Experiment no: 8

Title:
Perform encoding categorial feature on given dataset

Theory:
overview: it understand what is categorial data encoding
it have different encoding techniques and when to use them

Introduction:
The performance of M.L model not only depends on model and hyperparameters, but also on how we process and feed different types of variables to model. Since most M.L models only accept numerical variables preprocessing categorial becomes an necessary step. We need to convert these categorial variables to number such that model is able to understand and extract valuable information.

Content: 1) What is categorial data?
Since we are going to be working on categorial variable in this article, here is a quick refresher on basic with couple of example. categorial variables are usually represented as strings or categories and are finite in number. Here are few example

ordinal data: The categories have an inherent order.
Nominal data: The categories do not have an inherent order.

2) Label encoding or ordinal encoding:
We use this categorial data encoding technique

when categorical feature is ordinal. In this case ordering order is important. Hence encoding should reflect sequence. In label encoding each label is converted into integer value. We will create variable that contains categories representing education qualification of person

* one Hot encoding:

We use this categorical data encoding technique when features are nominal. In one hot encoding for each level of a categorical feature, we create a new variable. Each category is mapped with a binary variable containing either 0 to 1. Here, 0 represent absence and 1 represents presence of that category.

These newly created binary feature are known as dummy variables. The number of dummy variables depends on levels present in categorical variable.

* Dummy encoding:

dummy coding scheme is similar to one-hot encoding. categorical data encoding method transforms categorical variable into a set of binary variable. In case of one-hot encoding, for N-categories in a variable it uses N binary variable. The dummy encoding is small improvement over one hot encoding. Dummy encoding uses N-1 features to represent N labels categories.

2

* Drawbacks of one Hot and dummy encoding.

One hot encoder and dummy encoder are two powerful and effective encoding schemes. They are also very pop-ular among data scientist but may not be as effective when.

i) A large number of level are present in data. If there are multiple categories in feature variable in such a case we need similar number of dummy variables to encode data.
ii) If we have multiple categorical features in dataset similar situations will occur & again we will ends to have several dummy features each representing.

categorical feature and their multiple categories.

In both above cases, these two encoding schemes introduce sparsity in dataset several columns having 0's & one of them having 1s. In other words it creates multiple dummy feature in dataset without adding much information

* effect encoding:

This encoding technique is also known as deviation encoding can term encoding effect encoding is almost similar to dummy encoding, with little difference

The row containing only 0's in dummy encoding as -1 in effect encoding. In dummy encoding example, city Banglore at index 4 was encoded as 0000. whereas in effect coding it is represented -1 -1 -1 -1.

Effect encoding is an advanced technique.

* Hash encoder :-

To understand Hash encoding it is necessary to know about hashing is transformation of arbitrary size input in form of fixed size value. We use hashing algorithms to perform hashing operation further hashing is one way process

Hashing has several applications like data retrieval check data corruption and in data encryption also we have multiple hash function available.

* Binary encoding :-

Binary encoding is combination of hash encoding and one hot encoding. In this encoding scheme categorical feature is first convert into numerical using ordinal encoder.

Binary encoding is memory efficient encoding scheme as it uses fewer features than one hot encoding further it reduces curse of dimensionality for data with high cardinality

* Target encoding :

Target encoding is bayesian encoding. Bayesian encoders use information from dependent variable to encode categorical data

In target encoding we calculate mean of target

category and replace category variable with mean value. In case categorical target variable posterior probability replaces each category.

We perform target encoding for train data only & code test data using results obtaining from training dataset. Although very efficient coding system, it has following issues responsible for diteriorating model performance.

I) It can lead to target leakage or overfitting. To address overfitting we can use different techniques

1) In leave one out encoding, current target value is reduced from overall mean of target to avoid leakage.

2) In another method we may introduce expression noise in target statistic value of this nouse.

II) The second issue, we may face is improper distribution of categories in train & test data. In such case, categories may assume extreme values.

conclusion:

To summarize encoding categorical data is one unavoidable part of feature engineering. It is very imp. to know what encoding scheme we should use. Having and considerable dataset to use. In this practical, we have seen various encoding techniques along with their issues & suitable use cases.