# 1.) Observations on the data

1. AGE: number and continuous.
2. JOB: string - bunch of options
3. MARITAL: string - 3 options (single, married, divorced)
4. EDUCATION: string - 4 options (primary, secondary, tertiary, unknown)
5. DEFAULT: string - binary YES/NO
6. BALANCE: number
7. HOUSING: string - binary YES/NO
8. LOAN: string - binary YES/NO
9. CONTACT: - string - 3 options (cellular, telephone, unknown)
10. DAY: number
11. MONTH: string - 12 options - continuous
12. DURATION: number
13. CAMPAIGN: number
14. PDAYS: number
15. PREVIOUS: number
16. POUTCOME: string - 4 options (success, failure, unknown, other)
17. Y: string - binary YES/NO

There weren't any number columns that had missing data such that column values at to be converted based on inspection of the data frame with "info". The non-binary string columns are categorical features which are not considered continuous, therefore, they need to be converted into binary fields.

Except for "month", the other categorical strings don't seem continuous. Used the sk-learn.preprocessing.OneHotEncoder to change categorical data into individual binary columns.

# 2.) Evaluating the cross-validation scores

For bank-full.csv:

```
Decision Tree CV Score: 0.859038639379
Random Forest CV Score: 0.895622722249
```

For bank-additional-full.csv:

```
Decision Tree CV Score: 0.889579574205
Random Forest CV Score: 0.908759994817
```

The "additional" set had 3 extra features with a few slightly different features like the "education" feature had different labels and with more specificity. The result appeared that the RandomForest score between the "bank-full" and "bank-additional-full" wasn't extremely better but the decision tree score was much better for the "bank-additional-full" set. Perhaps suggesting that the extra fields allowed for better individual decision trees.