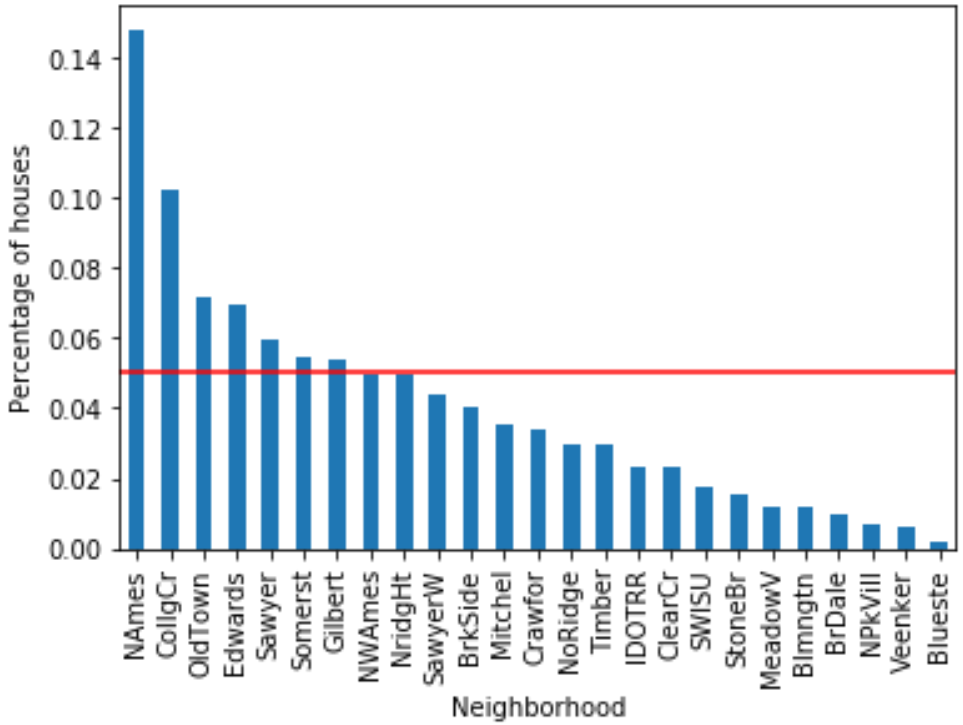
Train In Data

Rare Label Encoding

Grouping Rare Labels

- Rare labels are those that appear only in a tiny proportion of the observations in a dataset



Grouping Rare Labels

- Rare labels are those that appear only in a tiny proportion of the observations in a dataset

Some Scenarios

- Variables with one predominant category
- Variables with few categories
- Variables with high cardinality

Street	
Grvl	0.004892
Pave	0.995108

ExterQual	
Ex	0.029354
Fa	0.011742
Gd	0.332681
TA	0.626223

Exterior2nd	
AsbShng	0.016634
AsphShn	0.000978
Brk Cmn	0.003914
BrkFace	0.017613
CBlock	0.000978
CmentBd	0.038160
HdBoard	0.137965
ImStucc	0.007828
MetalSd	0.133072
Other	0.000978
Plywood	0.109589
Stone	0.003914
Stucco	0.015656
VinylSd	0.345401
Wd Sdng	0.138943
Wd Shng	0.028376

Some Scenarios

- Variables with one predominant category

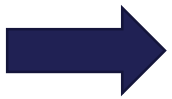
Street	
Grvl	0.004892
Pave	0.995108

Street	
Rare	0.004892
Pave	0.995108

Some Scenarios

- Variables with few categories
  - Worth corroborating value of rare labels

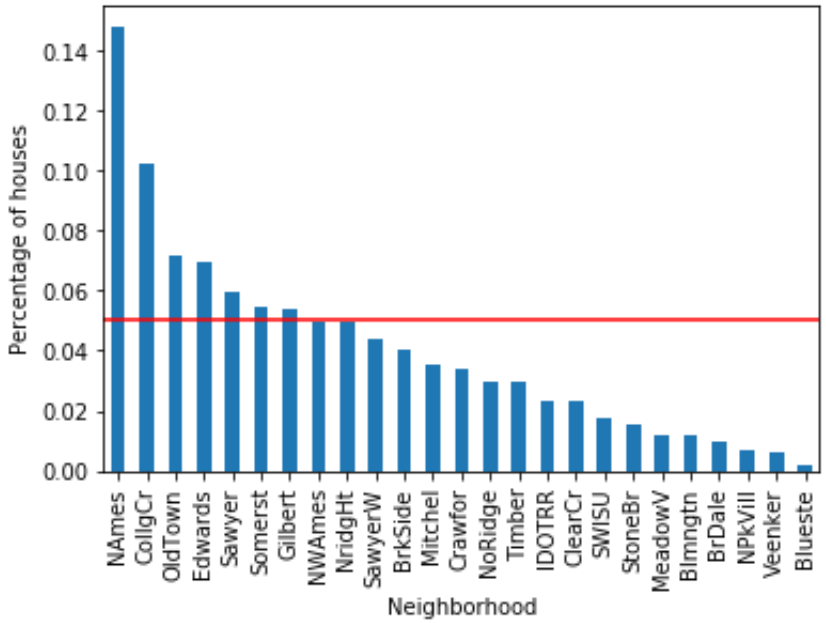
ExterQual	
Ex	0.029354
Fa	0.011742
Gd	0.332681
TA	0.626223

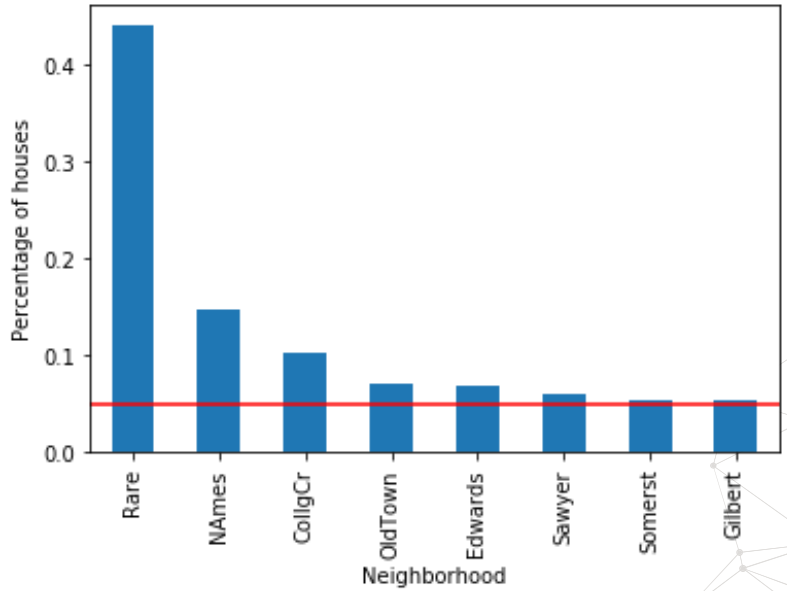


ExterQual	
Rare	0.04
Gd	0.332681
TA	0.626223

Some Scenarios

- Variables with high cardinality





Code should capture frequent categories

- This way, categories that are new in test set, will be treated as rare and put into the Rare group
- Model will know how to handle those categories as well, even though they were not present in the train set

Train In Data

THANK YOU

www.trainindata.com