EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 18 DISCUSSIO..

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 18
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You work for a large hotel chain and have been asked to assist the marketing team in gathering predictions for a targeted marketing strategy. You need to make predictions about user lifetime value (LTV) over the next 20 days so that marketing can be adjusted accordingly. The customer dataset is in BigQuery, and you are preparing the tabular data for training with AutoML Tables. This data has a time signal that is spread across multiple columns. How should you ensure that AutoML fits the best model to your data?

A. Manually combine all columns that contain a time signal into an array. Allow AutoML to interpret this array appropriately. Choose an automatic data split across the training, validation, and testing sets.

B. Submit the data for training without performing any manual transformations. Allow AutoML to handle the appropriate transformations. Choose an automatic data split across the training, validation, and testing sets.

C. Submit the data for training without performing any manual transformations, and indicate an appropriate column as the Time column. Allow AutoML to split your data based on the time signal provided, and reserve the more recent data for the validation and testing sets.

D. Submit the data for training without performing any manual transformations. Use the columns that have a time signal to manually split your data. Ensure that the data in your validation set is from 30 days after the data in your training set and that the data in your testing sets from 30 days after your validation set.

Show Suggested Answer

by gcp2021go at *July 1, 2021, 2:50 p.m.*

Comments

Type your comment...

Submit

kkd14 Highly Voted 3 years, 3 months ago

Should be D. As time signal that is spread across multiple columns so manual split is required.

upvoted 24 times

sensev 3 years, 2 months ago

Also think it is D, since it mentioned that the time signal is spread across multiple columns.

upvoted 4 times

GogoG 3 years ago

Correct answer is C - AutoML handles training, validation, test splits automatically for you when you specify a Time column. There is no requirement to do this manually.

upvoted 5 times

george_ognyanov 2 years, 12 months ago

Correct answer is D. It clearly says the time signal data is spread across different columns. If it weren't then C would be correct and your point would be valid. However, in this case the answer is D 100%.

https://cloud.google.com/automl-tables/docs/data-best-practices#time

upvoted 9 times

irumata 2 years, 9 months ago

this comment is only about time information in different columns, not about time itself. C is correct as for me

upvoted 1 times

irumata 2 years, 9 months ago

but if time signal means time mark not the business signal the D is the correct - very controversial

upvoted 1 times

Load full discussion...

Werner123 7 months, 4 weeks ago

I think the answer is C. In this case I am interpreting time signal as the features that hold predictive power as a function of time i.e. time signal. There is no indication to how much data is available so using the 30 days after mark is not wise. You only have 30 days worth of data for validation set. If you have a few years worth of data this seems like a unnecessary small validation set.

upvoted 4 times

DucLee3110 Highly Voted 3 years, 3 months ago

C
You use the Time column to tell AutoML Tables that time matters for your data; it is not randomly distributed over time. When you specify the Time column, AutoML Tables use the earliest 80% of the rows for training, the next 10% of rows for validation, and the latest 10% of rows for testing.
AutoML Tables treats each row as an independent and identically distributed training example; setting the Time column does not change this. The Time column is used only to split the data set.
You must include a value for the Time column for every row in your dataset. Make sure that the Time column has enough distinct values, so that the evaluation and test sets are non-empty. Usually, having at least 20 distinct values should be sufficient.
https://cloud.google.com/automl-tables/docs/prepare#time

upvoted 14 times

salsabilsf 3 years, 2 months ago

From the link you provided, I think it's A :

The Time column must have a data type of Timestamp.

During schema review, you select this column as the Time column. (In the API, you use the timeColumnSpecId field.) This selection takes effect only if you have not specified the data split column.

If you have a time-related column that you do not want to use to split your data, set the data type for that column to Timestamp but do not set it as the Time column.

upvoted 2 times

Dirtie_Sinkie Most Recent 3 weeks, 1 day ago

D could work, but I'm still leaning towards C

upvoted 1 times

nktyagi 2 months, 2 weeks ago

Selected Answer: C

AutoML handles training, validation, test splits automatically for you when you specify a Time column. There is no requirement to do this manually.

upvoted 1 times

PhilipKoku 4 months, 2 weeks ago

Selected Answer: D

D)D is correct, as this would satisfy the days criteria mentioned in the question. 30 days is more than 20 days, and the prediction model can be used on a validation dataset to validate the results for the next 20 days.

upvoted 1 times

guilhermebutzke 9 months ago

Selected Answer: D

thinking that "spread across multiple columns" seems like "columns with redundant information," and considering how AutoML can deal with correlated columns, I think option C is the best choice, with no need for a manual split.

