Train In Data

One hot encoding of top categories

One hot encoding of top categories: definition

- Performing one hot encoding, only considering the most frequent categories

- in the winning solution of the KDD 2009 cup: "Winning the KDD Cup Orange Challenge with Ensemble Selection", the authors limit one hot encoding to the 10 most frequent labels of the variable.

One hot encoding of top categories: example

Variable = City

London → 1000 observations

Manchester → 500 observations

Leeds → 200 Observations

Yorkshire, Milton-Keynes, Cambridge => 10 observations each

London	Manchester	Leeds
1	0	0
0	0	1
1	0	0
0	1	0
0	0	1
0	1	0
0	0	0

One hot encoding of top categories: Advantages

- Straightforward to implement
- Does not require hrs of variable exploration
- Does not expand massively the feature space
- Handles new categories in test set
- Suitable for linear models

One hot encoding of top categories: Limitations

- It does extend the feature space to some degree
- Does not add extra information while encoding
- Does not keep the information of the ignored labels

THANK YOU

www.trainindata.com