Google Discussions

Exam Professional Data Engineer All Questions

View all questions & answers for the Professional Data Engineer exam

Go to Exam

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 164 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 164

Topic #: 1

[All Professional Data Engineer Questions]

You are working on a linear regression model on BigQuery ML to predict a customer's likelihood of purchasing your company's products. Your model uses a city name variable as a key predictive component. In order to train and serve the model, your data must be organized in columns. You want to prepare your data using the least amount of coding while maintaining the predictable variables. What should you do?

A. Create a new view with BigQuery that does not include a column with city information.

B. Use SQL in BigQuery to transform the state column using a one-hot encoding method, and make each city a column with binary values.

C. Use TensorFlow to create a categorical variable with a vocabulary list. Create the vocabulary file and upload that as part of your model to BigQuery ML.

D. Use Cloud Data Fusion to assign each city to a region that is labeled as 1, 2, 3, 4, or 5, and then use that number to represent the city in the model.

Show Suggested Answer

by AWSandeep at *Sept. 2, 2022, 6:25 p.m.*

Comments

Type your comment...

Submit

cajica Highly Voted 2 years, 2 months ago

Selected Answer: D

If we're rigorous, as we should because it's a professional exam, I think option B is incorrect because it's one-hot-encoding the "state" column, if the answer was "city" column, then I'd go for B. As this is not the case and I do not accept an spelling error like this in an official question, I would go for D.

upvoted 10 times

sergiomujica 1 year, 7 months ago

I think it should say citiy instead of state... it is a typooi in the transcription of the question

upvoted 3 times

knith66 1 year, 9 months ago

you are right, OHE is mentioned for state in option B, but in option B it is also mentioned to use binary conversion for the city column. an additional method can be used which is applicable for the conversion.

upvoted 1 times

cetanx 1 year, 11 months ago

But also for D, assigning each city to a numbered region could lose important information, as cities within the same region might have different characteristics affecting customer purchasing behavior (from Chat GPT).

upvoted 1 times

MaxNRG Highly Voted 1 year, 4 months ago

Selected Answer: B

One-hot encoding is a common technique used to handle categorical data in machine learning. This approach will transform the city name variable into a series of binary columns, one for each city. Each row will have a "1" in the column corresponding to the city it represents and "0" in all other city columns. This method is effective for linear regression models as it enables the model to use city data as a series of numeric, binary variables. BigQuery supports SQL operations that can easily implement one-hot encoding, thus minimizing the amount of coding required and efficiently preparing the data for the model.

upvoted 6 times

MaxNRG 1 year, 4 months ago

A removes the city information completely, losing a key predictive component.

C requires additional coding and infrastructure with TensorFlow and vocabulary files outside of what BigQuery already provides.

D transforms the distinct city values into numeric regions, losing granularity of the city data.

By using SQL within BigQuery to one-hot encode cities into multiple yes/no columns, the city data is maintained and formatted appropriately for the BigQuery ML linear regression model with minimal additional coding. This aligns with the requirements stated in the question.

upvoted 3 times

MaxNRG 1 year, 4 months ago

https://cloud.google.com/bigquery/docs/auto-preprocessing#one_hot_encoding

upvoted 2 times

clouditis Most Recent 4 months, 3 weeks ago

Selected Answer: D

D it is, Question clearly says least amount of coding!

upvoted 1 times

baimus 7 months ago

Selected Answer: B

This is B. It's easier to one hot in bigquery than to do it in datafusion and then import the values back into bigquery.

upvoted 1 times

barnac1es 1 year, 7 months ago

Selected Answer: B

One-Hot Encoding: One-hot encoding is a common technique for handling categorical variables like city names in machine learning models. It transforms categorical data into a binary matrix, where each city becomes a separate column with binary values (0 or 1) indicating the presence or absence of that city.

Least Amount of Coding: One-hot encoding in BigQuery is straightforward and can be accomplished with SQL. You can use SQL expressions to pivot the city names into separate columns and assign binary values based on the city's presence in the original data.

