**Abstract**

The data generated by online advertising industry for digital ads contains many categorical variables. Predicting click through rate (CTR) for digital ads is vital for efficiency of ads marketplace (real time bidding for digital ads) and often involves dealing with these categorical variables. Since, most of machine learning algorithms are designed to work with numerical data; we need efficient numerical encoding of these categorical variables. A common challenge in online advertising is the presence of high cardinality in these categorical variables. This report attempts to identify efficient encoding techniques for categorical variables by studying their impact on predictive ability of classification algorithms for CTR. The experiment involves eight encoding techniques (one hot, ordinal, binary, frequency, target, hash, quantile, weight of evidence) and three classification algorithms (logistic regression, gradient boosted trees and random forest). The results from the analysis yield that weight of evidence encoding yields the best auc-roc for predicting CTR and thus works optimal for digital ads data. Moreover, Weight of Evidence (WoE) encoding can be used optimally for both low and high cardinal variables. Additionally, user device and website/app attributes such as device IP, device model, site id, app id, site domain and app domain play a prominent role in predicting CTR.

**1. Introduction**

**1.1 Significance of predicting CTR**

This project aims to tackle practical challenges associated with analyzing and predicting click through rate (CTR) for digital advertisements. CTR is an industry-wide used metric in online advertising to assess performance of digital ads. In online advertising industry, a digital ad is a transaction between two parties; advertisers (who want to promote/sell a product) and publishers (who allow their website, mobile app, search engine etc. to publish the advertisement). The advertisers have the option to pay fee to the publishers either per impression (every time an ad is shown to the user) or per click (every time a user clicks on the ad). Majority of advertisers prefer to pay per click basis and bid for the ads on real time auction bidding platform provided by publishers. Thus, it is critically important for involved parties (publishers and all advertisers) to accurately estimate probability of an ad being clicked (CTR) for a fair price transaction on the bidding platform.

**1.2 Key Challenges**

One of the key challenges in predicting CTR is the presence of high cardinality (number of unique levels) categorical variables. A conventional CTR dataset for digital ads contains attributes for users, ads and context. Very often these attributes are categorical in nature and possess high cardinality. For instance, identification variables for website or mobile app, user device id/model may exhibit multitude of values leading to issues of high cardinality. Also, the context attributes such as location, country etc. and ad attributes such as banner position etc. are categorical in nature. Since most machine learning algorithms are designed to work with numerical data, it’s crucial to find appropriate numerical representations of categorical variables. One Hot Encoding is a popular encoding technique for categorical variables. However, it leads to high dimensional data which further requires increased memory and computation in presence of high cardinal variables. To address these challenges, the report examines eight numerical encoding techniques (one hot, ordinal, binary, frequency, target, hash, quantile and weight of evidence) for categorical variables. The effect of these encoding techniques on predictive ability of various machine learning classification models is studied to determine optimal encoding technique for high cardinal variables.

**1.3 Research Goals**

The report addresses following three research questions to deal with high cardinal variables and predict CTR for digital ads:

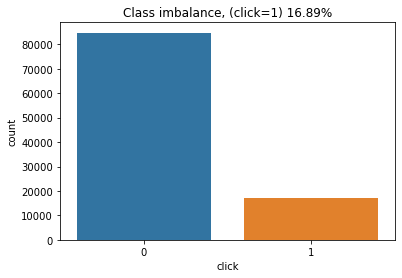
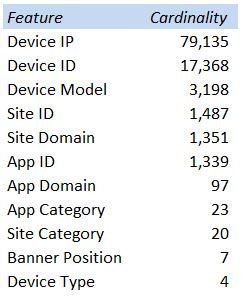
1) Which encoding techniques are optimal for digital advertising data?

2) Should the optimal encoding techniques be used for only high cardinal data and default to One Hot Encoding for low cardinal variables?

3) Using the optimal encodings for categorical variables, which features are most helpful in predicting CTR for digital ads?

