CATEGORICAL FEATURES

A categorical feature is a feature with a finite number of discrete (unordered) *levels* $c_1, \dots, c_k$, e.g., *House.Style=2Story*$\overset{?}{>}$*SFoyer*.

- Categorical features are very common in practical applications.
- Except for few machine learning algorithms like tree-based methods, categorical features have to be encoded in a preprocessing step.

Encoding is the creation of a fully numeric representation from a categorical feature.

- Choosing the optimal encoding can be a challenge, especially when the number of levels k becomes very large.

ONE-HOT ENCODING

- Convert each categorical feature to k binary $(1/0)$ features, where k is the number of unique levels.
- One-Hot encoding does not loose any information of the feature and many models can correctly handle binary features.
- Given a categorical feature x_j with levels $c_1, \dots, c_k$, the new features are

$$\tilde{x}_{j,c} = \mathbb{I}(x_j)_c \quad c = c_1, \dots, c_k.$$

One-Hot encoding is often the go-to choice for the encoding of categorical features!

ONE-HOT ENCODING: EXAMPLE

Original slice of the dataset:

SalePrice	Central.Air	Bldg.Type
189900	Y	1Fam
195500	Y	1Fam
213500	Y	TwnhsE
191500	Y	TwnhsE
236500	Y	TwnhsE

One-Hot Encoded:

SalePrice	Central.Air.N	Central.Air.Y	Bldg.Type.1Fam	Bldg.Type.2fmCon	Bldg.Type.Duplex	Bldg.Type.Twnhs	Bldg.Type.TwnhsE
189900	0	1	1	0	0	0	0
195500	0	1	1	0	0	0	0
213500	0	1	0	0	0	0	1
191500	0	1	0	0	0	0	1
236500	0	1	0	0	0	0	1

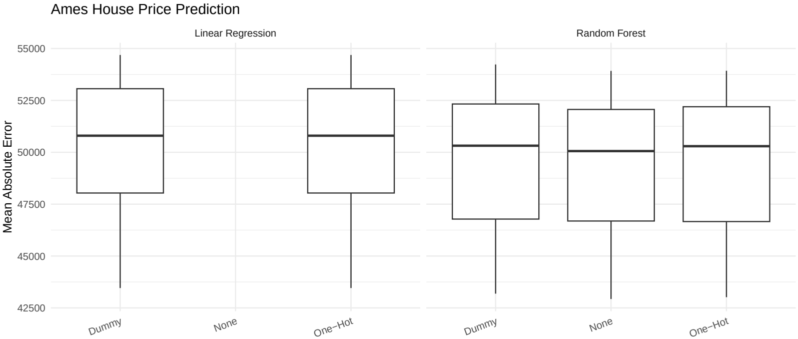
DUMMY ENCODING

- Dummy encoding is very similar to one-hot encoding with the difference that only $k-1$ binary features are created.
- A *reference* category is defined as all binary features being 0, i.e.,

$$\tilde{x}_{j,1} = 0, \ldots, \tilde{x}_{j,k-1} = 0.$$

- Each feature $\tilde{x}_{j,1}$ represents the *deviation* from the reference category.
- While using a reference category is required for stability and interpretability in statistical models like (generalized) linear models, it is not necessary, rarely done in ML and can even have negative influence on the performance.

AMES HOUSING - ONE-HOT VS. DUMMY ENCODING



- Result of linear model depends on actual implementation, e.g., R's `lm()` produces a *rank-deficient fit* warning and recovers by dropping the intercept.

ONE-HOT ENCODING: LIMITATIONS

- One-Hot encoding can become extremely inefficient when number of levels becomes too large, as one additional feature is introduced for every level.
- Assume a categorical feature with $k = 4000$ levels, by using dummy encoding 4000 new features are added to the dataset.
- These additional features are very sparse.
- Handling such *high-cardinality categorical features* is a challenge, possible solutions are
  - specialized methods such as *factorization machines*,
  - **target/impact encoding**,
  - clustering feature levels or
  - feature hashing.