Homework III

You can submit in groups of 2!

All assignments need to be submitted via github classroom, assignment
https://classroom.github.com/assignment-invitations/964f09ce4ee8d2d754a0e45b6b4ed637
The homework is due 04/03/17.

The repository will contain the following files:

- data.csv : the actual data to be used
- hw3_starter_notebook.ipynb : the notebook to be used for providing your final analysis
- Data_dictionary.txt : file containing feature descriptions

You are required to update the “hw3_starter_notebook.ipynb” notebook. It should contain your entire code in the steps defined in the notebook. Feel free to add comments and visualizations in the same notebook. Please DO NOT submit any other document apart from the notebook for grading as it will not be considered.

Use travis to run your notebook and add an “assert” statement inside the notebook that ensures the outcome is the actual error you report (Say, if you claim your algorithm is 90% accurate, add an assert statement that score returned by your model is at least 90%). Travis will simply run the notebook and check that it ran without errors.

We recommend that you fork the homework repository and run Travis on your own repository - this way you don’t have to wait for other students submissions to finish on travis.

Task

A banking institution ran a direct marketing campaign based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be subscribed or not. Your task is to predict whether someone will subscribe to the term deposit or not based on the given information.

Your analysis should include the following sections [as specified in the started notebook]

- **Data Cleaning [10 points]**
    - This involves a first level look into the data and some standard pre-processing independent of the model to be used.
- **Model Set1 [35 points]**
    - In section you are supposed to test models apart from tree-based models, like SVM, Logistic Regression, etc.

    - You should perform the necessary feature engineer, model validation, feature selection and model selection for all models you choose from set 1
    - You can select a max of 5 models from this set for final ensemble
- **Model Set2 [35 points]**
    - In section you are supposed to test tree-based models, like decision tree, random forest, gradient boosted trees, etc.
    - You should perform the necessary feature engineer, model validation, feature selection and model selection for all models you choose from set 2.
    - You can select a max of 5 models from this set for final ensemble
- **Model Ensemble [20 points + 10 Bonus points]**
    - In this step, we expect you to use the models created before and create new predictions.
    - You should try model averaging and "poor man's stacking" but we encourage you to think of different ensemble techniques as well. We will judge your creativity and improvement in model performance using ensemble models and you can potentially earn 10 bonus points here.
- **Bonus: Resampling Techniques [up to 30 Extra Credit]**
    - Evaluate different resampling techniques and the "Easy Ensemble" technique.

Feel free to include your thought process as comments in the notebook.

Report the test error using area under the ROC curve. You can try other metrics as you deem appropriate, but reporting ROC AUC is required.

Important Notes:

- **Benchmark scores** have been uploaded on Kaggle leaderboard. These are the most basic models. If you're getting less than that, you're probably doing something wrong. Your aim is to improve on these benchmarks as much as possible.
- Remember that not all features provided to you should be used for the model. You should think about which of them are available before making the promotional call. For instance, **"duration" feature should definitely NOT be used** as we don't know the duration of the call before the call has been made.
- You can only make **5 submissions per day** on Kaggle. So plan accordingly.