Feature Engineering and Selection…
Chapter 5: Encoding Categorical Variables, pt 1

Agenda

Encoding Categorical Predictors

“The approach to including the predictors depends on the type of model. A large majority of models require that all predictors be numeric.”

- **5.1 Creating Dummy Variables for Unordered Categories**
- **5.5 Encodings for Ordered Data**
- **5.7 Factors versus Dummy Variables in Tree-Based Models**
- **5.2 Encoding Predictors with Many Categories**
- **5.3 Approaches for Novel Categories**
- **5.4 Supervised Encoding Methods**

Next week will cover 5.6 Creating Features from Text Data

5.1 Creating Dummy Variables for Unordered Categories

- "The mathematical function required to make the translation is often referred to as a *contrast* or *parameterization* function. An example of a contrast function is called the "reference cell" or "treatment" contrast, where one of the values of the predictor is left unaccounted for in the resulting dummy variables. Using Sunday as the reference cell, the contrast function would create *six* dummy variables"
  - Sometimes called design variables

Original Value	Dummy Variables					
	Mon	Tues	Wed	Thurs	Fri	Sat
Sun	0	0	0	0	0	0
Mon	1	0	0	0	0	0
Tues	0	1	0	0	0	0
Wed	0	0	1	0	0	0
Thurs	0	0	0	1	0	0
Fri	0	0	0	0	1	0
Sat	0	0	0	0	0	1

One-hot encoding is similar, but has value for each level

One-Hot Encoding

Mon	1	0	0	0	0	0	0
Tue	0	1	0	0	0	0	0
Wed	0	0	1	0	0	0	0
Thu	0	0	0	1	0	0	0
Fri	0	0	0	0	1	0	0
Sat	0	0	0	0	0	1	0
Sun	0	0	0	0	0	0	1

7x7

https://2.bp.blogspot.com/-r9EY2uYSTjY/Wnn-nNbw1cI/AAAAAAAAAN4/DXj8ZmaJm3IivJp_imEG7JtbQ7UvvgcygCLcBGAs/s1600/Untitled%2B1%25281%2529.png

5.5 Encodings for Ordered Data

- With "*polynomial contrasts*, we can investigate multiple relationships (linear, quadratic, etc.) simultaneously by including these in the same model."
  - … regular dummy encoding would not capture the "order" of a variable with categories such as {"low", "medium", "high"}

Alternatives to polynomial contrasts:

- Just use "unordered" factors
- "Translate the ordered categories into a single set of numeric *scores* based on context-specific information" (e.g. 1, 2, … 10 for ordering things from bad to great)

Table 5.3: An example of linear and quadratic polynomial contrasts for an ordered categorical predictor with three levels.

Original Value	Dummy Variables	
	Linear	Quadratic
low	-0.71	0.41
medium	0.00	-0.82
high	0.71	0.41

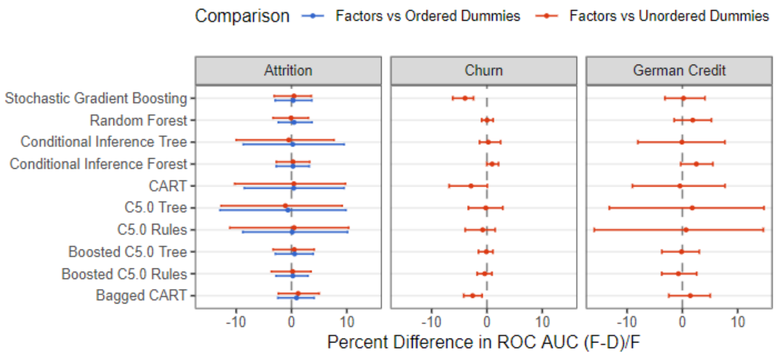
5.7 Factors versus Dummy Variables in Tree-Based Models

- (in R) many models do not require conversion to dummy variables, and can handle factors (tree-based methods especially) E.g.

```
if day in {Sun, Sat} then ridership = 4.4K
 else ridership = 17.3K
```

```
if day = Sun then ridership = 3.84K
  else if day = Sat then ridership = 4.96K
    else ridership = 17.30K
```

- Performance comparison across different models and datasets:



- Not huge differences in performance, though is still data & model specific

Recommendation:

- Usually prefer using factors where possible
  - Computation is faster
  - Things like variable importance are more intuitive and easier to measure

5.2 Encoding Predictors with Many Categories

Zero-variance predictor: variable or level that contains a single value

Approaches:

Near zero variance:
19:1 is common ratio to call something NZV

- Create full dummy set, remove zero-variance predictors or near-zero variance predictors
- Pool together into “other” category
- The “hashing trick”

Note: for rarely occurring levels, is possible something is not zero-variance on ENTIRE dataset, but is zero-variance (or NZV) on individual samples

5.3 Approaches for Novel Categories

Matters more during deployment (when new levels may come-up)

<u>Approaches:</u>

- Assign to “other” category
- Retrain model
- (if using hashing trick) “hash” in same with other levels

5.4 Supervised Encoding Methods

Regression example

- Effect/likelihood encoding: “the effect of the factor level on the outcome is measured and this effect is used as the numeric encoding”
- “For example, for the Ames housing data, we might calculate the mean or median sale price of a house for each neighborhood from the training data and use this statistic to represent the factor level in the model.”

```
# A tibble: 28 x 2
   Neighborhood        Neighborhood_effect_price
   <fct>                                   <dbl>
 1 North_Ames                            145097.
 2 College_Creek                         201803.
 3 Old_Town                              123992.
 4 Edwards                               130843.
 5 Somerset                              229707.
 6 Northridge_Heights                    322018.
 7 Gilbert                               190647.
 8 Sawyer                                136751.
 9 Northwest_Ames                        188407.
10 Sawyer_West                           184070.
# ... with 18 more rows
```

