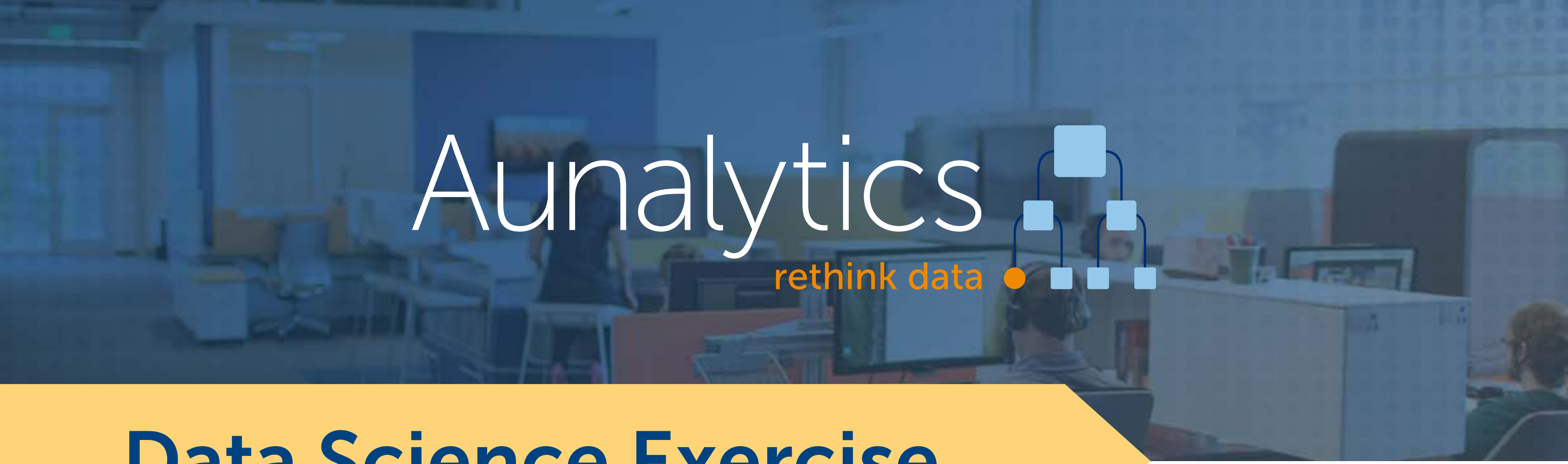

Aunalytics

rethink data

Data Science Exercise

Objective

Demonstrate your ability to solve a data science problem based on the information available and by making reasonable assumptions. 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 to chirag.mandot@aunalytics.com.
- 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.
- This exercise should take anywhere between 4 - 8 hours.
- **Note:** Even if the problem is not completely solved, please upload your work for the interview. Our objective is to analyze the process rather than the outcome.

Problem Statement*

1. The prediction task is to determine whether a person makes over 50K a year. Explain the performance of the model using accuracy, AUROC 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 each of the problem statements*

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 datasets attached to the email

Training set: au_train.csv

Testing set: au_test.csv

Contact

If you have questions, please contact:
Chirag Mandot (chirag.mandot@aunalytics.com)