Homework Assignment 4

Due May 30, 2019 5pm

Instructions:

This is an individual homework. Please do not work in groups. I want to make sure everyone has an understanding of how to do this.

In this homework, you'll build an interactive clustering tool. The goal is to incorporate unsupervised learning into the pipeline you've been developing, which will be useful for initial data exploration as well as exploring the high risk predictions from your supervised models.

Coding Assignment:

Develop a data exploration notebook that is interactive and uses clustering methods. Feel free to just use k-means for this homework.

The functionality should include:
1. load data from a csv
2. process the data so it's ready for clustering
3. Given a k, generate k clusters using one of the clustering methods (k-means is fine for this)
4. For each cluster:
    A. Provide summary stats for the cluster
    B. Describe (using statistics, graphs, or any other visualizations) what types of data points are in this cluster
    C. what are the distinctive features of data points in this cluster (you might want to use decision trees here)
5. Allow the user to
    A. merge several clusters into one
    B. recluster with a new k
    C. Split a specific cluster into many (with a specific number of new clustering)

Analysis:

Data: projects_2012_2013.csv

6. Once you've set up clustering code, you should apply it to the data from the previous homework to understand what types of projects are submitted added and also what type of projects are predicted as high risk of not getting fully funded (output of your last homework).

Report:

You should also write a short report (1-2 pages) that describes what types of clusters you found and what were the characteristics of projects in those clusters for:

1. the overall submitted projects

2. the top 5% of predicted projects from your test set that were not likely to be fully funded by your model.