CSE 519 -- Data Science (Fall 2017)
Prof. Steven Skiena
Homework 3: Data Integration and Modeling
Due: Monday, October 16, 2017

This homework will investigate data integration and model building in IPython. Our goal is to go deeper by working with a data set where you have some basic sense of familiarity as a result of Homework 2.

This homework is again based Zillow Prize Challenge on Kaggle, revolving around predicting the price that a particular real estate property (usually a home) will sell for. More than just data exploration, you must also join the challenge and submit your model before the deadline, to get a score back from Kaggle. **The final contest entries must be submitted to Kaggle by October 16 -- each student must submit two entries for scoring before that date**.

Be aware that Kaggle will be publishing a new training set on October 2. Please use that for all your experiments if you can.

Tasks

1. Build a scoring function to rank houses by "desirability", presumably a notion related to cost or value. Identify what the ten most desirable and least desirable houses in the Kaggle data set are, and write a one page description of which variables your function used and how well you think it worked. (10%)
2. Define a house "pairwise distance function", which measures the similarity of two properties. Like a distance metric, similar pairs of very similar properties should be distance near zero, with distance increasing as the properties grow more dissimilar. Experiment with your distance function, and write a one page discussion evaluating how well you think it worked. Your function should include geographic as well as property-specific variables. (10%)
3. Using your distance function and an appropriate clustering algorithm, cluster the houses using your distance function into 10 to 100 classes, as you see best. Present a dot-plot/map (with tiny dots colored to reflect the clustering) illustrating the clusters your method produced. Write a one page discussion/analysis of what your clusters seem to be capturing, and how well they work (15%)
4. Identify at least one external data set which you can integrate into your price prediction analysis to make it better. Perhaps it can be financial, such as the historical effects of interest rates, consumer confidence, etc. on housing prices. Perhaps it can be geographic, like the crime rate, educational scores, income levels, etc. Write a one page discussion/analysis on whether this data helps with the prediction tasks (20%)
5. Finally, build the best prediction model you can to solve the Zillow task. Use any data, ideas, and approach that you like. Predict the logerror for instances at file

"sample_submission.csv". Report the score/rank you get. You are allowed to merge your prediction teams to the extent that Zillow allows it. Write a 2-3 page report about how it works, an evaluation, and any interesting experiences along the way. (20%)

6. Do a permutation test to determine a *p*-value of how good your predictions of logerror are. You can use whatever metric you wish to score your model (like mean absolute error). For a large enough sample of the evaluation data, compare how your model ranks by this metric on the real data compared to 100 (or more) random permutations of the logerror assigned to the real data records. What fraction of permutations produce at least error at least as good at the real data set? If necessary, sample your data so these 100+ runs do not take too much time. (15%)
7. **Submit your results on the real test data to Kaggle before deadline.** Write the result into a csv file and submit it to the website. Actually, submit two for your two best models, to the extent that Zillow allows it. (10%)

Rules of the Game

1. I will allow you to work together in small teams (2-3) on the modelling to **the extent that Zillow allows you merge teams with someone**. Make sure you merge before any deadlines.
2. Your written analysis should be embedded in the notebook with figures and output.
3. There are some public discussions and demos relevant to this problem on Kaggle. It is okay for students to read these discussions, but they must write the code and analyze the data by themselves.
4. Our class Piazza account is an excellent place to discuss the assignment. Check it out at piazza.com/stonybrook/fall2017/cse519. You can also send email to the TA at cse519@cs.stonybrook.edu if you have questions about IPython, github, etc.

Submission

To submit your homework successfully, you need:

1. Create a folder named "HW3" under your repo at GitHub. Coding, debugging and saving your IPython notebook (named "HW3.ipynb") in folder "HW3".
2. Only the commits before the deadline will be considered to be graded. Thus make sure you have successfully pushed you files to the server (use git command "pull" to test).
3. Note that GitHub can show the IPython notebook directly, which means the content showing when you open HW3/HW3.ipynb should the same with what in your computer (browser). So we will grade directly based on the content of this file on you repo.
4. Remember to submit your predicting results to Kaggle and save the screenshot of your ranking and score into HW3 directory.
5. Please search Google for a solution your problem first or consult a fellow student before asking the TA. It will help you a lot.