CSE 519 -- Data Science (Fall 2019)
Prof. Steven Skiena
Homework 3: Data Integration and Modeling
Due: Tuesday, October 22, 2019

This homework will investigate data integration and model building in Python. Our goal is to go deeper by working with a data set where we have a very concrete idea of what we are working with.

This homework is based on the Ames Housing Dataset on Kaggle, revolving around predicting the price that a particular real estate property (usually a home) will sell for. This is a continuously ongoing competition intended for beginners to become acquainted with data science. Since the dataset is relatively small and every field is descriptive, we will be focused on analyzing the data itself.

Data downloading

First of all, you need to join the challenge and download the data here. The description of the data can also be found at this page.

Tasks

1. Select a set of 10-15 of the most interesting variables. Do a pairwise Pearson correlation analysis on all pairs of these variables. Show the result with heat map and find out most positive and negative correlations. You can use the seaborn library to plot the heatmap, with instructions found here.
2. Produce five other informative plots revealing aspects of this data. For each plot, write a paragraph in your notebook describing what interesting properties your visualization reveals. These must include:
    - at least one line chart
    - at least one scatter plot or data map
    - at least one histogram or bar chart
3. Build a handcrafted 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 description of which variables your function used and how well you think it worked.
4. 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 discussion evaluating how well you think it worked. When did it do well and when badly?
5. Using your distance function and an appropriate clustering algorithm, cluster the houses using your distance function into 5 to 20 classes, as you see best. Present a

visualization illustrating the clusters your method produced. How well do your clusters reflect neighborhood boundaries? (do not use neighborhood in your distance function) Write a discussion/analysis of what your clusters seem to be capturing, and how well they work.

6. Set up a simple linear regression model on one or more variables to predict the pricing as a function of other variables. How well/badly does it work? Which variable is the most important one?
7. Identify at least one external data set which you can integrate into your price prediction analysis to make it better. Write a discussion/analysis on whether this data helps with the prediction tasks.
8. For ten different variables (some likely good, some likely meaningless) from the data set, build single-variable regression models, and for each one do a permutation test to determine a *p*-value of how good your predictions of the housing prices are. Use root-mean-squared error of the log(price) to score your model. In other words, compare how your model ranks by this metric on the real data compared to 100 (or more) random permutations of the housing priced assigned to the real data records.
9. Finally, build the best prediction model you can to solve the task. Use any data, ideas, and approach that you like. Predict the pricing for instances at file "sample_submission.csv". Report the score/rank you get.
10. **Submit your results on Kaggle.** Write the result into a csv file and submit it to the website. You should do this for every model you develop. Report the rank, score, number of entries, for your highest rank. Include a snapshot of your best score on the leaderboard as confirmation. Be sure to provide a link to your Kaggle profile.

Rules of the Game

1. This assignment must be done **individually by each student**. It is not a group activity.
2. If you do not have much experience with Python and the associated tools, this homework will be a substantial amount of work. Get started on it as early as possible!
3. All of your written responses will be put in the appropriate place in your notebook template. Get the template notebook form from Google Classroom!! You are allowed to add more cells, but definitely fill out the cells we give.
4. We will discuss topics like linear regression in detail only after the HW is due. Muddle along for now, and we will understand the issues better when we discuss them in the course.
5. To ensure that you are who you are when submitting your models, have your Kaggle profile show your face as well as a Stony Brook affiliation.
6. 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.

7. You will submit your code so we can run it through MOSS to detect copying and plagiarism. Do your own work!!
8. Our class Piazza account is an excellent place to discuss the assignment. Check it out at piazza.com/stonybrook/fall2019/cse519.

Submission

Submit everything through Google classroom. As mentioned above, you will need to upload:

1. The Jupyter notebook all your work is in (.ipynb file), derived from the provided template
2. Python file (export the notebook as .py)
3. PDF (export the notebook as a pdf file)

These files should be named with the following format, where the italicized parts should be replaced with the corresponding values:

1. cse519_hw3_*lastname_firstname_sbuid*.ipynb
2. cse519_hw3_*lastname_firstname_sbuid*.py
3. cse519_hw3_*lastname_firstname_sbuid*.pdf