DAT10 SF: HOMEWORK 6 ASSIGNMENT

Assigned: Wednesday, November 12th
Due: Monday, November 17th, midnight
Review due: Wednesday, November 19th.

The purpose of this homework is to gain deeper understanding or decision trees and random forests, as well as learn about online MLaaS.

DATA & CONTEXT

For this assignment we will use the Bank Marketing dataset that you can find here:

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

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

There are four datasets:
1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010
2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).
The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

HOMEWORK QUESTIONS

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

2. Learn about label encoders here: http://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features and transform the features to numerical features

3. Build a simple decision tree model to predict the classification goal.

4. Evaluate the result of the classification with cross-validation.

5. Extend the analysis and cross-validation to bank-additional-full.csv. How does the performance change?

6. Improve your model by using an ensemble method (RandomForest or ExtraTrees). How does the cross-validation performance improve?

7. Optional: read about learning curves here: http://scikit-learn.org/stable/auto_examples/plot_learning_curve.html and plot the learning curves for your best model.

8. Optional: register on BigML.com, upload the dataset and run a model. How does the result compare with your result?