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 140 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 140
Topic #: 1

[All Professional Machine Learning Engineer Questions]

You work for a retailer that sells clothes to customers around the world. You have been tasked with ensuring that ML models are built in a secure manner. Specifically, you need to protect sensitive customer data that might be used in the models. You have identified four fields containing sensitive data that are being used by your data science team: AGE, IS_EXISTING_CUSTOMER, LATITUDE_LONGITUDE, and SHIRT_SIZE. What should you do with the data before it is made available to the data science team for training purposes?

A. Tokenize all of the fields using hashed dummy values to replace the real values.

B. Use principal component analysis (PCA) to reduce the four sensitive fields to one PCA vector.

C. Coarsen the data by putting AGE into quantiles and rounding LATITUDE_LONGTTUDE into single precision. The other two fields are already as coarse as possible.

D. Remove all sensitive data fields, and ask the data science team to build their models using non-sensitive data.

Show Suggested Answer

by mil_spyro at *Dec. 13, 2022, 6:54 p.m.*

Comments

Type your comment...

Submit

bobjr 5 months ago

Selected Answer: C

The best approach is C. Coarsen the data by putting AGE into quantiles and rounding LATITUDE_LONGITUDE into single precision. The other two fields are already as coarse as possible.

Here's why:

Preserves Utility: Coarsening the data reduces its sensitivity while retaining some of its informational value for modeling. Age quantiles and approximate location can still be useful features for certain types of models.
Minimizes Risk: By removing the exact age and precise location, you significantly reduce the risk of re-identification or misuse of sensitive information.
Practicality: Coarsening is a relatively simple technique to implement and doesn't require complex transformations or additional model training.

pen_spark

upvoted 3 times

pico 11 months, 3 weeks ago

Selected Answer: D

This approach involves not providing the sensitive fields (AGE, IS_EXISTING_CUSTOMER, LATITUDE_LONGITUDE, and SHIRT_SIZE) to the data science team for model training. Instead, the team can focus on building models using non-sensitive data. This helps to mitigate the risk of exposing sensitive customer information during the development and training process.

While options A, B, and C propose different methods of obfuscating or transforming the sensitive data, they may introduce complexities and potential risks. Tokenizing with hashed dummy values (option A) may not be foolproof in terms of security, and PCA (option B) may not effectively retain the necessary information for accurate modeling. Coarsening the data (option C) might still retain some level of identifiable information, and it may not be sufficient for ensuring the privacy of sensitive data.

upvoted 1 times

LFavero 8 months, 1 week ago

why would you remove potential important features from the training?

upvoted 2 times

M25 1 year, 5 months ago

Selected Answer: A

Went with A

upvoted 3 times

TNT87 1 year, 8 months ago

Selected Answer: D

D. Remove all sensitive data fields, and ask the data science team to build their models using non-sensitive data. This is the best approach to protect sensitive customer data. Removing the sensitive fields is the most secure option because it eliminates the risk of any potential data breaches. Tokenizing or coarsening the data may still reveal sensitive information if the hashed dummy values can be reversed or if the coarsening can be used to identify individual customers. PCA can also be a useful technique to reduce dimensionality and protect privacy, but it may not be appropriate in this case because it is not clear how the sensitive fields can be combined into a single PCA vector without losing information.

upvoted 1 times

tavva_prudhvi 1 year, 3 months ago

Removing all sensitive data fields (Option D) would likely limit the effectiveness of the machine learning model, as important predictive variables would be excluded from the training process. It is important to balance privacy considerations with the need to train accurate models that can provide valuable insights and predictions.

upvoted 1 times

pico 11 months, 3 weeks ago

But in option A, Hashing can result in information loss. While the original values are hidden, the hashed values might not retain the same level of information, which can impact the effectiveness of the machine learning models.

upvoted 1 times

Scipione_ 1 year, 8 months ago

Selected Answer: A

B -> possible in general but not suitable in this case since you don't know AGE, IS_EXISTING_CUSTOMER, LATITUDE_LONGITUDE, and SHIRT_SIZE are the first components in PCA.
C -> You are changing data which could be highly correlated with the output

D -> like C explanation

Answer 'A' uses hashing so you encript the data without losing relevant information

upvoted 4 times

ares81 1 year, 10 months ago

Selected Answer: A

Hashing --> A

upvoted 4 times

TNT87 1 year, 10 months ago

Selected Answer: A

Answer A

upvoted 3 times

hiromi 1 year, 10 months ago

Selected Answer: A

A (by experience)

upvoted 3 times

> **hiromi** 1 year, 10 months ago
>
> https://cloud.google.com/blog/products/identity-security/take-charge-of-your-data-how-tokenization-makes-data-usable-without-sacrificing-privacy
>
> upvoted 4 times

mymy9418 1 year, 10 months ago

Selected Answer: A

I think hash should be better

upvoted 2 times

mil_spyro 1 year, 10 months ago

Selected Answer: D

Removing the sensitive data fields is the safest and most effective way to ensure that customer data is not used in the training of your models.

upvoted 3 times

> **hiromi** 1 year, 10 months ago
>
> see https://cloud.google.com/blog/products/identity-security/take-charge-of-your-data-how-tokenization-makes-data-usable-without-sacrificing-privacy
>
> upvoted 1 times

> **tavva_prudhvi** 1 year, 3 months ago
>
> Removing all sensitive data fields (Option D) would likely limit the effectiveness of the machine learning model, as important predictive variables would be excluded from the training process. It is important to balance privacy considerations with the need to train accurate models that can provide valuable insights and predictions.
>
> upvoted 1 times

Social Media

Facebook , Twitter
YouTube , Reddit
Pinterest

EXAMTOPICS

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.

ExamTopics doesn't offer Real Microsoft Exam Questions. ExamTopics doesn't offer Real Amazon Exam Questions. ExamTopics Materials do not contain actual questions and answers from Cisco's Certification Exams.

CFA Institute does not endorse, promote or warrant the accuracy or quality of ExamTopics. CFA® and Chartered Financial Analyst® are registered trademarks owned by CFA Institute.