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 62 DISCUSSIO..

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 62
Topic #: 1
[All Professional Machine Learning Engineer Questions]

While conducting an exploratory analysis of a dataset, you discover that categorical feature A has substantial predictive power, but it is sometimes missing. What should you do?

A. Drop feature A if more than 15% of values are missing. Otherwise, use feature A as-is.

B. Compute the mode of feature A and then use it to replace the missing values in feature A.

C. Replace the missing values with the values of the feature with the highest Pearson correlation with feature A.

D. Add an additional class to categorical feature A for missing values. Create a new binary feature that indicates whether feature A is missing.

Show Suggested Answer

by Vedjha at *Dec. 7, 2022, 11:19 p.m.*

## Comments

Type your comment...

Submit

⊟ wish0035 Highly Voted 1 year, 10 months ago

Selected Answer: D

ans: D

A => no, you don't want to drop a feature with high prediction power.
B => i think this could confuse the model... a better solution could be to fill missing values using an algorithm like Expectation Maximization, but using the mode i think is a bad idea in this case, because if you have a significant number of missing values (for example >10%) this would modify the "predictive power". you don't want to lose predictive power of a feature, just guide the model to learn when to use that feature and when to ignore it.
C => this doesn't make any sense for me. not sure what i would do that.
D => i think this could be a really good approach, and i'm pretty sure it would work pretty well a lot of models. the model would learn that when "is_available_feat_A" == True, then it would use the feature A, but whenever it is missing then it would try to use other features.

upvoted 14 times

frangm23 1 year, 6 months ago

I guess I would go with D, but it confuses me the fact that in option D, it doesn't say that NaN values are replaced (only that there's a new column added) and this could lead to problems like exploding gradients.
Plus, Google encourages to replace missing values. https://developers.google.com/machine-learning/testing-debugging/common/data-errors
Any thoughts on this?

upvoted 2 times

PhilipKoku Most Recent 5 months ago

Selected Answer: D

D) Good approach

upvoted 1 times

MultiCloudIronMan 7 months, 1 week ago

Selected Answer: B

Google encourages filling missing value and using mode is one of the examples given. D only tell the obvious - data is missing!

upvoted 1 times

fragkris 11 months ago

Selected Answer: D

B and D are correct, but I decided to go with D.

upvoted 1 times

Mickey321 11 months, 3 weeks ago

Selected Answer: D

highly predictive

upvoted 1 times

ichbinnoah 11 months, 4 weeks ago

Selected Answer: B

Definitely not D, it does not even solve the problem of NA values.

upvoted 2 times

andresvelasco 1 year, 1 month ago

Options B or D
But isnt there an inconsistency in option D? if you replace missing values with a new category ("missing") why would you haveto create an extra feature?

upvoted 1 times

Liting 1 year, 4 months ago

Selected Answer: D

Agree with wish0035, answer should be D

upvoted 1 times

PST21 1 year, 4 months ago

By creating a new class for the missing values, you explicitly capture the absence of data, which can provide valuable information for predictive modeling. Additionally, creating a binary feature allows the model to distinguish between cases where feature A is present and cases where it is missing, which can be useful for identifying potential patterns or relationships in the data.

upvoted 2 times

amtg 1 year, 4 months ago

Selected Answer: B

By imputing the missing values with the mode (the most frequent value), you retain the original feature's predictive power while handling the missing values

upvoted 1 times

**Scipione_** 1 year, 5 months ago

**Selected Answer: D**

Both B and D are possible, but the correct answer is D because of the feature high predictive power.

upvoted 2 times

**M25** 1 year, 6 months ago

**Selected Answer: D**

Went with D

upvoted 1 times

**tavva_prudhvi** 1 year, 7 months ago

I think, its D.
Option B of imputing the missing values of feature A with the mode of feature A could be a reasonable approach if the mode provides a good representation of the distribution of feature A. However, this method may lead to biased results if the mode is not representative of the missing values. This could be the case if the missing values have a different distribution than the observed values.

Similarly, When a categorical feature has substantial predictive power, it is important not to discard it. Instead, missing values can be handled by adding an additional class for missing values and creating a new binary feature that indicates whether feature A is missing or not. This approach ensures that the predictive power of feature A is retained while accounting for missing values. Computing the mode of feature A and replacing missing values may distort the distribution of the feature and create bias in the analysis. Similarly, replacing missing values with values from another feature may introduce noise and lead to incorrect results.

upvoted 2 times

**BenMS** 1 year, 8 months ago

**Selected Answer: D**

If our objective was to produce a complete dataset then we might use some average value to fill in the gaps (option B) but in this case we want to predict an outcome, so inventing our own data is not going to help in my view.

Option D is the most sensible approach to let the model choose the best features.

upvoted 1 times

**hiromi** 1 year, 10 months ago

**Selected Answer: B**

B
"For categorical variables, we can usually replace missing values with mean, median, or most frequent values"
Dr. Logan Song - Journey to Become a Google Cloud Machine Learning Engineer - Page 48

upvoted 4 times

> **tavva_prudhvi** 12 months ago
>
> While this approach may seem reasonable, it can introduce bias in the dataset by over-representing the mode, especially if the missing values are not missing at random.
>
> upvoted 1 times

**Pancy** 1 year, 10 months ago

B. Because the important feature is already known. By using mode, contribution of other features will not be missed

upvoted 2 times

**ares81** 1 year, 11 months ago

Mode is the way to go for categorical features. B, for me.

upvoted 3 times

Load full discussion...