Dell Data Scientist and Big Data Analytics Foundations 2023

1. What is an appropriate assignment for a data scientist?

A. Conduct customer surveys

B. Develop predictive models

C. Define an OLAP database schema

D. Monitor key performance indicators

Answer(s): B

---

2. How is dimensionality defined in a "bag of words" document representation?

A. Average number of words per sentence in the document

B. Frequency of repeated words in the document

C. Number of unique terms in the document

D. Total number of words in the document

Answer(s): C

---

3. A call center for a large electronics company handles an average of 35, 000 support calls a day. The head of the call center would like to optimize the staffing of the call center during the rollout of a new product due to recent customer complaints of long wait times.

A. Goal 2 and 4

B. Goals 1, 2, 3, 4

C. Goals 2, 3, 4

D. Goal 1 and 3

Answer(s): A

---

4. During a study to understand the population growth of a certain bacterial culture, you plot the data and identify a quadratic growth trend over time.

A. Square root

B. Add a constant

C. Square

D. Cube

Answer(s): A

---

5. You have been assigned to do a study of the daily revenue effect of a pricing model of online transactions.

A. Aggregate all data to the monthly level in order to create a monthly revenue model.

B. Report back to the business owner that the current data model does not support the business question.

C. Disregard revenue as a driver in the pricing model, and create a daily model based on pricing and transactions only.

D. Interpolate a daily model for revenue from the monthly revenue data.

Answer(s): B

---

6. The web analytics team uses Hadoop to process access logs. They now want to correlate this data with structured user data residing in their massively parallel database.

A. Chukwa

B. Sqoop

C. Pig

D. Scribe

Answer(s): B

---

7. Based on the exhibit, what is a likely issue with the data?

A. Saturated data; indicating potential issues with data definitions

B. Incomplete data; indicating potential issues with data transmission

C. No obvious concerns with the data is visible

D. Mis-scaled data; indicating potential issues with data entry

Answer(s): A

---

8. What describes the use of UNION clause in a SQL statement?

A. Operates on queries and potentially decreases the number of rows

B. Operates on queries and potentially increases the number of rows

C. Operates on tables and potentially decreases the number of columns

D. Operates on both tables and queries and potentially increases both the number of rows and columns

Answer(s): B

---

9. What is the difference between the array and list data structures in R?

A. Arrays can contain different data types;Lists can contain only the same data type

B. Arrays are N-dimensional;Lists are only 2-dimensional

C. Arrays are only 2-dimensional;Lists are N-dimensional

D. Arrays contain only the same data type;Lists can contain different data types

Answer(s): D

---

10. You have completed your model and are handing it off to be deployed in production.

A. The production team are technical, and they need to understand how the processes that they support work, so give them the same presentation that you prepared for the analysts.

B. The production team supports the processes that run the organization, and they need context to understand how your model interacts with the processes they already support. Give them the executive summary.

C. The production team supports the processes that run the organization, and they need context to understand how your model interacts with the processes they already support. Give them the same presentation that you prepared for the project sponsor.

D. The production team needs to understand how your model will interact with the processes they already support. Give them documentation on expected model inputs and outputs, and guidance on error-handling.

Answer(s): D

---

11. Based on the exhibit, the table shows the values for the input Boolean attributes A, B, and C.

A. Tree B

B. Tree A

C. Tree D

D. Tree C

Answer(s): A

---

12. In which lifecycle stage are test and training data sets created?

A. Discovery

B. Model building

C. Data preparation

D. Model planning

Answer(s): B

---

13. You have run the association rules algorithm on your data set, and the two rules {banana, apple} =>

A. {grape, apple, orange} must be a frequent itemset.

B. {banana, apple} => {orange} must be a relevant rule.

C. {grape} => {banana, apple} must be a relevant rule.

D. {banana, apple, grape, orange} must be a frequent itemset.

Answer(s): A

---

14. Which word or phrase completes the statement? Data-ink ratio is to data visualization as _____.

A. Seasonality is to ARIMA

B. Data scientist is to big data

C. K-means is to Naive Bayes

D. Confusion matrix is to classifier

Answer(s): D

---

15. You fit a Logistic Regression model to your training data and notice that the variable X has an infinite magnitude coefficient.

