Welcome to ExamTopics

EXAMTOPICS
- Expert Verified, Online, Free.

MENU

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 48 DISCUSSIO..

Actual exam question from Google's Professional Machine Learning Engineer
Question #: 48
Topic #: 1
[All Professional Machine Learning Engineer Questions]

You started working on a classification problem with time series data and achieved an area under the receiver operating characteristic curve (AUC ROC) value of
99% for training data after just a few experiments. You haven't explored using any sophisticated algorithms or spent any time on hyperparameter tuning. What should your next step be to identify and fix the problem?

A. Address the model overfitting by using a less complex algorithm.

B. Address data leakage by applying nested cross-validation during model training.

C. Address data leakage by removing features highly correlated with the target value.

D. Address the model overfitting by tuning the hyperparameters to reduce the AUC ROC value.

Show Suggested Answer

by Paul_Dirac at *June 27, 2021, 3:10 a.m.*

Comments

Type your comment...

Submit

Paul_Dirac Highly Voted 3 years, 3 months ago

Ans: B (Ref: https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9)
(C) High correlation doesn't mean leakage. The question may suggest target leakage and the defining point of this leakage is the availability of data after the target is available.(https://www.kaggle.com/dansbecker/data-leakage)

upvoted 27 times

Jarek7 1 year, 3 months ago

This ref doesn't explain WHY we should use NCV in this case - it just explains HOW to use NCV when dealing with time series.
Cross-validation, including nested cross-validation, is a powerful tool for model evaluation and hyperparameter tuning, but it does NOT DIRECTLY ADDRESS data leakage. Data leakage refers to a situation where information from the test dataset leaks into the training dataset, causing the model to have an unrealistically high performance. Nested cross-validation can indeed help provide a more accurate estimation of the model's performance on unseen data, but IT DOESN'T SOLVE the underlying issue of data leakage if it's already present.

upvoted 5 times

John_Pongthorn Highly Voted 1 year, 7 months ago

Selected Answer: C

C: this is correct choice 1000000000%
This is data leakage issue on training data
https://cloud.google.com/automl-tables/docs/train#analyze
The question is from this content.
If a column's Correlation with Target value is high, make sure that is expected, and not an indication of target leakage.

Let 's explain on my owner way, sometime the feature used on training data use value to calculate something from target value unintentionally, it result in high correlation with each other.
for instance , you predict stock price by using moving average, MACD , RSI despite the fact that 3 features have been calculated from price (target).

upvoted 8 times

black_scissors 1 year, 4 months ago

I agree. Besides, when a CV is done randomly (not split by the time point) it can make things worse.

upvoted 2 times

Foxy2021 Most Recent 1 week ago

Select answer: C. --reason--- While B (nested cross-validation) helps improve the evaluation process and prevents over-optimistic performance estimates, it doesn't tackle the root cause of data leakage. Data leakage is often caused by features that are too closely tied to the target—in this case, the unusually high AUC suggests that the model is gaining unfair information.

upvoted 1 times

chirag2506 3 months, 4 weeks ago

Selected Answer: B

B is the correct option

upvoted 1 times

PhilipKoku 4 months, 2 weeks ago

Selected Answer: C

C) Is the best answer

upvoted 1 times

girgu 4 months, 4 weeks ago

Selected Answer: C

Nested cross validation will not work for time series data. Time series data require the expanding widow training data set. Seems most likely the issue is high correlation in columns.

upvoted 1 times

AnnaR 5 months, 4 weeks ago

B: correct.
considering c, but why should we remove a feature of highly predictive nature?? for me, this does not explain the problem of overfitting... a highly predictive feature is also useful for good performance evaluated on the test set.
--> Decide for B!

upvoted 2 times

gscharly 6 months ago

Selected Answer: B

agree with Paul_Dirac

upvoted 1 times

b1e8fee 9 months, 4 weeks ago

b7a8fae 9 months, 4 weeks ago

Selected Answer: B

I initially went with B- however after reading this: https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/ I think C is right. Quoted from the link: "Nested cross-validation is an approach to model hyperparameter optimization and model selection that attempts to overcome the problem of overfitting the training dataset.". Overfitting is exactly our problem here. Correlated features in the dataset may be a sign of data leakage, but they are not necessarily.

upvoted 1 times

Sum_Sum 11 months, 1 week ago

Selected Answer: B

I think its B. GPT4 makes a good argument about C:
While this is a valid approach to handling data leakage, it might not be sufficient if the leakage is due to reasons other than high correlation, such as temporal leakage in time-series data.

upvoted 1 times

pico 1 year, 1 month ago

Selected Answer: A

Option A: This option is a reasonable choice. Switching to a less complex algorithm can help reduce overfitting, and using k-fold cross-validation can provide a better estimate of how well the model will generalize to unseen data. It's essential to ensure that the high performance isn't solely due to overfitting.

upvoted 1 times

pico 1 year, 1 month ago

Option B: Nested cross-validation is primarily used to estimate model performance accurately and select the best model hyperparameters. While it's a good practice, it doesn't directly address the overfitting issue. It helps prevent over-optimistic model performance estimates but doesn't necessarily fix the overfitting problem.

Option C: Removing features highly correlated with the target value can be a valid step in feature selection or preprocessing. However, it doesn't directly address the overfitting issue or explain why the model is performing exceptionally well on the training data. It's a separate step from mitigating overfitting.

Option D: This option is incorrect. Tuning hyperparameters should aim to improve model performance on the validation set, not reduce it.

In summary, the most appropriate next step is Option A:

upvoted 2 times

atlas_lyon 1 year, 2 months ago

Selected Answer: B

B: If splits are done chronologically(as it is always advised), Nested CV should work
C: High correlation with target means we have to check if this is strong explanatory power or data leakage. dropping the features won't help us distinguish in those cases but may help reveal independence contribution of remaining features

upvoted 1 times

tavva_prudhvi 1 year, 2 months ago

Selected Answer: B

Option C is a good step to avoid overfitting, but it's not necessarily the best approach to address data leakage.

Data leakage occurs when information from the validation or test data leaks into the training data, leading to overly optimistic performance metrics. In time-series data, it's important to avoid using future information to predict past events.

Removing features highly correlated with the target value may help to reduce overfitting, but it does not necessarily address data leakage.

Therefore, applying nested cross-validation during model training is a better approach to address data leakage in this scenario.

upvoted 2 times

Jarek7 1 year, 3 months ago

Selected Answer: C

https://towardsdatascience.com/avoiding-data-leakage-in-timeseries-101-25ea13fcb15f
Directly says: "Dive straight into the MVP, cross-validate later!"
MVP stands for Minimum Viable Product

upvoted 1 times

Liting 1 year, 3 months ago

Selected Answer: B

Agree with Paul_Dirac. Also it is recommended to use nested-cross-validation to avoid data leakage in time series data.

upvoted 1 times

black_scissors 1 year, 4 months ago

Selected Answer: C

There can be a feature causing data leakage which might have been overlooked. In addition, when cross-validation is done randomly, the leakage can be even bigger.

upvoted 1 times

M25 1 year, 5 months ago

Selected Answer: B

Went with B

upvoted 1 times

Load full discussion...

Start Learning for free

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.

© 2024 ExamTopics
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.