Train In Data

Binary encoding and Feature Hashing

One hot encoding: definition

- Create 1 binary variable per category.
- Each single derived variable has a meaning on its own.

COLOR
Red
Blue
Green
Yellow

RED	BLUE	GREEN	YELLOW
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Binary encoding: definition

- Use binary code, a collection of 0s and 1s, to encode the meaning of the variable.
- Individually derived variable lack human readable meaning.

COLOR
Red
Blue
Green
Yellow

VAR1	VAR2
1	0
0	1
1	1
0	0

Feature hashing: definition

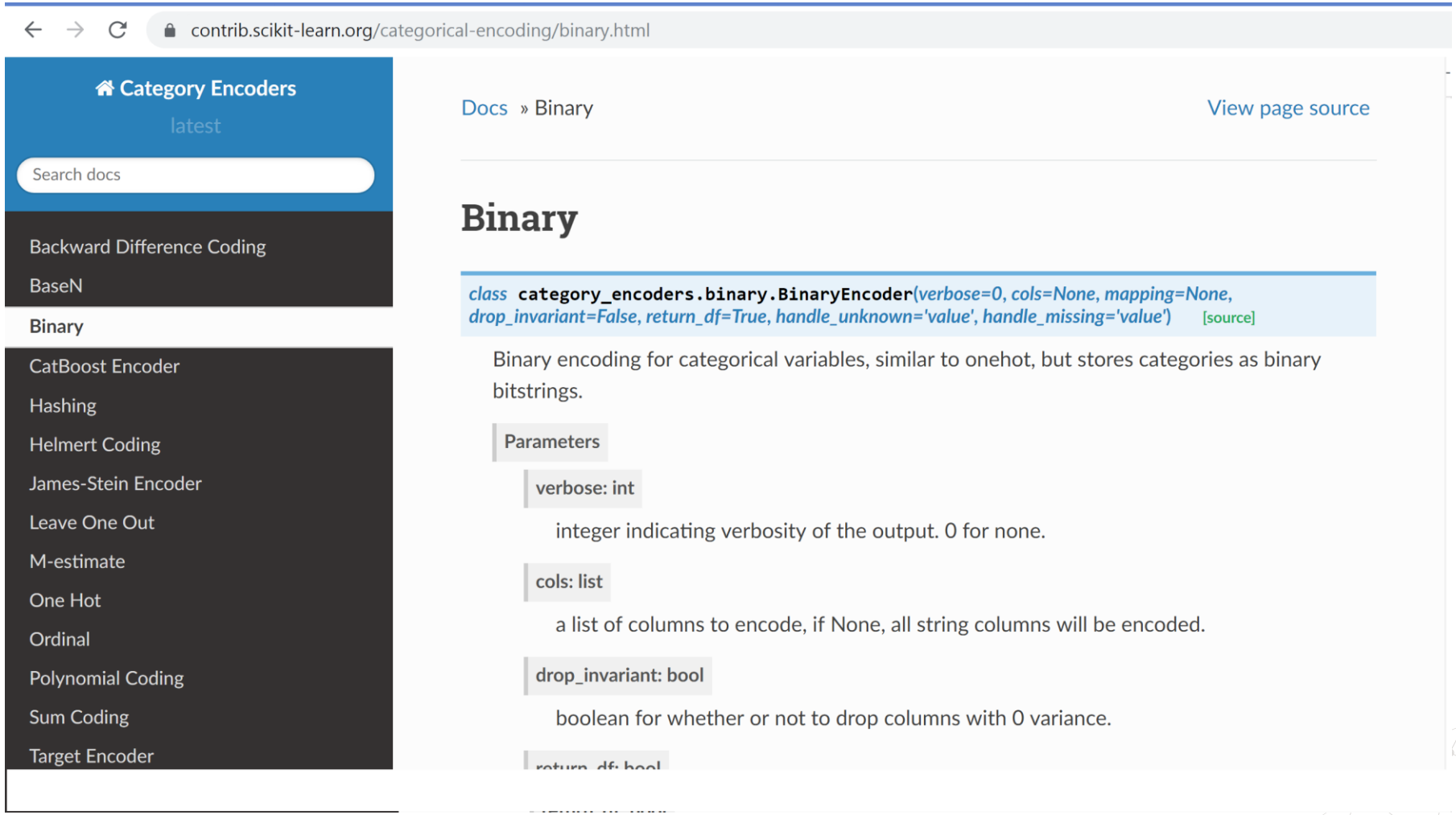
- Use a hashing method, any of choice, to encode the variables
- May lead to different labels taking the same value