Predictive Power: One-hot encoding retains the predictive power of city information while making it suitable for linear regression models, which require numerical input.

upvoted 4 times

knith66 1 year, 9 months ago

Selected Answer: B

One hot encoding for state and binary values for each city will allow me to choose the B option.

upvoted 3 times

tavva_prudhvi 1 year, 9 months ago

I guess Option D loses the granularity of the city-level information, as multiple cities will be grouped into the same region and represented by the same number. This can result in a loss of important predictive information for your linear regression model.

On the other hand, if we use one-hot encoding to create binary columns for each city. This method preserves the city-level information, allowing the model to capture the unique effects of each city on the likelihood of purchasing your company's products. Additionally, it can be done directly in BigQuery using SQL, which requires less coding and is more efficient.

upvoted 2 times

blathul 1 year, 10 months ago

Selected Answer: B

One-hot encoding is a common technique used to represent categorical variables as binary columns. In this case, you can transform the city variable into multiple binary columns, with each column representing a specific city. This allows you to maintain the predictive city information while organizing the data in columns suitable for training and serving the linear regression model.

By using SQL in BigQuery, you can perform the necessary transformations to implement one-hot encoding.

upvoted 4 times

KC_go_reply 1 year, 10 months ago

Selected Answer: B

- A is wrong since it drops the city which is a key predictor.
- C is wrong since we want to keep it simple, and not use Tensorflow here.
- D is wrong since there is no specific reason to use Data Fusion, and also this encoding here is ordinal, which doesn't make sense for something non-quantitative such as cities - we want one-hot coding instead.

Therefore, B must be the correct answer.

upvoted 3 times

ckanaar 1 year, 7 months ago

It could be argued that a specific reason to use Data Fusion is the minimal coding requirement.

upvoted 1 times

leandrors 1 year, 10 months ago

Selected Answer: D

Cloud Datafusion: least amount of coding

upvoted 4 times

knith66 1 year, 9 months ago

OHE is better that datafusion considering least amount coding

upvoted 1 times

tavva_prudhvi 1 year, 9 months ago

While it's true that Cloud Data Fusion can simplify data integration tasks with a visual interface, it might not be the best choice in this specific scenario as using Cloud Data Fusion to assign each city to a region might result in a loss of important predictive information due to the grouping of cities

upvoted 1 times

vaga1 1 year, 11 months ago

Selected Answer: B

A doesn't include the city column.
C is not low code.
D is not a one hot encoding, but an ordinal one on the city column.

B applies a one hot encoding on the state column and a binary encoding on the city column, which works for me.

upvoted 1 times

mialll 2 years ago

Selected Answer: B

https://cloud.google.com/bigquery/docs/reference/standard-sql/bigqueryml-auto-preprocessing#one_hot_encoding

upvoted 2 times

juliobs 2 years, 1 month ago

Selected Answer: D

D uses the least amount of coding... even if the model is not good.
B encodes the "state", not the "city".

upvoted 4 times

dconesoko 2 years, 4 months ago

Selected Answer: B

Manually bigquery ml does preprocessing for you however if one wants to do a manual processing one can use the ML.ONE_HOT_ENCODER function. It just acts as an analytical funciton.

upvoted 2 times

zellck 2 years, 5 months ago

Selected Answer: B

B is the answer.

https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-auto-preprocessing#one_hot_encoding
One-hot encoding maps each category that a feature has to its own binary feature where 0 represents the absence of the feature and 1 represents the presence (known as a dummy variable) creating N new feature columns where N is the number of unique categories for the feature across the training table.

upvoted 3 times

ovokpus 2 years, 5 months ago

Selected Answer: B

The Cloud Data Fusion method will add unecessary weights to categories with higher value labels, which will skew the model. The best practice for encoding nominal categorical data is to one-hot-encode them into binary values. That is conveniently done in BigQuery:

https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-auto-preprocessing#one_hot_encoding

upvoted 4 times

Atnafu 2 years, 5 months ago

D
Cloud Data Fusion is a fully managed, code-free data integration service that helps users efficiently build and manage ETL/ELT data pipelines.

upvoted 1 times

> **dconesoko** 2 years, 4 months ago
>
> Does it come with an out of the box one hot encoding template ?
>
> upvoted 1 times

Load full discussion...