**1.4 Description of digital ads dataset**

The dataset used in this analysis has been made available on Kaggle by online advertising company Avazu. The dataset contains 10 days history of CTR on digital ads along with associated categorical variables. All the variables in dataset are categorical variables which represent attributes about users, ads and context. A binary variable ‘click’ taking values in {0,1} signifies if the user clicked on ad. Other attributes include identifiers for banner position of ad, website id/domain/category, advertisement id/domain/category, and device id/IP/model/type. In addition, there are other eight categorical features (C14-C21) which are anonymized by Avazu for data privacy purpose. It is worth highlighting that the original dataset contained records for 40 million ads. Processing such huge dataset caused Python to run out of memory and would require big data tools such as Spark. So to implement the project in Python, 120000 records were randomly sampled from the original data. The imbalance in clicks and cardinality (number of unique levels) of few selected categorical variables is mentioned below for reference:

**2. Methodology**

In order to answer first research goal and determine optimal encoding for categorical variables, the report considers eight encoding techniques and their effect on performance of three classification models. The inherent assumption is that the optimal encoding technique would perform well for all classification models. For each combination of encoding and classification model, hyper-parameters of classification model are tuned using 5 fold cross validation on training set (85% of data) and then auc-roc is computed on test set (15% of data). The encodings are compared based on auc-roc of three classification models on test set. The aim is to find the encoding which achieves high CTR prediction auc-roc on test set across most of classification models. To answer second research goal, the aforementioned exercise is carried out on two sets of training data; (i) the encoding technique in question applied to all variables (ii) the encoding technique in question applied to high cardinal variables (cardinality>=25) and one hot encoding applied to low cardinal variables (cardinality<25). To answer the third research question, the classification model with highest auc-roc on test set using the optimal encoding technique is selected. This optimal model is inspected for inference on features. A detailed description of encoding techniques and classification models considered is mentioned below.

**2.1 Encoding Techniques**

(a) One Hot Encoding: One hot encoding, dummy encoding or indicator encoding is the most commonly used encoding technique for categorical variables. It expands a categorical variable into dummy columns taking values in {0, 1} to indicate whether or not that dummy column is true for that observation. Each dummy column corresponds to one of levels of the categorical variable. If the cardinality of categorical variable is high, it leads to large sparse datasets.

(b) Ordinal Encoding: This replaces the levels of categorical variable by ordinal number. It’s easy to implement and doesn’t lead to sparse datasets unlike one hot encoding. However, it introduces arbitrary order among levels of categorical variable which doesn’t make sense practically.

(c) Binary Encoding: First, the levels of categorical variable are converted to ordinal numbers. Then the ordinal numbers are converted to their binary representations which are then split into different columns. This can be particularly helpful where we don’t want to use one hot encoding on high cardinality categorical variable as it leads to fewer dimensions than one hot encoding.

(d) Frequency Encoding: Frequency encoding maps each level of categorical variable to its relative frequency in the dataset. The inherent assumption being that high frequency of a level is associated with target. This approach is similar to n-grams approach in Natural Language Processing to encode tokens based on relative frequency.

(e) Target Encoding: This encodes each level of categorical variable by its conditional target mean (if target is continuous) or conditional target probability (is target is binary).

(f) Hash Encoding: The hash encoding first converts each level of a categorical variable into an integer using a hash function. This integer is transformed to indicator representation, similar to binary encoding, using a given hash size. This can lead to collision where multiple levels of categorical variable can map to one indicator representation. Smaller hash size leads to more collision unlike larger hash size and can lead to worse performance.

(g) Quantile (50%) Encoding: This approach is similar to target encoding except that instead of target mean, target median is encoded for each level of categorical variable.

(h) Weight of Evidence (WoE) Encoding: It is similar to target encoding for binary target. Instead of encoding mean probability of target like target encoding, WoE encodes log(odds) of binary target for each level of categorical variable. This is particularly useful for imbalanced target datasets.

