Kaggle Ames Housing Project Reports

Emily Goren, Andrew Sage, Haozhe Zhang

Report (2) (on April 14, 2017)

Current Progress

Method	Cross-validation RMSE
Elastic Net (tuned to Ridge Regression)	27267.2
Random GLM (order 2)	609239.4
Random Forest	27928.9
Conditional Random Forest	30489.9
PLS	32350.0

Kaggle Scoreboard: 0.12746 (#987) using average of random forest and PLS predictions. We carefully tuned and cross-validated the Elastic Net, PLS, random forest, and conditional random forest predictions individually. We will start to build essemble predictions based on current models.

Report (1) (on April 7, 2017)

Our team "CycloneSTAT" made a primitive attempt of submission to Kaggle this week. The current public score is 0.13037 with a ranking of #1149 (out of 2244 teams). This week we focused on data cleaning, data exploration, handling missing values and feature selection. We will start to do repeated cross validation to select models.

Data Cleaning

Including but not limited to:

- Change MSSubClass to a factor;
- Combine condition1 and condition2 variables into an indicator for each level.
- For numeric variables, replace `NA`'s with the median value;
- For factor variables, make NA's into their own factor level. Most of these are "not applicable"" so they'll likely be correlated with similar variables, e.g. garage area and garage quality;
- Change the two condition variables into binary indicators for each condition, etc.

Feature Selection

We used a feature selection R package called "Boruta" (see Kursa & Rudnicki, J Stat Software, 2010) to identify a total of 56 "important" features for prediction. The importance ranking of these features were also investigated in a linear model, boosted tree model, and random forest (all untuned). Most of the features we considered were the raw variables provided in the training set.It seemed that the models without the unimportant variables did better, but we didn't test this very extensively.

Predictive Methods

We used "caret" package to select tuning parameters for PLS, PCR, RF and tree methods, and used these models to make predictions on the test data. At this point, we just took a straight average of the estimates produced by the 4 techniques. We also fitted an elastic net model using the Boruta selected features and their first order interactions (tuning to include ridge and lasso penalties).