5.4 Supervised Encoding Methods

Classification example

Effect/likelihood encoding (examples from OkC data)

- "If the outcome event occurs with rate p, the *odds* of that event is defined as p/(1−p)p/(1−p). As an example, with the OkC data, the rate of STEM profiles in Mountain View California is 0.53 so that the odds would be 1.125."
    - *"Logistic regression models the log-odds of the outcome as a function of the predictors."*

Shrinkage methods: "if the *quality* of the data within a factor level is poor, then this level's effect estimate can be biased towards an overall estimate that disregards the levels of the predictor. "Poor quality" could be due to a small sample size or, for numeric outcomes, a large variance within the data for that level. Shrinkage methods can also move extreme estimates towards the middle of the distribution."

- Bayesian and empirical Bayesian methods can be used… (as can regularization techniques)

Table 5.2: Supervised encoding examples for several towns in the OkC Data.

	Data		Log-Odds	
Location	Rate	n	Raw	Shrunk
belvedere tiburon	0.086	35	-2.367	-2.033
berkeley	0.163	2676	-1.637	-1.635
martinez	0.091	197	-2.297	-2.210
mountain view	0.529	255	0.118	0.011
san leandro	0.128	431	-1.922	-1.911
san mateo	0.277	880	-0.958	-0.974
south san francisco	0.178	258	-1.528	-1.554
<new location>				-1.787

5.4 Supervised Encoding Methods

Classification example, word/entity embedding approaches

- “Similar to the dimension reduction methods described in the next chapter, the idea is to estimate a smaller set of numeric features that can be used to adequately represent the categorical predictors”
  - R er
    di

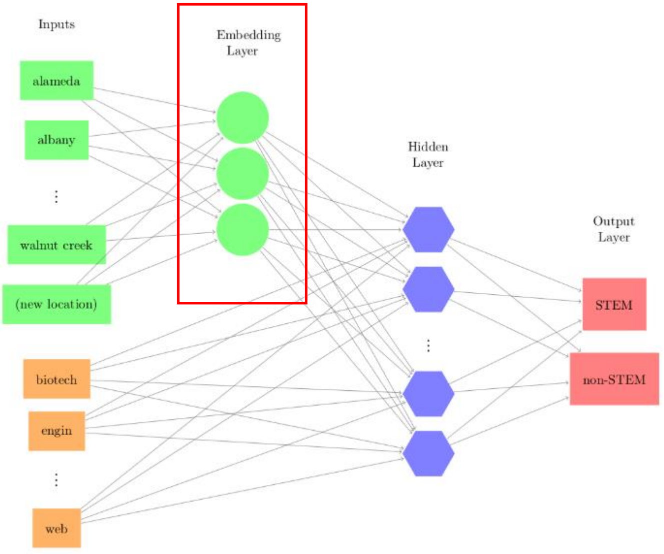
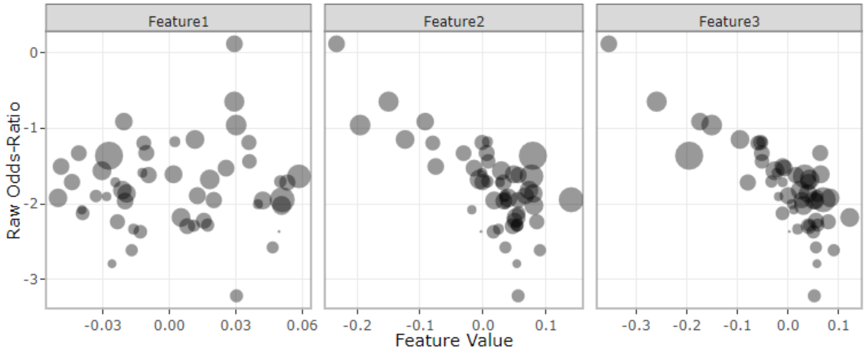


Table 5.2: Supervised encoding examples for several towns in the OkC Data.

	Data		Log-Odds		Word Embeddings		
Location	Rate	n	Raw	Shrunk	Feat 1	Feat 2	Feat 3
belvedere tiburon	0.086	35	-2.367	-2.033	0.050	-0.003	0.003
berkeley	0.163	2676	-1.637	-1.635	0.059	0.077	0.033
martinez	0.091	197	-2.297	-2.210	0.008	0.047	0.041
mountain view	0.529	255	0.118	0.011	0.029	-0.232	-0.353
san leandro	0.128	431	-1.922	-1.911	-0.050	0.040	0.083
san mateo	0.277	880	-0.958	-0.974	0.030	-0.195	-0.150
south san francisco	0.178	258	-1.528	-1.554	0.026	-0.014	-0.007
<new location>				-1.787	0.008	0.007	-0.004



5.4 Supervised Encoding Methods

caution

- (like everything) be careful of over-fitting when using supervised encodings methods
  - “it is strongly recommended that either a different data sets be used to estimate the encodings and the predictive model or that their derivation is conducted inside resampling so that the assessment set can measure the overfitting (if it exists).”