A. X is strongly correlated with the outcome for a subset of the data

B. X takes on only one value for most of the training data

C. X is completely uncorrelated with the outcome

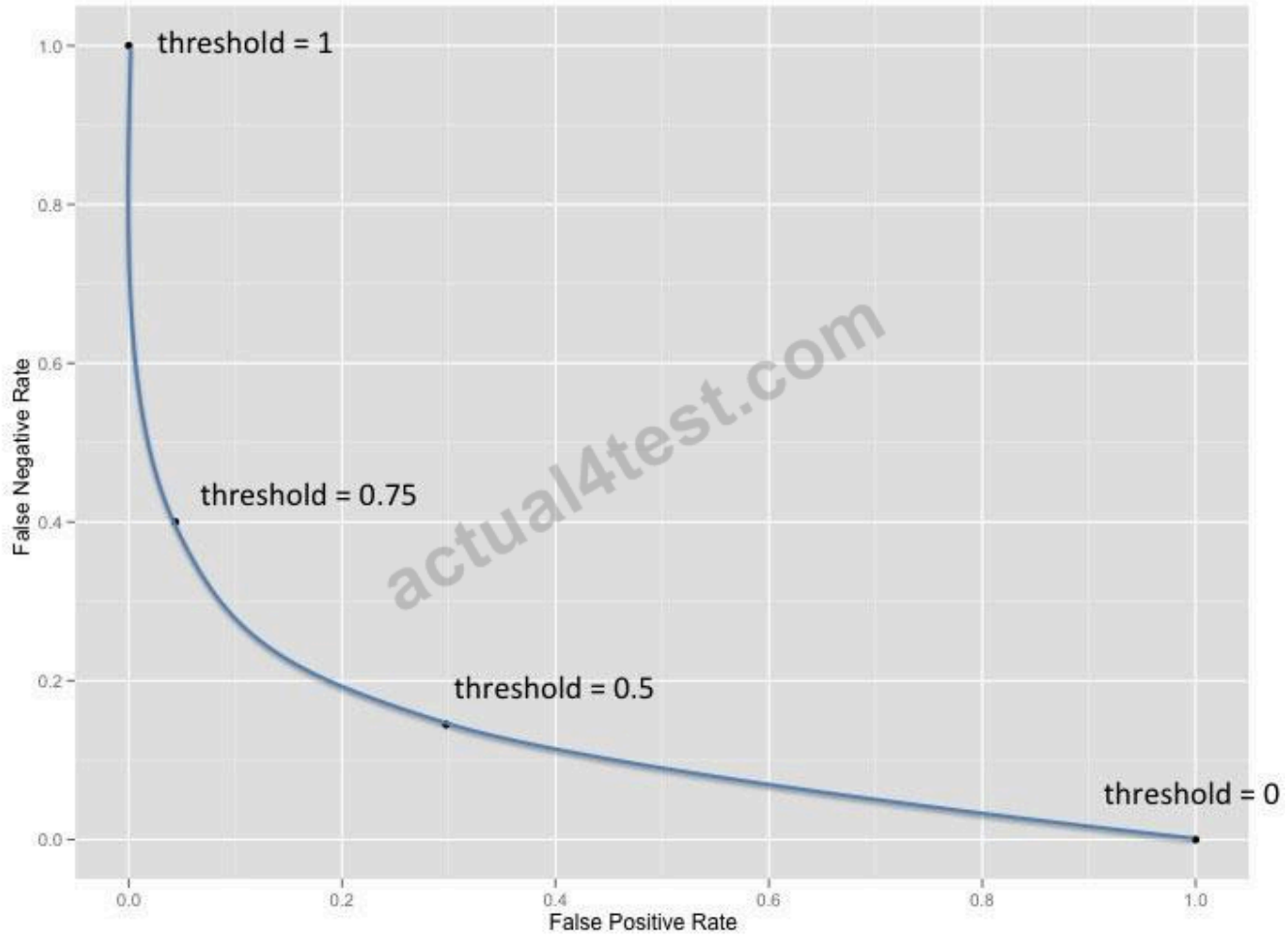
D. X is more correlated with the outcome than any of the other variables

