Two main components:

* Data preprocessor
  + Inputs: data, labels, instructions for each column
  + Splits data into categorical / continuous data
  + Impute values or leave out values based on the instructions from user
  + select features based on feature importance
* Model builder
  + Inputs: data from first step, user model preference
  + Take the data and model type and train on that.
  + Perform cross validation with stratified, unstratified, other options
  + Perform feature selection and hyperparameter tuning for specified model
  + Build multiple models based on user input
* Output from model
  + This will take the constructed models and give outputs for test data or just return the pretrained models so that they can be used immediately
  + Maybe even do some model selection to get best performing model
  + The data used to train the model will also be provided
  + Output a trained model that can be used to predict batches right away

Data Descriptions:

* Features: feature names
* Continuous / Categorical: tell whether data is continuous or categorical
  + CONT – continuous
  + CAT – categorical
  + TEXT – convert text data TODO later
* Encoding – how to encode categorical data
  + MAP – map to unique numbers
  + OHE – one hot encoding
  + LOO – leave one out counting
  + Bin - Binary: first the categories are encoded as ordinal, then those integers are converted into binary code, then the digits from that binary string are split into separate columns.  This encodes the data in fewer dimensions that one-hot, but with some distortion of the distances.
  + LVL – Level encoding. For each possible value, threshold levels and assign anything less to be a 0 and 1 to be a 1.
  + FWD – forward difference – 1 for current class, -1 for lesser class
  + BWD – backward difference – 1 for current class, -1 for next class
  + NONE – don’t do anything
* Scaling – specify how to scale continuous features
  + SCL0 – scale between (0, 1) - (x – min) / (max – min)
  + SCL1 – scale between (-1, 1) –
  + STD – standardize – (x – xbar) / sigma
  + NRM1 – scale by dividing by 1 norm
  + NRM2 – scale by dividing by 2 norm
  + NRMI – scale by dividing by Inf norm
  + NONE – don’t do anything
* Missing Values – continuous features
  + DEL – delete entries with a missing value
  + MEAN – use the mean of all values
  + MED – use the median of all values
  + MODE – use the most frequent value
  + NONE – don’t do anything
  + MIN – add the minimum value
  + MAX – add the maximum value
  + UNIQ – use a uniq value for all missing entries
  + Look at some other python libraries for imputation