Aunalytics Data Science Exercise

Objective:
Demonstrate your ability to solve a data science problem based on the information available and by making reasonable assumption. The output work should exhibit machine learning, feature engineering, statistics and visualization skills. The use of different types of modeling approaches is encouraged

Instructions:

- Solve the following problem statement using jupyter notebooks with python kernel
- Upload the jupyter notebooks to github for a walk-through during in person interview
- Send the github link to the above notebook a day before the interview i.e. before *11.59 PM 15 Feb 2018* in this case
- Make sure the jupyter notebook is self explanatory wherever needed with appropriate markdowns
- Feel free to make your own assumptions in case of any confusion
- Exercise would take anywhere between 4 - 8 hours
- *Note : Even if the problem is not completely solved, please upload your work for the interview. Our objective to analyze the process rather than the outcome*

Problem Statement*

1. Prediction task is to determine whether a person makes over 50K a year. Explain the performance of the model using accuracy, auc roc curve and confusion matrix. Feel free to add any other metric you see fit
2. Perform a segmentation study on the dataset to display useful information using any visualization library

**Upload a separate jupyter notebook for both the problem statement*

Dataset:

- **age**: continuous
- **workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
- **fnlwgt**: continuous
- **education**: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- **education-num**: continuous
- **marital-status**: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
- **occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
- **relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
- **sex**: Female, Male
- **capital-gain**: continuous

- **capital-loss**: continuous
- **hours-per-week**: continuous
- **native-country**: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
- **class**: >50K, <=50K

Find the dataset attached to the email
Training set: *au_train.csv*
Testing set: *au_test.csv*

Contact:
Incase of any questions, please contact:
Chirag Mandot : chirag.mandot@aunalytics.com