Data Science Class

Final Project

Submitted November 7, 2013

Jacob Bollinger posted a course update to Data Science: Section 2.

"Hello All,  
  
Here is the assignment due Thursday 10/31. The idea here is to begin to solidify your final projects. If you have questions or are in need of ideas please refer here:[http://cs229.stanford.edu/](http://email.schoology.com/wf/click?upn=-2FGqA7CRgMvumIizGz0Mc86Da6Jl4HbbNYJpmrPCb-2FiM-3D_N01VTKG1mXbSdp0UbJ4269B3Lzeg0tz7xsfDdsD-2BLm1E7n4edMIlhmQl6hawM02gTCD2HHWmfBxSThr2Gse3l1KUZh2htEq-2BWdJN-2BIPvDxSQhrLyY6KAGqDe4SWtkErDbwYdHyfr0w3qX7yZgVLQ50lPVj6geRJMn0oseADUshYVfJ6cHkJSA9PtZcN1aoKj8Tn26Mg6kI8Ms3Y9s5e8f9maT-2F1-2BQOnVFeGWYz8TmIx9EOwA7nuJ-2BE7QzoMded-2Ba" \o "http://cs229.stanford.edu/" \t "_blank)  
  
Final project:  
  
Acquire a dataset of your choice - You can use an api, download .csv files, scrape and parse html, use data from your company, etc. The fallback (least recommended but easiest) option is to use a kaggle dataset.  
  
Analyze and visualize the dataset - Provide statistics on the features with respect to the classes to be predicted. Which features best predict your classes?   
  
Feature extraction, engineering, and selection - Using methods discussed in class (or others), engineer and/or extract new features. Do these new features discriminate between your classes? Use a feature selection method.  
  
Build a model and provide analysis of the accuracy of you model using metrics/methods discussed in class.  
  
Future directions - How can you continue to improve your model?  
  
This assignment:  
  
Provide a justification of your model. What are you trying to predict and why it is important? Has anyone attempted something similar in the past? What is your dataset? How will you acquire your dataset? What are the potential problems you might run in to?   
  
Feel free to email me or Alice if you have questions."

To view the update, click the link below:  
[http://classroom.generalassemb.ly/course/60576373/update/71680805](http://email.schoology.com/wf/click?upn=KDpy7i7wfr4YaSX6R1G91WtQDZUsvTkdLPQ0L-2FUmEJucNgC9MANMvPiFktPuJuq0uWYN-2F9iUbNiuxrB1GQscOTk5sw10E5DsFxDaIdpYIe4-3D_N01VTKG1mXbSdp0UbJ4269B3Lzeg0tz7xsfDdsD-2BLm1E7n4edMIlhmQl6hawM02g5be6IxkxoMFEMXN3a-2F3djl1MiOaNwofXBWEPnZtKeaUSAgCg-2BqTM5ihHZEDQuyAkgJ-2BVwbMCXKBKdImZ0-2FFbKYl5KiiZJ2tMajd93lf-2BQD7mZdKc7SQfl7DA3VhppwfvRWibscypiN3CKxHZVHlNghX78jGMKd1NE59b1MyQAqxd1W2iy0-2BpEPuhyMlPkI2B" \o "http://classroom.generalassemb.ly/course/60576373/update/71680805" \t "_blank)

Objective: factors that can be used to predict changes in residential property prices. Specifically median sold price will be used as the dependent variable.

**Data**: Zillow data available <http://www.zillow.com/blog/research/data/#bulk> The data is available at the State, Metro, County, City, Zip code and Neighborhood levels. I plan on using the City level data. The files are easily downloadable as csv files from the Zillow website. Zillow provides almost 40 different files providing information as far back as 1998 on metrics such as percent of homes selling for loss, sale price to list ratio, homes sold as foreclosures, etc. I have also downloaded other possible predictors of property prices from a number of other websites such as unemployment data [http://stats.bls.gov/web/metro/laummtrk.htm](http://stats.bls.gov/web/metro/laummtrk.htm" \t "_blank), S&P historical data [http://finance.yahoo.com/q/hp?s=%5EGSPC+Historical+Prices](http://finance.yahoo.com/q/hp?s=%5EGSPC+Historical+Prices" \t "_blank), Income data [http://www.census.gov/hhes/www/income/data/historical/household/](http://www.census.gov/hhes/www/income/data/historical/household/" \t "_blank), Pop numbers [http://www.census.gov/popest/data/historical/2010s/vintage\_2011/city.html](http://www.census.gov/popest/data/historical/2010s/vintage_2011/city.html" \t "_blank)

[http://www.census.gov/popest/data/historical/2000s/vintage\_2009/index.html](http://www.census.gov/popest/data/historical/2000s/vintage_2009/index.html" \t "_blank), Crime data [http://www.ucrdatatool.gov/Search/Crime/Local/JurisbyJurisStepTwo.cfm](http://www.ucrdatatool.gov/Search/Crime/Local/JurisbyJurisStepTwo.cfm" \t "_blank), Counts of businesses [http://www.census.gov/econ/susb/data/susb2008.html](http://www.census.gov/econ/susb/data/susb2008.html" \t "_blank), FEMA’s natural disasters [http://www.fema.gov/disasters/grid/state-tribal-government/86?field\_disaster\_type\_term\_tid\_1=All](http://www.fema.gov/disasters/grid/state-tribal-government/86?field_disaster_type_term_tid_1=All" \t "_blank)

Needless to say there are many possible features. One objective of this analysis would be to do pca on the data and identify fewer variables but increase the prediction power.

The second objective would be to build a predictive model for change in median sold price. I am not clear yet whether I want to be able to predict change in median sold price or the median price itself.

The Zillow data is available for many years as indicated above, I am not sure if I will be able to do any time series analysis.

I plan on dividing the City data into test and training set.