DAT13 SF: HOMEWORK 3 ASSIGNMENT

Assigned: As of March 28, this is being assigned as an optional exercise. It may become an "official" homework assignment after we learn more about trees.
Due: TBD
Submission Method: Please push completed homework assignments to your personal homework repo on Github and submit the URL via the Google Form: http://goo.gl/forms/QBZBG4P3bm

The purpose of this homework is to gain hands-on experience with decision trees.

DATA

For this assignment we will use the Bank Marketing dataset. The dataset(s) that you need for this exercise are included in the github repo with this assignment:

- **bank.csv**: This file is a 10% sample of the total dataset and includes 17 features
- **bank-additional-full.csv**: This file includes the full 41,188 prospect records and includes 20 features

The original source of the dataset is the UCI Machine Learning Repository:

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

This data is from direct marketing campaigns of a Portuguese banking institution. The goal of this direct marketing campaign was to sell bank term deposits (CDs, or certificates of deposit in the US). Often, the prospective client was contacted multiple times. If the prospect purchased the product, then that prospect was labeled "yes". If the prospect did not purchase the product, the that prospect was labeled "no".

The classification goal is to predict if the client will purchase (yes/no) a bank term deposit. In the dataset(s), the class label column header is simply "y".

HOMEWORK QUESTIONS

Answer these question in an iPython notebook. Show your code.

1. Use the file bank.csv to explore the dataset. Observe the features: Are they numbers? Are they strings? Are they binary? Are they continuous?

2. Learn about label encoders at the following link and use what you learn to transform the features to numerical features.

3. Build a decision tree model to predict whether a prospect will buy the product.

4. Evaluate the accuracy of your decision tree model using cross validation.

5. Repeat the analysis and cross validation with the file bank-additional-full.csv. How does the performance of the model change (with the additional training examples and additional features)?