Intro to Machine Learning: Individual Prediction Project

Due Aug 4 at 10pm (CT)

When building predictive models, collecting additional data/variables (or "features") and/or transforming or combining variables to create new features ("feature engineering") can give bigger improvements than choosing a more complicated model or method. The goal of this project is to explore this phenomenon.

We'll continue working with the Austin houses data from take-home problem 3. Your goal here is to build predictive models of house prices. The complete data set consists of over 13,000 houses in the Austin area. The data was scraped from Zillow, meaning that an automated process captured the data from Zillow's public web site. (Because this was done by a computer and not a human, it can introduce errors!).

You will need two datasets for this project: `austinhouses.csv`, which is complete, and `austinhouses_holdout.csv`, which has the same variables but is missing the house prices. This is the "validation" set I'll use to evaluate your predictions.

1. In take-home problem 3 you used a subset of the variables. Now we'll increase the set of variables you consider to include all the predictors except the street address and description.
    a. Begin by verifying that each of these variables are appropriate to include as predictors. If any are not, explain why.
    b. Consider how each of these should enter your models. For example, are there numeric columns that correspond to *categorical* variables? Would it make sense to aggregate or recode any of the categorical variables to define new variables? Would it make sense to combine or transform any of the variables at this stage, based on what you know about the housing market?
2. Repeat take-home problem 3 with the expanded set of variables. How does your estimated out of sample prediction error change for each of the methods? (Make sure you use the same training/testing split.)
3. Prediction contest! Choose a set of predicitons to enter the contest. You can choose your best performer from part 2 above (and receive full credit for this part), or you can go wild here – use any method you wish, whether or not we covered it in class, *as long as you can describe it in a few paragraphs.* You will submit 1) a < 1 page description of your approach, b) a csv file containing your predicted prices, with one row for each case in the same order as `austinhouses_holdout.csv`, and c) the R script that generates your predictions. (You may also submit an augmented dataset if you added or modified features outside of R.)

The winner of the contest will receive fame, glory, and a special prize!

Some tips for getting the best predictions for the contest

1) Use prediction errors (residuals) to help you refine your approach. Summarize some of the cases with large prediction errors; is there anything special about them you could use to define new features? Plot prediction errors against the variables in your dataset, in space using lat/long, try to predict them using interpretable shallow trees, etc. Use what you find to define new features: Anything you can use to predict the prediction errors is a good candidate to include as a new feature! *Avoid evaluating on your testing data at this stage, or you'll overfit!*
2) You are free to add variables to this dataset! How could you do this? Here are two ideas:
    a. Add neighborhood-level features using the zipcode. One good source for zip-level data is IPUMS, which has a point-and-click interface to make things easier for you. A short tutorial for merging

datasets in R is available here. Or use other tools you know – SQL, VLOOKUP in Excel, etc.

b. Inspect the description variable. Can you come up with ways to extract more information from the description (e.g. the presence or absence of words or word pairs, the length of the description, etc)?