Full House price prediction

Housing price prediction challenge from Kaggle

CDIPS Data Science Workshop Aug. 5, 2017

Shuyang Li, Wyatt Hartman, Xiwang Li

Mentor Andy Vargas

What are you looking for in a home?

Neighborhood

n BR, *n* bath
n floors
n SQFT
HW FLR
Yard, BSMNT

House Prices: Advanced Regression

kaggle

1460 sale prices 930 witheld
80 features

37 continous
43 categorical

Eames, Iowa housing dataset
De Cock 2011 J. Stat. Educ.

Making *“the best”* predictions

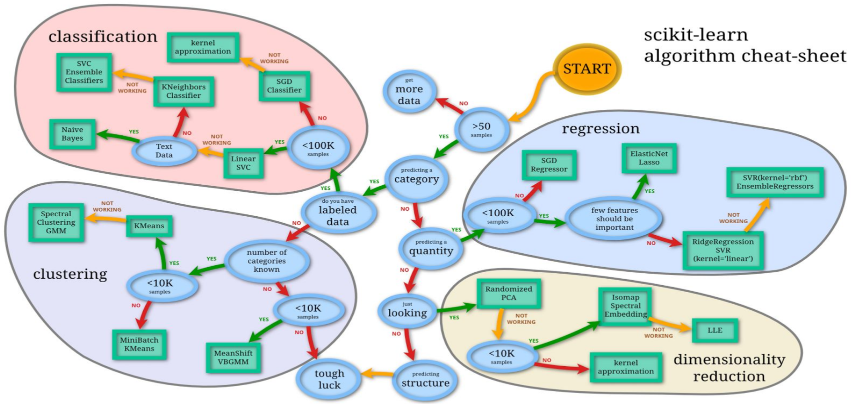
Goals and objectives

Best fit to the data

Appropriate algorithm

Parsimonious ?

Interpretable ?



Feature Engineering

Automagic

Transform skewed

NA filling

pd.getdummies

Manual Intervention…

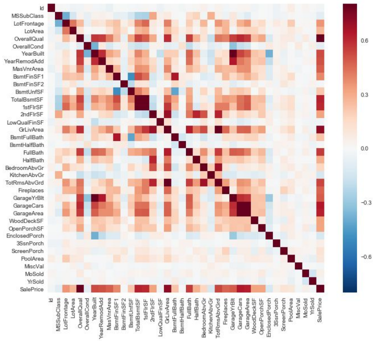
“Manual” feature engineering

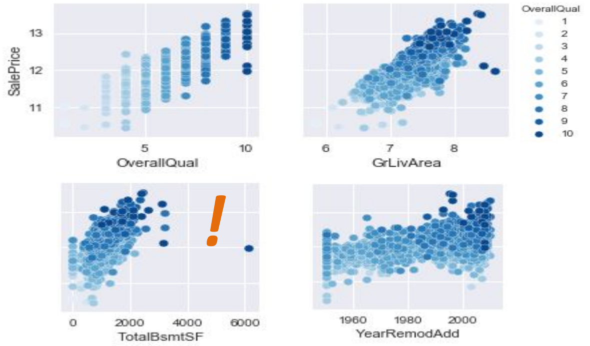
ID important features

Correlation heatmap

Top single features

Co-linear predictors





Recode / combine features

NA ⟶ “none”, 0, mode/average

Ranked cat’s. ⟵ ⟶ num.

Combine related:

- Total: Sqft., Baths
- Quality: Exter, Kitch, Bsmt, Gar.

