Click Through Rate (CTR) prediction for digital ads under high cardinality

Background & Motivation

- CTR prediction essential to bid price for ads
- CTR $= \frac{\#\ user\ clicks}{\#\ impressions}$
- Impression: every time ad is shown on internet
- Digital ads data exhibits high cardinality; One Hot Encoding leads to high dimensionality, requiring increased memory and computation
- Determine optimal encoding for high cardinal categorical variables

Research Goals

- Which encoding techniques are optimal for categorical variables?
- Should optimal encoding be used only for high cardinal features, leaving low cardinal features to one hot encoding?
- Which features are most useful in predicting CTR for digital ads?

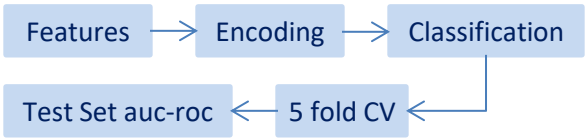
Dataset Familiarity

- CTR data for 10 days provided by Avazu, online advertising company, on Kaggle
- All features categorical, representing attributes for users, ads and context
- Imbalanced dataset, 17% of ads clicked
- Eight features anonymized by Avazu, for privacy
- 120k records randomly sampled from 40 million records, due to computational challenge

Feature	Cardinality	Feature	Cardinality
Device IP	79,135	App ID	1,339
Device ID	17,368	App Domain	97
Device Model	3,198	App Category	23
Site ID	1,487	Site Category	20
Site Domain	1,351	Banner Position	7

Methodology

- Examine effect of 8 encoding techniques on 3 classification models



- Due to imbalanced data, observation weights inversely proportion to target frequency
- 2 sets of experiment; Encoding for all features and only high cardinal features (cardinality>25)
- Inference via optimal encoding & classification

Encoding Techniques

- One Hot: expand into dummy columns of {0,1}
- Ordinal: replace levels by ordinal number
- Binary: split binary representation of ordinal
- Frequency: group less frequent levels
- Target: replace level by mean/cond. probability
- Hash: Level into integer using hashing function
- Quantile (50%): replace level by target median
- Weight of Evidence (WoE): replace level by conditional log(odds)

Results

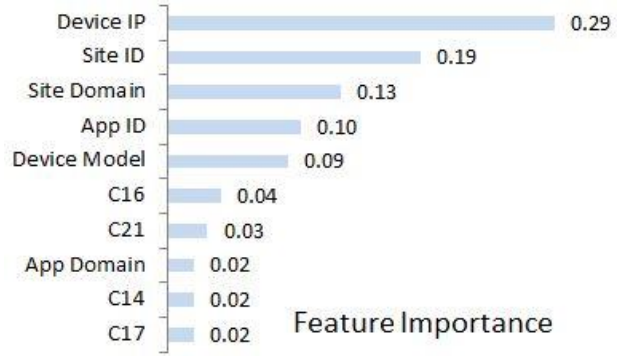
High Carinality Threshold = 0

Encoding/Model	Gradient Boosting	Logistic Regression	Random Forest
Binary Encoding	0.72	0.67	0.71
Frequency Encoding	0.70	0.63	0.71
Frequency Encoding 01	0.71	0.64	0.70
Frequency Encoding 10	0.71	0.63	0.70
Hashing Encoding	0.71	0.65	0.70
One Hot Encoding	0.70	0.72	0.67
Ordinal Encoding	0.70	0.62	0.70
Quantile Encoding	0.62	0.61	0.62
Target Encoding	0.69	0.64	0.69
WoE Encoding	0.71	0.71	0.72

Results (Cont'd)

High Carinality Threshold = 25

Encoding/Model	Gradient Boosting	Logistic Regression	Random Forest
Binary Encoding	0.72	0.68	0.70
Frequency Encoding	0.70	0.66	0.70
Frequency Encoding 01	0.71	0.66	0.70
Frequency Encoding 10	0.71	0.66	0.69
Hashing Encoding	0.71	0.67	0.70
One Hot Encoding	0.70	0.72	0.68
Ordinal Encoding	0.69	0.65	0.69
Quantile Encoding	0.64	0.62	0.65
Target Encoding	0.69	0.64	0.70
WoE Encoding	0.71	0.71	0.72



Conclusion

- WoE encoding is optimal for low and high cardinal variables
- Domain and identifier for website/app, user device IP/model are most useful to predict CTR
- Optimal encoding should be model/data agnostic; extend research to new models/data
- Future work includes optimal encoding for each feature and deep learning based encodings

Duke University

Duke TRINITY COLLEGE OF ARTS AND SCIENCES | DEPARTMENT of STATISTICAL SCIENCE