COLOR
Red
Blue
Green
Yellow

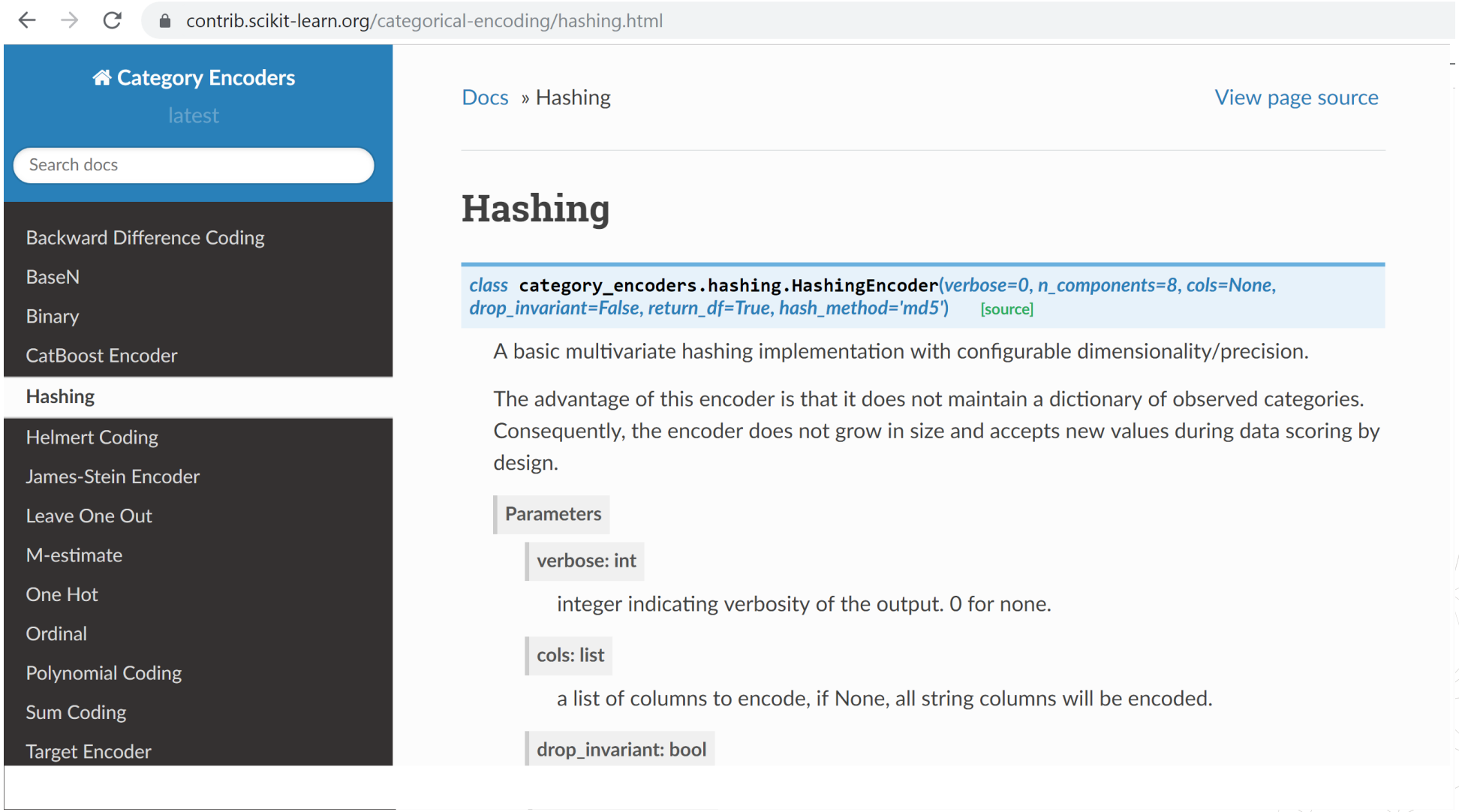
HASH
2
1
0
1

VAR1	VAR2	VAR3
0	0	1
0	1	0
1	0	0
0	1	0

Binary encoding: Category encoders



Feature hashing: Category encoders



Train In Data

THANK YOU

www.trainindata.com