, Virginia etc..), encode ordinal data, etc...
4. Explore the Data plot it, are there anomalies? can you see patterns? How to plot if you have >3 features?
5. Model Data build, fit and validate a model
6. Evaluate Model not good enough? Go to step 2, consider ensembling multiple models
7. Communicate and visualize results

See https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations

Critical Bits

1. Domain Expertise is essential. For instance:
   - When evaluating medical images for tumors, it helps if you can recognize a tumor.
   - When exploring clustered botanical data its useful to be able to verify that groupings make sense.
   - When checking engine sensor data, it helps if you have some idea of what typical running parameters are and how anomalous behavior presents.
2. Communication and presentation skills are the most important of all.
   - If you cannot convince stakeholders to follow you, then you have wasted your time

See https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations