Welcome to ExamTopics

EXAMTOPICS

- Expert Verified, Online, **Free**.

MENU

Google Discussions

Exam Professional Machine Learning Engineer All Questions

View all questions & answers for the Professional Machine Learning Engineer exam

Go to Exam

EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 152 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 152

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are developing an ML model using a dataset with categorical input variables. You have randomly split half of the data into training and test sets. After applying one-hot encoding on the categorical variables in the training set, you discover that one categorical variable is missing from the test set. What should you do?

A. Use sparse representation in the test set.

B. Randomly redistribute the data, with 70% for the training set and 30% for the test set

C. Apply one-hot encoding on the categorical variables in the test data

D. Collect more data representing all categories

Show Suggested Answer

by TNT87 at *Feb. 16, 2023, 9 a.m.*

Comments

Type your comment...

Submit

baimus 1 month, 1 week ago

Selected Answer: C

I've very grudgingly ticket C, as the question is missing "handle the missing category by one hot encoding all zeros for the missing feature column". It otherwise doesn't make sense as will have the wrong amount of entries.

upvoted 1 times

fitri001 6 months ago

Selected Answer: C

The correct approach is to handle the missing category during one-hot encoding of the test data. Here's how to address this issue:

Identify the Missing Category: After applying one-hot encoding to the training set, compare the categories (unique values) present in the training data with the categories in the test data. This will reveal the missing category.

Add a Column for the Missing Category in the Test Data: Include a new column in the test data specifically for the missing category. Initialize the values in this column with 0.

Apply One-Hot Encoding to the Test Data: Now that the test data includes a column for the missing category, proceed with one-hot encoding the categorical variables in the test data. This will ensure the test data has the same structure as the encoded training data.

upvoted 2 times

baimus 1 month, 1 week ago

But your description includes a missing critical step that the question is missing to make it make sense.

upvoted 1 times

CHARLIE2108 8 months, 2 weeks ago

Selected Answer: C

Answer C

upvoted 1 times

Nxtgen 1 year, 2 months ago

Selected Answer: C

Answer options analysis:

C. Since one categorical variable is missing from the test set, (As I understand: “a categorical variable is in the test but not in train”) apply one hot encoding (trained with the train set?) to the test set, for the variables not present in train we just would obtain an array of all 0’s, so that would be OK.
D. That data collection could be not feasible depending on the real-world-problem.
B. Randomness would not always fix the problem.
A. Not recommended to use different representations for train/test. Sparse representation doesn't magically recover missing categories; it's a way to efficiently store data with a large number of zeros.

I would go with C.

upvoted 3 times

SamuelTsch 1 year, 3 months ago

Selected Answer: C

C but not really sure

upvoted 1 times

Scipione_ 1 year, 5 months ago

Selected Answer: C

You must apply one hot enconding alsto for the test dataset. However, i find this answer incomplete.

upvoted 2 times

baimus 1 month, 1 week ago

Yeah 100% - it's missing the "but make sure it deals with the missing category by adding a "missing" or something to it so the one hot representation has the right number of items.

upvoted 1 times

nescafe7 1 year, 5 months ago

Selected Answer: D

Add data to the test set to get the same OHE

upvoted 2 times

tavva_prudhvi 1 year, 3 months ago

Option D (collecting more data) may not be feasible or necessary if the missing category is not significant or if one-hot encoding is sufficient to handle it.

upvoted 2 times

M25 1 year, 5 months ago

Selected Answer: B

“Rows are selected for a data split randomly, but deterministically. (…) Training a new model with the same training data results in the same data split.” https://cloud.google.com/vertex-ai/docs/tabular-data/data-splits#classification-random. “Randomly redistribute data” [Option B] with different fractions, will result in a different data split. Having a higher fraction split of 70% for the training set will additionally help the model to better generalize (compared to only 50%), thus perform better when testing, the ultimate goal.

upvoted 2 times

maukaba 12 months ago

https://cloud.google.com/vertex-ai/docs/tabular-data/data-splits#classification-random
I think it's applicable to VertexAI auto ML only

upvoted 1 times

M25 1 year, 5 months ago

Sparse representation is one “in which only nonzero values are stored”, excluding [Option A]: https://developers.google.com/machine-learning/crash-course/representation/feature-engineering#sparse-representation. Applying “one-hot encoding” to the columns will not help finding the missing column, thus excluding [Option C]. No indication provided for a need to “collect more data”, excluding [Option D].

upvoted 1 times

julliet 1 year, 5 months ago

it is possible that category is very rare and that is the reason we don't have it in the test. So I guess we should just apply the train data transformations and use one-hot

upvoted 2 times

Gudwin 1 year, 5 months ago

Selected Answer: C

By using a sparse representation, you will be losing the information contained in the missing categorical variable. This could lead to the model making incorrect predictions on the test set.

upvoted 2 times

wrosengren 1 year, 6 months ago

I agree with formazioneQl that if a different one hot encoding is used for the test set compared to the train set then the results would be poor. However, there is no problem with not having all combinations in the test set if all possibilities are present in the training set. So assuming that we are using the same mapping of data in the train and test set, I would vote C. If we don't encode the test set, the variable is meaningless anyways. So I would lean C.

upvoted 1 times

formazioneQI 1 year, 6 months ago

Selected Answer: A

Since one categorical variable is missing from the test set, C would result in a different number of columns in the training and test sets.

upvoted 3 times

tavva_prudhvi 1 year, 3 months ago

Option A (sparse representation) may not work well in this case, as it can lead to sparsity issues and affect the model's performance.

upvoted 1 times

TNT87 1 year, 7 months ago

C. Apply one-hot encoding on the categorical variables in the test data.

When using one-hot encoding on categorical variables, each unique value of the variable is represented as a separate binary variable. Therefore, it is important to ensure that the same set of binary variables is present in both the training and test datasets. Since one categorical variable is missing in the test set, the recommended approach is to apply one-hot encoding on the categorical variables in the test set to ensure that the same set of binary variables is present in both datasets.

upvoted 2 times

TNT87 1 year, 8 months ago

Selected Answer: C

Answer C

upvoted 1 times