

(*a*) exploratory data analysis is used to investigate the data.

(*b*) indicates where early data analysis might take place - evaluating simple summary measures or identifying predictors that have strong correlations with the outcome. Iterate between visualization and analysis until the modeler feels confident that the data are well understood.

(*c*) the first draft for how the predictors will be represented in the models is created based on the previous analysis.

(*d*) several different modeling methods might be evaluated with the initial feature set followed by *hyperparameters*.

(*e*) Models are numerically evaluated on the data to understand their performance characteristics. Summary measures for each model, such as model accuracy, are used to understand the level of difficulty for the problem and to determine which models appear to best suit the data.

(*f*) Based on these results, more EDA can be conducted on the model results, such as residual analysis. The properties that are poorly predicted can be examined to understand if there are any systematic issues with the model. \

(*g*) Another round of feature engineering might be used to compensate for these obstacles.

(*h*) By this point, it may be apparent which models tend to work best for the problem at hand and another, more extensive, round of model tuning can be conducted on fewer models. After more tuning and modification of the predictor representation, the two candidate models (#2 and #4) have been finalized.

(*i*) These models can be evaluated on an external test set as a final “bake off” between the models.

(*j*) The final model is then chosen and this fitted model will be used going forward to predict new samples or to make inferences.