Categorical variables encoding

Kyiv Kaggle Trainings

Motivation

- Many algorithms require numerical feature representation
- If feature is not ordered (age : <20, 20-40, 40>, income: <1000, 1000-5000, 5000>), we should do some transformation, not only label encoding with integer
- Good feature encoding could significantly raise your accuracy

One Hot Encoding

- Every category is a separate column
- Example:
  - Raw:

Id	Quarter	Party
1	Quarter1	Democrat
2	Quarter2	Republican
3	Quarter3	Missing

  - Preprocessed

Id	Quarter1	Quarter2	Quarter3	Quarter4	Democrat	Republican	Missing
1	1	0	0	0	1	0	0
2	0	1	0	0	0	1	0
3	0	0	1	0	0	0	1

Dummy encoding

- Almost the same as OHE, but n-1 features
- My assumption is that dummy encoding better for linear models, OHE – trees (better to try yourself)
- Example:
  - Raw:

Id	Quarter	Party
1	Quarter1	Democrat
2	Quarter2	Republican
3	Quarter3	Missing

  - Preprocessed

Id	Quarter1	Quarter2	Quarter3	Democrat	Republican
1	1	0	0	1	0
2	0	1	0	0	1
3	0	0	1	0	0

Binary encoding

- Steps:
  - 1) Labels -> integers
  - 2) Integers -> binary numbers
  - 3) Each digit is a separate feature
- Example:
  - Raw:

Id	Quarter	Party
1	Quarter1	Democrat
2	Quarter2	Republican
3	Quarter3	Missing

  - Preprocessed

Id	Quarter_1	Quarter_2	Party_1	Party_2
1	0	0	0	0
2	0	1	1	0
3	1	0	0	1

Information Value and Weight of Evidence

- Transformation is used for binary classification or you should do 'one against all' and repeat it for each class
- Steps:
  - 1) Calculate conditional probability for each category
  - 2) $WOE = \log \frac{P(good)}{P(bad)}$
  - 3) $IV = \sum_{i=0}^{n-1} (P_i(good) - P_i(bad)) \bullet WOE$

Age Group	Total Number of loans	Number of Bad loans	Number of Good Loans	% Bad loans	Name of Coarse Groups	Distribution of loans	Distribution Bad (DB)	Distribution Good (DG)	WOE	DG - DB	(DG -DB) *WOE
21-30	4821	206	4615	4.3%	G1	0.079	0.135	0.078	-0.553	-0.057	0.0318
30-36	10266	357	9909	3.5%	G2	0.169	0.235	0.167	-0.339	-0.067	0.0228
36-48	32926	776	32150	2.4%	G3	0.542	0.510	0.542	0.062	0.032	0.0020
48-60	12788	183	12605	1.4%	G4	0.210	0.120	0.213	0.570	0.092	0.0527
Total	60801	1522	59279							Information Value -->	0.1093

Combination of factors

- Combining features ☺ and then do some encoding
- Example
  - Raw

Id	Quarter	Party
1	Quarter1	Democrat
2	Quarter2	Republican
3	Quarter3	Missing

  - Preprocessed

Id	Quarter_Party
1	Quarter1_Democrat
2	Quarter2_Republican
3	Quarter3_Missing

Count Featurizer

- Microsoft algorithm
- Description https://phvu.net/2016/05/13/count-featurizer/

Mean Target

- Calculate mean target value for each factor

Weighted Mean Target

$$S_i = \lambda(n_i)\frac{n_{iY}}{n_i} + (1-\lambda(n_i))\frac{n_Y}{n_{TR}}$$

- λ is a monotonically increasing function on n_i bounded between 0 and 1

Stas Semenov approach

- Formula: (mean value * category size + global mean value * C) / (category size + C)
- Works almost always
- Task: Optimize C parameter

My realization

- GitHub opendatascience : https://github.com/open-data-science/datascience-swiss-knife/blob/master/handling%20categorical%20variables/cat_var_encoding.ipynb

Literature

1. https://www.kaggle.com/c/caterpillar-tube-pricing/forums/t/15748/strategies-to-encode-categorical-variables-with-many-categories
2. http://helios.mm.di.uoa.gr/~rouvas/ssi/sigkdd/sigkdd.vol3.1/barreca.ps
3. https://www.kaggle.com/rsakata/bnp-paribas-cardif-claims-management/xgboost-with-combination-of-factors/code
4. https://dato.com/learn/userguide/feature-engineering/count_featurizer.html
5. http://ucanalytics.com/blogs/information-value-and-weight-of-evidencebanking-case/
6. http://patsy.readthedocs.io/en/latest/categorical-coding.html
7. https://www.kaggle.com/forums/f/15/kaggle-forum/t/16927/how-to-deal-with-features-having-high-cardinality
8. https://www.quora.com/What-are-the-best-practices-for-coding-a-categorical-variable-for-a-linear-regression
9. http://www.kdnuggets.com/2015/12/beyond-one-hot-exploration-categorical-variables.html
10. http://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

Thank you