Answer(s): A

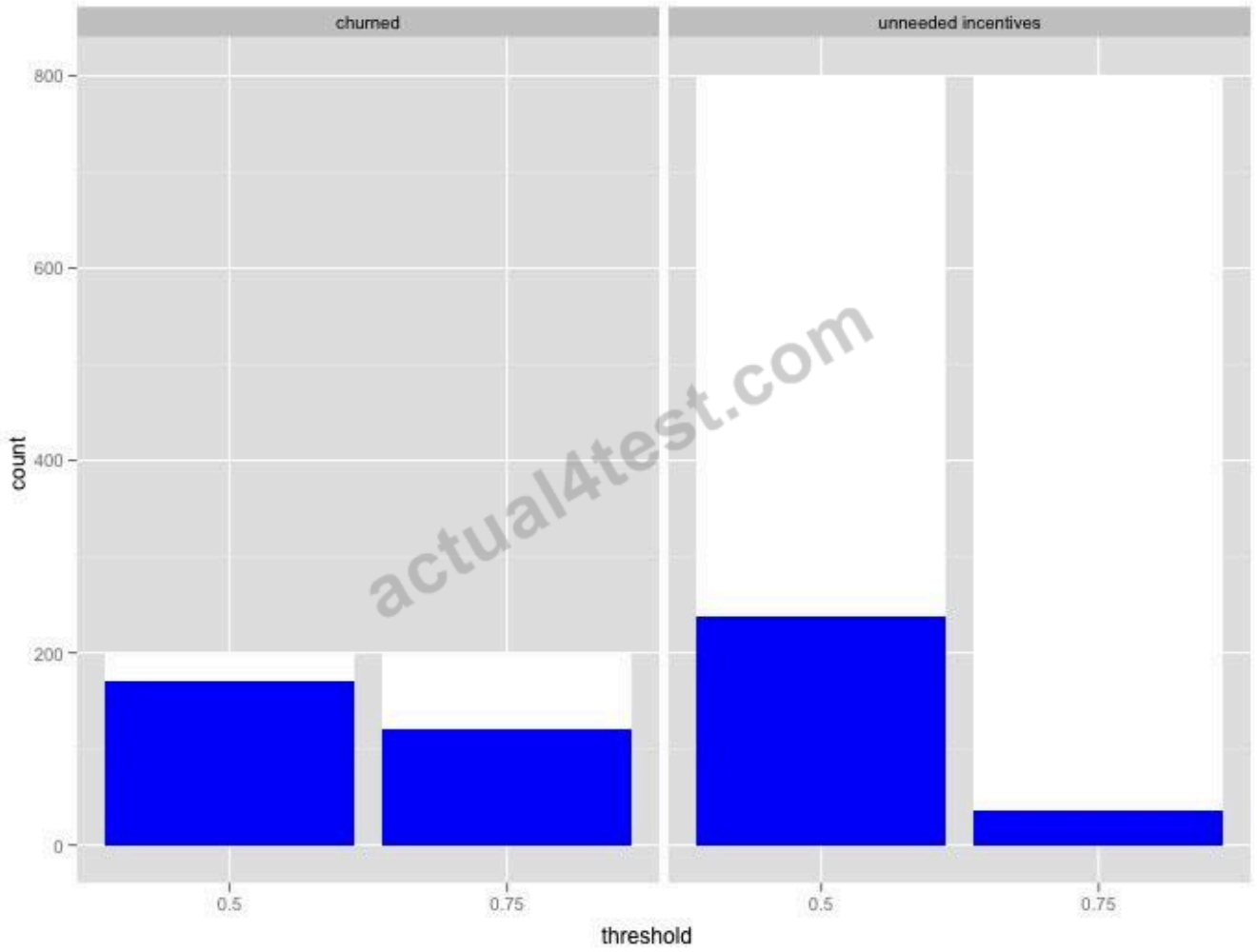
---

16. You have created a Logistic Regression model to predict customer churn for your company. The company's Marketing department wants to use your model to identify at-risk customers and offer incentives to keep them from leaving.

