DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-SCIENTIST Q&As

Databricks Certified Professional Data Scientist Exam

Pass Databricks DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-SCIENTIST Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.pass2lead.com/databricks-certified-professional-data-scientist.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Databricks Official Exam Center

QUESTION 1

While working with Netflix the movie rating websites you have developed a recommender system that has produced ratings predictions for your data set that are consistently exactly 1 higher for the user-item pairs in your dataset than the ratings given in the dataset. There are n items in the dataset. What will be the calculated RMSE of your recommender system on the dataset?

A. 1

B. 2

C. 0

D. n/2

Correct Answer: A

Explanation: The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. Basically, the RMSD represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSD is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent. RMSE is calculated as the square root of the mean of the squares of the errors. The error in every case in this example is 1. The square of 1 is 1 The average of n items with value 1 is 1 The square root of 1 is 1 The RMSE is therefore 1

QUESTION 2

What is the probability that the total of two dice will be greater than 8, given that the first die is a 6?

A. 1/3

B. 2/3

C. 1/6

D. 2/6

Correct Answer: B

QUESTION 3

Under which circumstance do you need to implement N-fold cross-validation after creating a regression model?

A. The data is unformatted.

B. There is not enough data to create a test set.

C. There are missing values in the data.

D. There are categorical variables in the model.

Correct Answer: B

QUESTION 4

Consider the following confusion matrix for a data set with 600 out of 11,100 instances positive:

In this case, Precision = 50%, Recall = 83%, Specificity = 95%, and Accuracy = 95%.

Select the correct statement

		Predicted Label	
		Positive	Negative
Known Label	Positive	500	100
	Negative	500	10,000

A. Precision is low, which means the classifier is predicting positives best

B. Precision is low, which means the classifier is predicting positives poorly

C. problem domain has a major impact on the measures that should be used to evaluate a classifier within it

D. 1 and 3

E. 2 and 3

Correct Answer: E

Explanation: In this case, Precision = 50%, Recall = 83%, Specificity = 95%: and Accuracy = 95%. In this case, Precision is low, which means the classifier is predicting positives poorly. However, the three other measures seem to suggest that this is a good classifier. This just goes to show that the problem domain has a major impact on the measures that should be used to evaluate a classifier within it, and that looking at the 4 simple cases presented is not sufficient.

QUESTION 5

In which phase of the analytic lifecycle would you expect to spend most of the project time?

A. Discovery

B. Data preparation

C. Communicate Results

D. Operationalize

Correct Answer: B

In the data preparation phase of the Data Analytics Lifecycle, the data range and distribution can be obtained. If the data is skewed, viewing the logarithm of the data (if it\\'s all positive) can help detect structures that might otherwise be overlooked in a graph with a regular, nonlogarithmic scale. When preparing the data, one should look for signs of dirty data, as explained in the previous section. Examining if the data is unimodal or multimodal will give an idea of how many distinct populations with different behavior patterns might be mixed into the overall population. Many modeling techniques assume that the data follows a normal distribution. Therefore, it is important to know if the available dataset can match that assumption before applying any of those modeling techniques.