Assessing house prices

Gonzalo Rubio and Breno Bahia

ECE625 Data Analysis
and Knowledge Discovery
April, 2018

Preprocessing

Kept 63,781 / 71524 observations

>> Complete missing values

> Lot size

> Site coverage

> 2016 house prices

> effective building year

> building count

Kept 17/35 variables

Created 5 new variables (22 predictors in total)

> Useful area

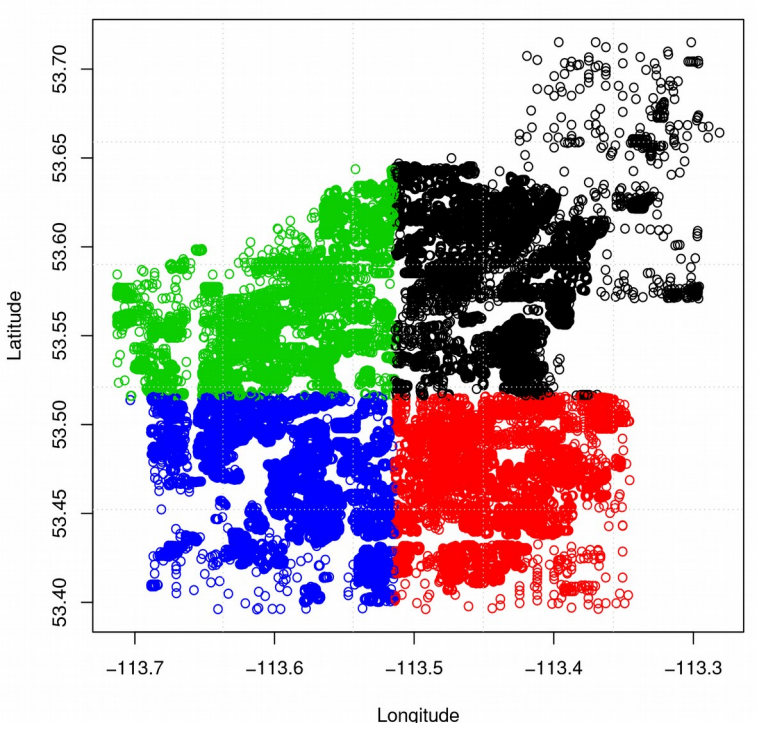
> Quadrant

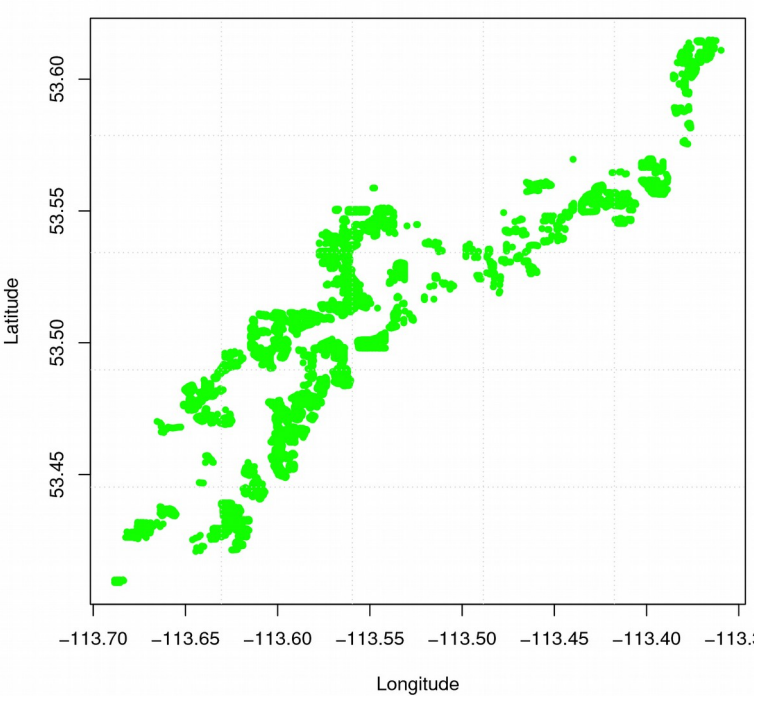
> School distance

> Job distance

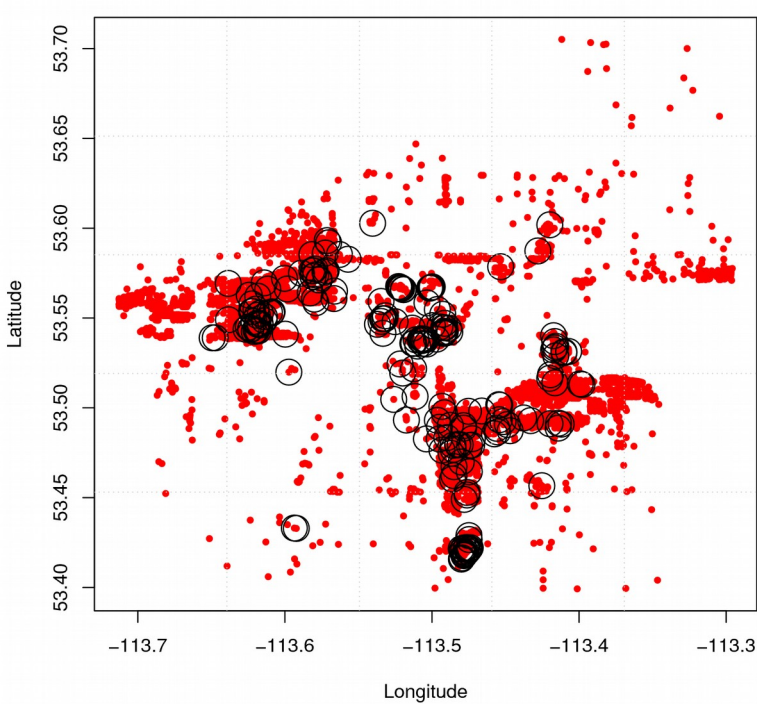
> Park distance

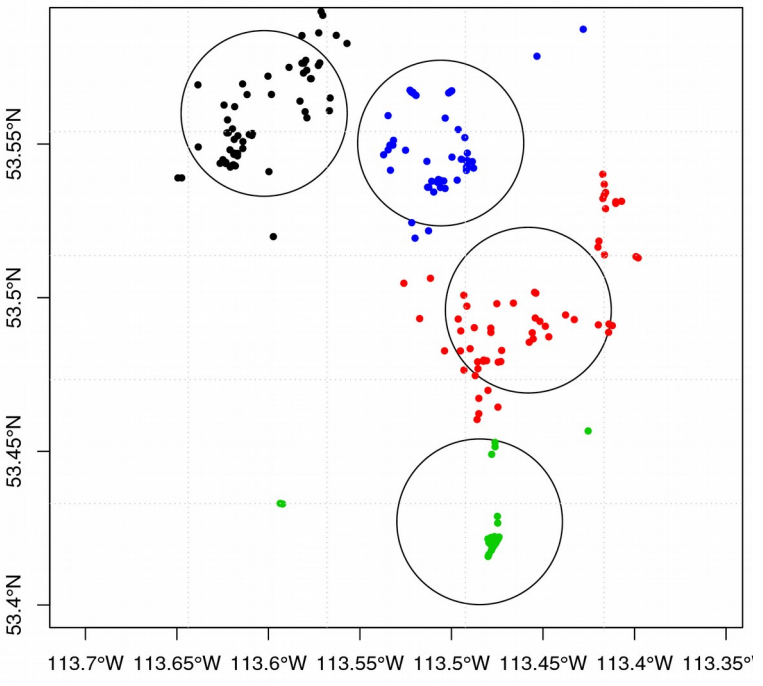
Feature addition and transformation



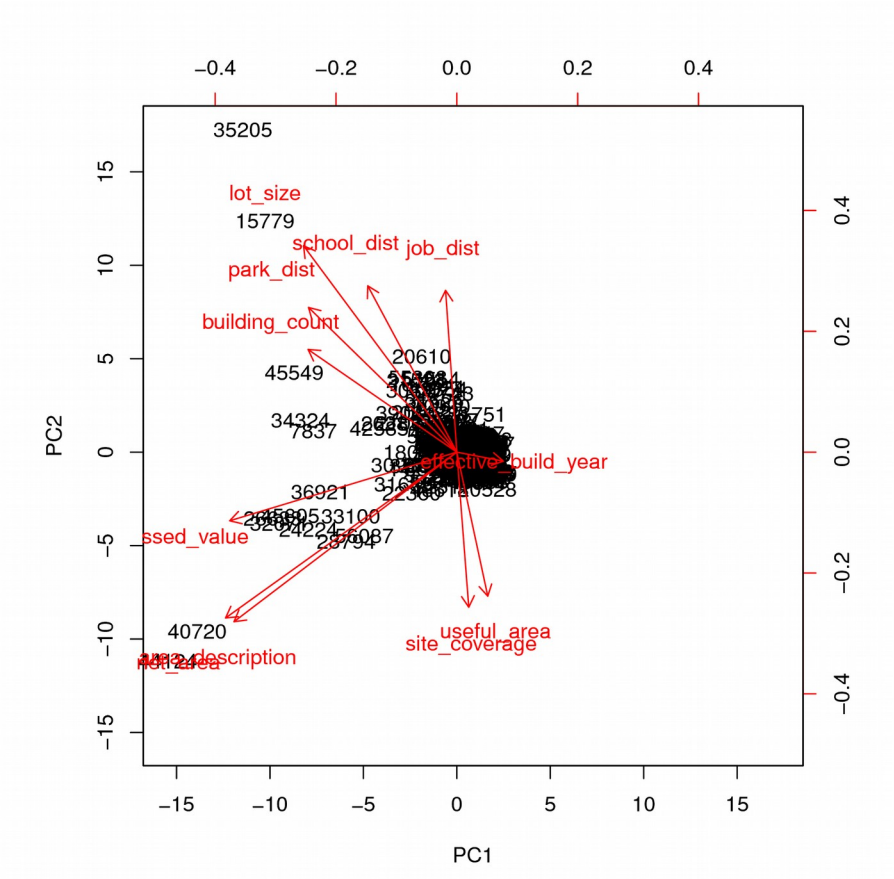


Feature addition and transformation





Feature addition and transformation



Objective I

Method	MSE (/10^7)	Var	Predicted median	Test median
LR	0.01919	1,385,302	814,283.8	845,742.9
Lasso	0.78385	8,853,563	864,602.1	845,742.9
GBM	0.00998	999,470.5	830,432.7	845,742.9

Non-linear terms, such as poly(gross_area,2) and log(lot_size), are added to the model.

Better MSE can be achieved by fitting different models for different categories (e.g., separate data based on display_type) but we stick to a unique model to the whole city.

GBM was fit with n.trees = 5000 and interaction.dept = 10.

Objective II

Method	Logistic Reg.	LDA	QDA	Pruned Tree
Accuracy rate	72.7 %	72.6%	68.4%	45.4%

Logistic regression consistently gave ~28% error rate for different combinations of predictors.

Almost Identical results for Logistic regression and LDA across all fits.

Classification Trees with 6 nodes.

Thank you