Decision Trees

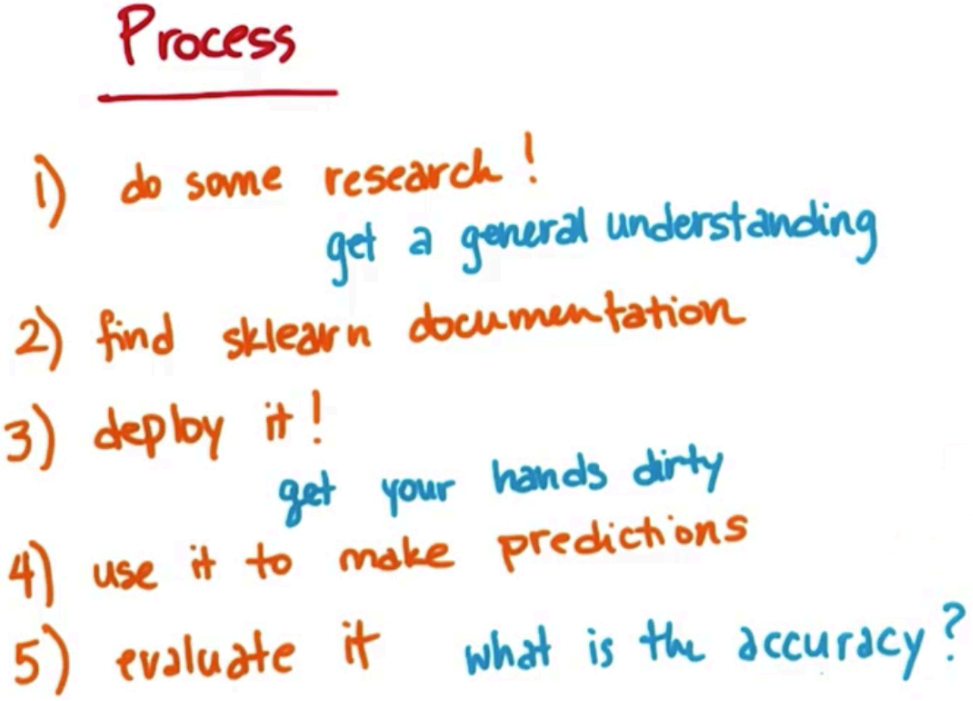
Ex1

- Your first email DT: accuracy
- The number of features for a smaller training set (in *email_preprocess.py*)
  selector = SelectPercentile(f_classif, percentile=10)
- Accuracy using 1% of features

A New Algorithm

Ex2

- K nearest neighbors
- Random forest
- Adaboost (also called boosted decision tree)



Ex3

- Size Of The Enron Dataset
- Features In The Enron Dataset
- Finding POIs In The Enron Data
- How Many POIs Exist?
- What is the total value of the stock belonging to James Prentice?
- Of these three individuals (Lay, Skilling and Fastow), who took home the most money (largest value of “total_payments” feature)?
- How many folks in this dataset have a quantified salary? What about a known email address?
- How many people in the E+F dataset (as it currently exists) have “NaN” for their total payments? What percentage of people in the dataset as a whole is this?

Ex4

Bonus Target And Features

- Extracting Slope And Intercept
- Regression Score: Training Data
- Regression Score: Test Data
- Regressing Bonus Against LTI
- Salary Vs. LTI For Predicting Bonus
- Outliers Break Regressions