**Machine Learning Experiment Checklist**  **2023-02-10**

* **Data Review**
  + The data sources are identified.
  + Does the output variable depend on time?
  + Do the records belong to groups (e.g., multiple medical test results or images per patient, with each result or image represented by a single record) or are they independent?
  + Were the data checked for class imbalance?
* **Data Cleaning**
  + Each row corresponds to a single sample and vice versa
  + Each variable has its own column
    - Variables are sometimes combined. For example, a column describing car transmissions may have string values like “6-cyl, RWD, conventional” and “4-cyl, AWD, hybrid”. The column should be separated into three separate variables, each with their own column.
  + All of values in each column are formatted consistently (e.g., state names are all spelled and capitalized in a consistent manner, dates are all formatted as YY-MM-DD)
  + All values in each column have the same types (e.g., categorical, int, float)
  + All values in each column have the same units (e.g., inches, USD)
  + Any missing values are identified, handled appropriately, and documented
  + Any outlier or unexpected values or records are identified, handled appropriately, and documented
* **Problem Definition**
  + Output variable and type (e.g., continuous, categorical, discrete counts, etc.) are stated
  + Any constraints on the output variable are stated (e.g., values must be positive or within a specific range)
  + Problem type (e.g., regression, classification, other) is identified
  + Does the customer require that the models achieve a minimum performance threshold?
  + Are there any constraints around the model training and deployment (e.g., available memory or compute power, infrastructure costs, maximum run time for inference, programming language)?
* **Experimental Setup**
  + Is it documented what data will be available at training and inference time, respectively?
  + Will you use a train-test split, train-test-validation split, or cross-fold validation?
  + Is the splitting consistent with what data will be available at training and prediction time in the real-world usage scenario?
  + Are the samples stratified by class label (classification) or output value (regression)?
  + If the problem is time dependent, are splits performed along the time dimension? Will all previous windows or only the previous window used for training?
  + If the records belong to groups (e.g., multiple images per patient), are entire groups assigned to the same split or window (e.g., records from the same group should not be in different splits)?
* **Feature Engineering and Transformation**
  + Are continuous variables scaled (mean of 0, standard deviation of 1)?
  + Are missing values imputed?
  + Are categorical variables one-hot encoded?
  + If transformations (e.g., standard scaler, text vectorization, PCA / SVD, etc.) are stateful, are the parameters fit using only the training set?
* **Model Training**
  + Did you pick models that appropriate for the problem type and constraints on the output value?
  + Did you pick models that can operate within the given resource requirements?
  + Were the models trained on training set (and not the test set)?
  + If the data are imbalanced, was the training set rebalanced before the model was trained?
  + Did you tune any hyper-parameters (e.g., regularization type and strength, number of iterators, learning rate, etc.) use cross-fold validation on the training set or a validation set if doing a three-way split?
* **Model Selection and Evaluation**
  + Are the chosen metrics appropriate for the problem type?
  + If there is class imbalance, are the chosen metrics robust to it?
  + Are the metrics interpreted correctly?
  + Are the predicted values within the expected range (e.g., no negative predictions when the output variable can only be positive)?
  + Is there evidence of bias or outlier predictions (e.g., does a plot of the fitted values versus the residuals indicate large errors for some records?)?
  + If required, was variable selection or regularization performed to simply the model?