Detect other languages: just 5, mixed with english. Decided to leave in dataset.

Adding other predictors based on the words people describe food

Dealing with NA via available package

Check the randomness of NA rows first. Find the distribution of NA rows (based on stars) does not differ significantly from overall distribution of stars. No evidence against the randomness of NA stars. Hence assumes that the missing data are Missing at Random (MAR), which means that the probability that a value is missing depends only on observed value and can be predicted using them.

MICE: missing at random, remove categorical variables (all predictors in this data are assumed to be numeric)

missForest

mi

<https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/>

<http://www3.imperial.ac.uk/pls/portallive/docs/1/1171924.PDF>

1st model: stars ~ X1 + X2 + X3 + X4 + X5 + X15 + X18 + X23 + X24 + X25 + X26 + X27 + X29 + X30 + X33 + X36 + X38 + X52 + X54 + X55 + X61 + X63 + X66 + X68 + X71 + X75 + X77 + X86 + X94 + X71:X75

model = lm(stars ~ X1 + X2 + X3 + X5 + X7 + X9 + X10 + X11 + X12 + X14 + X15 + X16 + X17 + X18 +

X19 + X20 + X21 + X22 + X23 + X24 + X25 + X26 + X27 + X28 + X29 + X30 + X31 + X32 +

X33 + X35 + X36 + X37 + X38 + X39 + X40 + X41 + X42 + X43 + X44 + X46 + X47 + X49 +

X50 + X52 + X54 + X55 + X56 + X57 + X58 + X59 + X61 + X62 + X63 + X65 + X66 + X67 +

X68 + X69 + X71 + X72 + X73 + X75 + X77 + X78 + X80 + X81 + X82 + X84 + X85 + X86 +

X87 + X88 + X90 + X91 + X92 + X93 + X94 + X95 + X96 + X98 + X99 + X100 + X101 +

X38:X94 + X94:X98 + X71:X75 + X52:X94 + X36:X94 + X36:X52 + X75:X94 + X52:X75 +

X36:X75 + X77:X95 + X36:X52:X94 + X52:X75:X94 + X36:X75:X94 + X33:X94:X95 +

X33:X77:X95 + X33:X77:X94:X95, data = Data)

Mallow cp and adjusted r^2 all based on SSE. when p (# of variables) are relatively small compared to n, those two criteria are expected to be similar

Thoughts: come up a method (similar to penalized regression), that add constraints on the predicted value of a model

10048: leverage point. Detailed review for every item ordered, a lot of positive words. Should divide the occurrence of all words by the total number of words in text.