Encoding Categorical Columns with lots of Unique values for Regression

This problem assumes that you have knowledge of regression. You have a dataset which has a numeric column named “Y” (the target column) as well as other numeric and/or categorical columns which are the predictors (also called as “X” columns). But what do you do when one of the X columns is a categorical column which has lots of unique values e.g. a column named “pin\_code” that has 15000 unique values? Categorical column are usually encoded using one-hot-encoding in regression- but that is not the best way to encode a column like pin\_code in regression. You will end up with 14999 new encoded columns.

Your problem is to find an encoding for columns like pin code in regression.

Write a function which takes as input:

An input dataset for regression where the first column is target column and the remaining columns as X.

Some X columns are numeric columns, and the remaining are categorical columns.

An integer I which refers to the column number in the data.

Returns:

If column I is not categorical, return nothing.

If column I is categorical, returns an encoding of the column. The encoded output will have the same number of rows as input data but can have one or more columns (e.g. one hot encoding on 10 unique values gives 9 columns with binary values).

Note: You are expected to provide an encoding strategy when a column might have lots of unique values.

Solution:

Here I have used the category\_encoders library which can be installed as follows:

Pip install category\_encoders

Conda install category\_encoders