



Machine Learning Project Checklist

Written by Dr. Jody-Ann S. Jones

▼ 💡 Frame the Problem

- ☐ What is the context?
- ☐ What is the research/business problem?
- ☐ What is your response/answer to the research/business problem?
- ☐ Describe the data (where is it coming from, what is it about, etc.?)

☐ Import Requisite Libraries

☐ Load and Check Dataset

▼ Exploratory Data Analysis (EDA)

Explore the following:

- ☐ Summary statistics
- ☐ Cardinality of categorical variables

- ☐ Distributions of numerical variables
- ☐ Identify Outliers
- ☐ Examine bivariate relationships between key features and the target
- ☐ Correlation among key variables

▼ Data Preprocessing & Feature Engineering

- ☐ Check for and treat duplicate values
- ☐ Check for and treat missing values (what to do with them? Drop them, impute them, etc.)
- ☐ Scale numerical features (Standardization, Normalization, etc.)
- ☐ Encode categorical variables (transforming categories into dummies, i.e. 1s and 0s or some other type of label encoding).
- ☐ **Assign X & y variables (if working on a supervised learning problem; otherwise, there is no y)**
- ☐ **Split Data into Training and Test Sets (if running cross-validation, you may need a validation set)**

▼ Choose and Train an Estimator (Algorithm)


- ☐ Import respective class
- ☐ Instantiate an object of said class
- ☐ Train (or Fit) Model

▼ Evaluate Model (a few options here, use one or more)

- ☐ Apply .transform method on values in the test set (if necessary)
- ☐ Predict values on the test set
- ☐ Respective estimator's score method
- ☐ Cross-Validation
- ☐ ~~Respective Metrics~~

▼ Fine Tune the Model (Model Optimization)

- ☐ Setup a Pipeline

- ☐ RandomizedSearchCV (for large datasets)
- ☐ GridSearchCV (for small to medium-sized datasets)
- ☐  **Deploy or Present Model**
- ☐ Test and Monitor Model in Production