A Beginner's Guide to The Machine Learning Workflow

datacamp

1 Project setup

1. Understand the business goals

Speak with your stakeholders and deeply understand the business goal behind the model being proposed. A deep understanding of your business goals will help you scope the necessary technical solution, data sources to be collected, how to evaluate model performance, and more.

2. Choose the solution to your problem

Once you have a deep understanding of your problem—focus on which category of models drives the highest impact. See this **Machine Learning Cheat Sheet** for more information.

2 Data preparation

1. Data collection

Collect all the data you need for your models, whether from your own organization, public or paid sources.

2. Data cleaning

Turn the messy raw data into clean, tidy data ready for analysis. Check out this **data cleaning checklist** for a primer on data cleaning.

3. Feature engineering

Manipulate the datasets to create variables (features) that improve your model's prediction accuracy. Create the same features in both the training set and the testing set.

4. Split the data

Randomly divide the records in the dataset into a training set and a testing set. For a more reliable assessment of model performance, generate multiple training and testing sets using cross-validation.

3 Modeling

1. Hyperparameter tuning

For each model, use hyperparameter tuning techniques to improve model performance.

2. Train your models

Fit each model to the training set.

3. Make predictions

Make predictions on the testing set.

4. Assess model performance

For each model, calculate performance metrics on the testing set such as accuracy, recall and precision.

4 Deployment

1. Deploy the model

Embed the model you chose in dashboards, applications, or wherever you need it.

2. Monitor model performance

Regularly test the performance of your model as your data changes to avoid model drift.

3. Improve your model

Continously iterate and improve your model post-deployment. Replace your model with an updated version to improve performance.