Total SF=

Machine Learning

Algorithm toolbox

Regression	Linear, Ridge, Lasso, Elastic, NN
Trees	Random Forest, XGBoost
Scoring	RMSE, 10 fold cross-validation

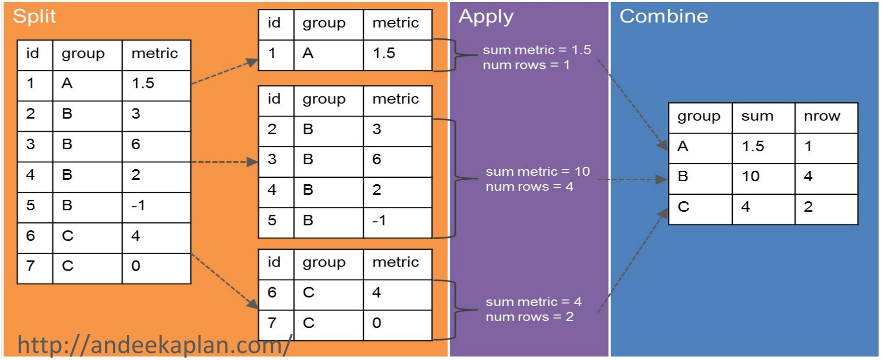
Split datasets

Test / train / witheld

Manual FE only

Continuous only / cat. only

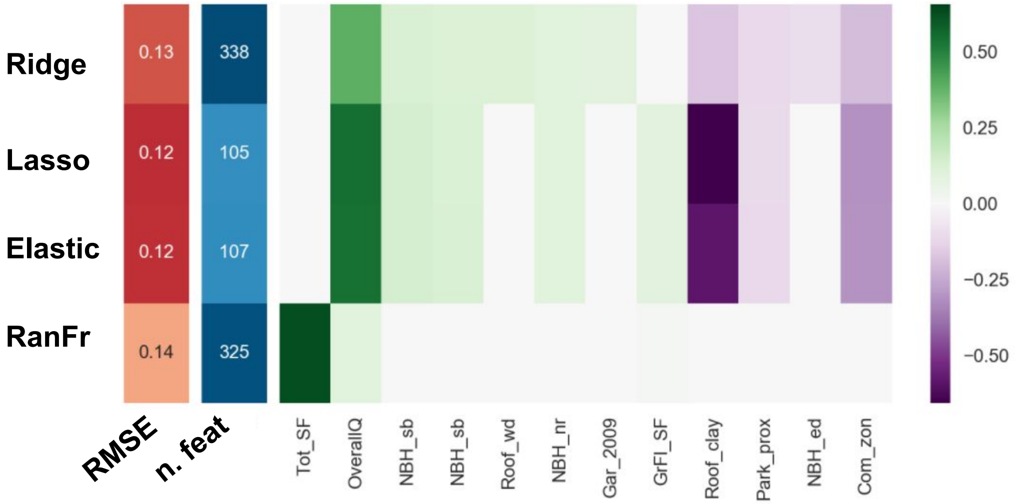
Full data set



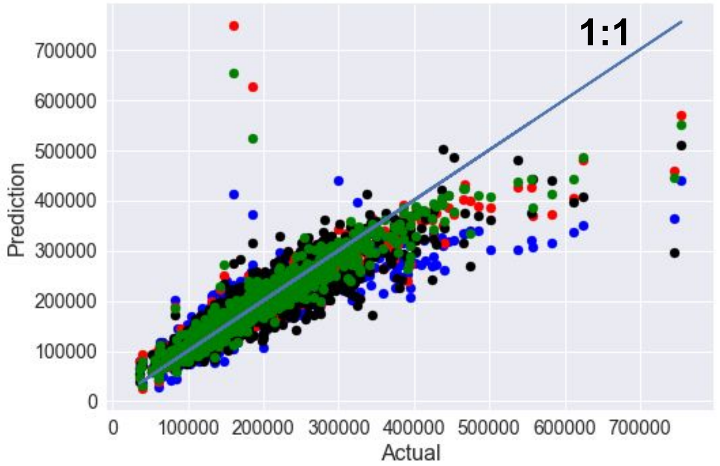
Evaluate models …

Results / Discussion

Algorithm scores, top features



By data subset (Kaggle score)



Subset, K RMSE

- 10_top, K=0.195
- 55_num, K=0.136
- 325_cat, K=0.204
- 380_all, K=0.123

Lessons learned and future development

Outliers are from 2008, economic recession

Piecewise estimation: classification + regression