**2.2 Classification Models**

This report examines effect of encoding techniques on predictive ability of three classification models:

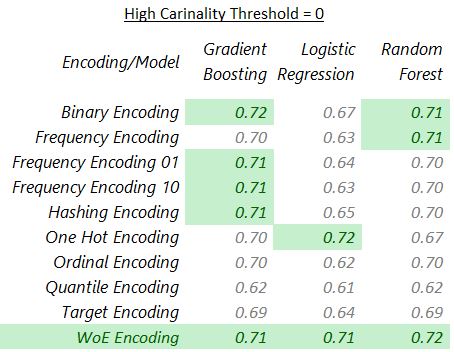
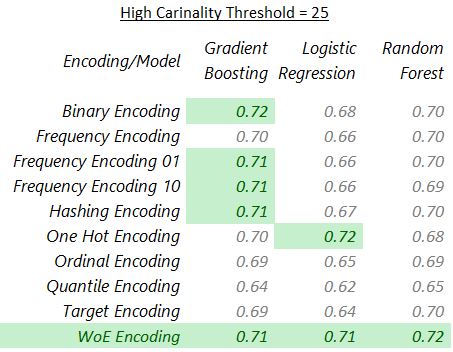
(a) Regularized Logistic Regression: The logistic regression with L1 penalty is used with all the above encodings to predict CTR. It maps linear combination in encoded categorical variables to CTR probability through logit link function. The L1 regularization helps feature selection and model fit simultaneously.

(b) Gradient Boosted Trees: These are state of the art ensemble methods which combine several weak learners into an ensemble for better prediction. It iteratively fits many trees and up-weights/down-weights incorrectly/correctly predicted observations from previous iteration, reducing model bias.

(b) Random Forest: These is an ensemble learning method which averages out the prediction from many decision trees. Each decision tree is built on a dataset which is sampled with repetition from original dataset. The predictions from these decision trees are averaged, reducing variance of model prediction.

**3. Results**

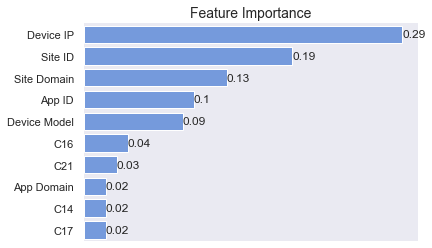
The following table shows auc-roc on test set for various encodings and classification models:

1) It is clearly evident that weight of evidence encoding performs best across all the classification models considered, making it the optimal choice atleast for digital ads dataset.

2) There is no incremental benefit to using one hot encoding for low cardinal variables and different encoding for high cardinal variables. Weight of Evidence encoding can be used optimally for low and high cardinal variables.

3) Random Forest outperforms gradient boosting and logistic regression when used with weight of evidence encoding. Random forest model with tuned hyper-parameters on 5 fold cross validation using weight of evidence encoding achieves highest (72%) auc-roc on test set. The feature importance (quantified using average decrease in tree impurity when split on the feature) for top 10 features using the optimal random forest model is shown below.

****

**4. Discussion**

The optimal encoding for digital ads dataset is Weight of Evidence (WoE) encoding as classification models can achieve highest prediction auc-roc using this encoding without transforming data into very high dimensions unlike one hot encoding. WoE encoding can be used for low cardinal variables in addition to high cardinal ones. The most important features to predict CTR are user attributes such as device IP/model and website or app attributes such as site/app id, site/app domain. Ideally, the optimal encoding should be model and dataset agnostic. In this report, the optimal encodings are determined with respect to three classification models and a specific CTR dataset. So to practically accommodate this, this study needs to expand to other classification techniques and datasets. Additionally, the report attempts to find an optimal encoding for the entire dataset which applies to all the variables. This work can be extended to determine different optimal encoding techniques for different features. Further, encoding using deep learning based embedding can be explored as it is widely used in the industry.

**5. References**

1) Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features: [Link](https://arxiv.org/abs/2104.00629)

2) Modern models for learning large-scale highly skewed online advertising data: [Link](https://escholarship.org/content/qt7mc0k1v8/qt7mc0k1v8_noSplash_68fe9a60ebefb849e7eb26b341d072b0.pdf)