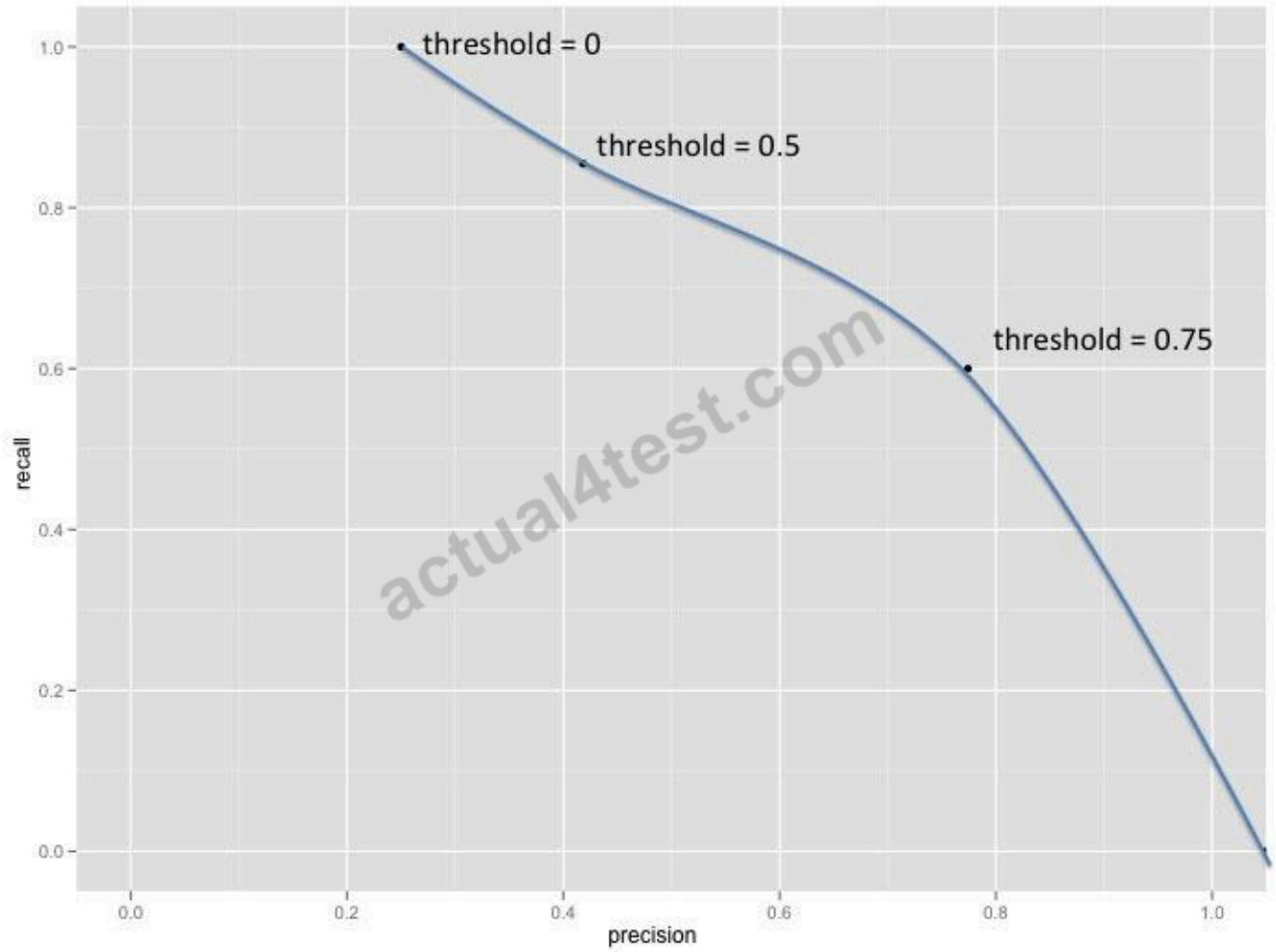
A. ○ **A.**



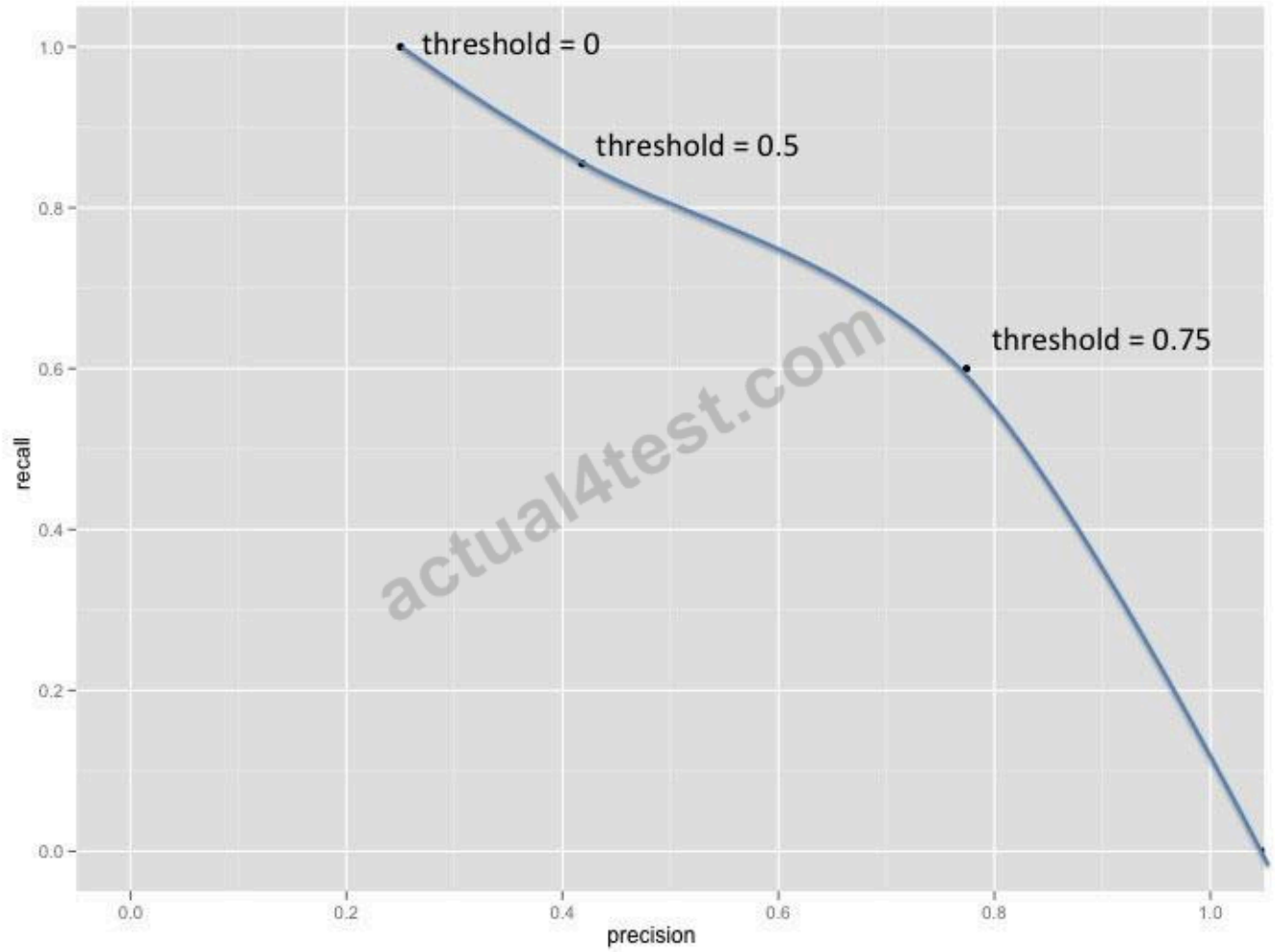
B. ○ **B.**



C. ○ **C.**



D. ○ **D.**



Answer(s): B

---

17. Which assumption makes the Naïve Bayesian classifier different from the general Bayesian model?

A. Features of a class are conditionally independent of one another

B. Number of features cannot be greater than the number of records

C. All variables need to be numeric

D. Fewer features can be used with the Naïve Bayes classifier

Answer(s): A

---

18. You have been assigned to perform a study of the daily revenue effect of a pricing model of online transactions. All data currently available to you has been loaded into your analytics database. This includes revenue data, pricing data, and online transaction data.

A. Disregard revenue as the key reason in the pricing model and create a daily model based on pricing and transactions only.

B. Aggregate all data to the monthly level in order to create a monthly revenue model.

C. Report back to the business owner that the current data model does not support the business question.

D. Interpolate a daily model for revenue from the monthly revenue data.

Answer(s): C

---

19. The web analytics team uses Hadoop to process access logs. They now want to correlate this data with structured user data residing in a production single-instance JDBC database. They collaborate with the production team to import the data into Hadoop.

A. Chukwa

B. Sqoop

C. Pig

D. Scribe

Answer(s): B

---

20. You have created a scatter plot using R from household income and education data as shown in the graphic.

A. Recreate the plot with a hexbin plot

B. Recreate the plot with a Bar plot

C. Add a rug to the plot

D. Add a Box and Whisker overlay to the plot

Answer(s): A