However, "time information is not contained in a single column" is the same thing as "time signal that is spread across multiple columns." I agree that D could be the best option.

Then, I tend to think that D is the best choice because the text could be more clearly expressed in redundant options.

upvoted 2 times

Mickey321 11 months, 1 week ago

Selected Answer: C

Either C or D but leaning towards C as not get the 30 days in D

upvoted 2 times

Sum_Sum 11 months, 1 week ago

Selected Answer: D

"data has a time signal that is spread across multiple columns" - I interpret as having > 1 timeseries column.
AutoML knows how to deal with a single column but not multiple
hence answer is D

upvoted 2 times

Krish6488 11 months, 2 weeks ago

Selected Answer: C

Since AutoML is good enough to perform the splits, C appears to be the right answer. Moreover, time information across multiple columns which requires manual split as per option D is different from the question's scenario where the time signal is spread across multiple columns which can be hours, months, days, etc. if we can define in AutoML the right time signal column, its enojugh to split the data and pick most recent data as test data and earliest data as test data

upvoted 1 times

atlas_lyon 1 year, 2 months ago

Selected Answer: D

A Wrong, Even if columns are combines into a 1D-array(column), the time signal should be noticed to autoML anyway. Automatic split cannot work since we need more than 20 days history
B Wrong, Without indicating time signal to AutoML, data would leak in (time leakage) in training/validation/test sets
C Wrong, but might be possible if time signal wouldn't have bee spread across multiple columns
D True, because time signal is spread accross multiple columns require to manually split the data. Since we want to predict LTV over the next 20 days, it is necessary to have at least 20 days history between the splits (30 seems okay: 10 days predictions) Validating and testing on the last 2 months seems reasonable for marketing purpose (usually seasonal).

upvoted 2 times

12112 1 year, 3 months ago

Why 30 days after each data sets, even though we need to predict only for 20 days?

upvoted 1 times

Liting 1 year, 3 months ago

Selected Answer: D

Agree with kkd14. D should be the correct answer.

upvoted 1 times

SamuelTsch 1 year, 3 months ago

Samuel1sch 1 year, 3 months ago

Selected Answer: C

As far as I understand, that AutoML table can handle time-signal column full automatically. Thus, I went to C.

upvoted 1 times

M25 1 year, 5 months ago

Selected Answer: D

Went with D

upvoted 1 times

hghdh5454 1 year, 7 months ago

C. Submit the data for training without performing any manual transformations, and indicate an appropriate column as the Time column. Allow AutoML to split your data based on the time signal provided, and reserve the more recent data for the validation and testing sets.

This approach ensures that AutoML can handle the time-based nature of the data properly. By providing the Time column, AutoML can automatically split the data in a way that respects the time-based structure, using more recent data for validation and testing. This approach is especially important for time-series data, as it helps prevent leakage of future information into the training set, ensuring a more accurate and reliable model.

upvoted 1 times

enghabeth 1 year, 8 months ago

Selected Answer: D

https://cloud.google.com/automl-tables/docs/data-best-practices#time

- If the time information is not contained in a single column, you can use a manual data split to use the most recent data as the test data, and the earliest data as the training data.

upvoted 1 times

John_Pongthorn 1 year, 9 months ago

Selected Answer: D

I go with D: https://cloud.google.com/automl-tables/docs/data-best-practices#time
Read it carefully at the last paragraph of the topic: If the time information is not contained in a single column, you can use a manual data split to use the most recent data as the test data, and the earliest data as the training data.

upvoted 1 times

Load full discussion...

We are the biggest and most updated IT certification exam material website.

Using our own resources, we strive to strengthen the IT professionals community for free.