Pipeline Steps:

1. Data Gathering
   * Can come from:
     1. Within the business itself
     2. Bought from a third party
     3. Publicly available sources
2. Data Analysis
   * Trying to understand the data
   * To know how the variables are related to each other and to what we want to predict
   * We have to know which variables we can use (as there may be regulations in your business)
3. Data Pre-processing
   * Missing values, encoding categorical variables, becoming aware (or dealing) with class imbalance, checking numerical features distribution, outlier detection (as some algorithms may perform poorly on those) feature scaling…
   * Rare labels: Labels that may only appear on the test or on the training data, we should take care of that somehow (in the future lessons)
   * If our distributions are not Gaussian, we may apply some transformations to make it Guassian (mean 0 and unit variance) by applying feature scaler
   * Outliers may cause overffiting in boosting algorithms
4. Dimensionality Reduction
   * Feature Selection and/or
   * Feature Extraction (non-interpretable results)
   * Less lines of code for error handling:
     1. Example: Lines to handle rare labels
   * Wrapper method: creates a model using for each combinations of subset of features
   * Wrapper method is “not model agnostic” because the subset of features that it finds for one model, may not be the best for another model
   * Embedded method is “not model agnostic” because the features that a random forest selects may not be the best for a linear model
   * In this course we are not going to use an embedded method, we will rather come up with a list of features that are the most important and make that list of features part of the pipeline
5. Machine Learning Model building
   * Try to come up with the best model for your problem
6. Model building Business uplift evaluation
   * How the evaluation of our models is related to the improvement in the business value
   * For example, if we were building a model for fraud, we need to evaluate how much we saved by using the model that we created
   * For example, if we were dealing with advertisement, how much more value the customers would bring if we used our model
7. Model deployment (actually, a data and machine learning pipeline deployment)
   * Meaning: Putting the model in the cloud (or other system) where it can be then accessed by the other systems in our business to get the model’s predictions
   * In deployment we need to deploy steps 3, 4, 5, which are Data Preprocessing, Dimensionality Reduction and Machine Learning Model building respectively

Meta assembling is using many models to create a new model (called meta model) which is then used to combine the results of all the models