Vendor:EMC

Exam Code:E20-007

Exam Name: Data Science and Big Data Analytics

Version: Demo

QUESTION 1

When is a Na?e Bayesian Classifier model for classification preferred versus a Logistic Regression model?

A. When using several categorical input variables with over 1000 possible values each

B. When an estimate of the probability of an outcome is needed, not just which class it is in

C. When all input variables are numerical

D. When some of the input variables might be correlated

Correct Answer: A

QUESTION 2

Consider these itemsets:

(hat, scarf, coat)

(hat, scarf, coat, gloves)

(hat, scarf, gloves)

(hat, gloves)

(scarf, coat, gloves)

What is the confidence of the rule (gloves -> hat)?

A. 75%

B. 60%

C. 66%

D. 80%

Correct Answer: A

QUESTION 3

Data visualization is used in the final presentation of an analytics project. For what else is this technique commonly used?

A. Data exploration

B. Descriptive statistics

C. ETLT

D. Model selection

Correct Answer: A

QUESTION 4

When is the GROUP BY ROLLUP clause used in an OLAP query?

A. All subtotals and grand totals are to be included in the output

B. Subtotals are only to be included in the output

C. Grand totals are only to be included in the output

D. Specific subtotals and grand totals for a combination of variables are only to be included in the output

Correct Answer: A

QUESTION 5

Refer to the exhibit.

Attribute	Info-Gain
Age	0.0310
Income	0.0100
Gender	0.0034
Credit Score	0.0456

You are building a decision tree. In this exhibit, four variables are listed with their respective values of info-gain.

Based on this information, on which attribute would you expect the next split to be in the decision tree?

A. Credit Score

B. Age

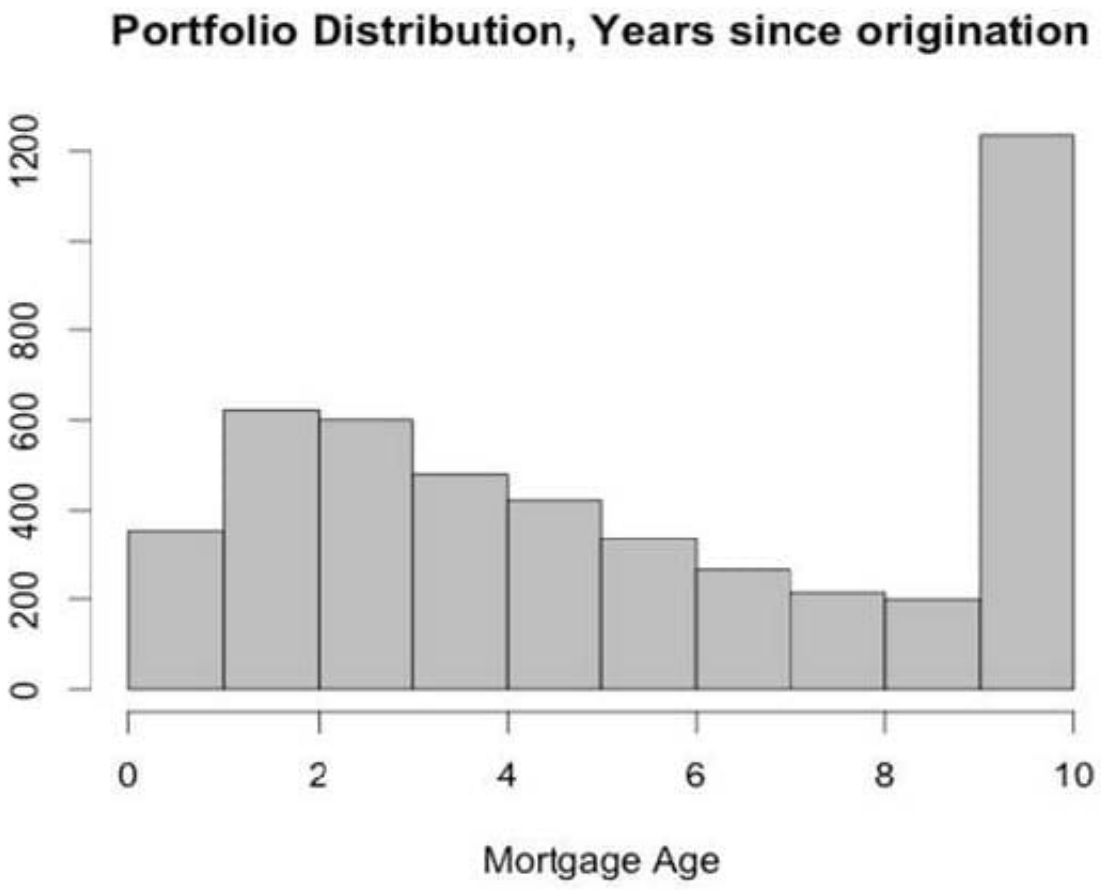
C. Income

D. Gender

Correct Answer: A

QUESTION 6

Refer to the exhibit.



Which type of data issue would you suspect based on the exhibit?

A. "Saturated" data, indicating potential issues with data definitions

B. Incomplete data, indicating potential issues with data transmission

C. Mis-scaled data, indicating potential issues with data entry

D. The exhibit does not raise any obvious concerns with the data.

Correct Answer: A

QUESTION 7

What describes a true property of a Logistic Regression method?

A. Robust with redundant variables and correlated variables

B. Handles missing values well

C. Works well with discrete variables that have many distinct values

D. Works well with variables that affect the outcome in a discontinuous way

Correct Answer: A

QUESTION 8

Refer to the exhibit.

Independent Variable	Coefficient	P-Value
A	0.45	0.0000
B	3.67	0.0000
C	1.23	0.0000

$R^2 = 0.10$

You are asked to write a report on how specific variables impact your client\\'s sales using a data set provided to you by the client. The data includes 15 variables that the client views as directly related to sales, and you are restricted to these variables only.

After a preliminary analysis of the data, the following findings were made:

1.

Multicollinearity is not an issue among the variables

2.

Only three variables--A, B, and C--have significant correlation with sales

You build a linear regression model on the dependent variable of sales with the independent variables of

A, B, and C. The results of the regression are seen in the exhibit.

Which interpretation is supported by the analysis?

A. Variables A, B, and C are significantly impacting sales, but are not effectively estimating sales

B. Variables A, B, and C are significantly impacting sales and are effectively estimating sales

C. Due to the R2 of 0.10, the model is not valid ?the linear regression should be re-run with all 15 variables forced into the model to increase the R2

D. Due to the R2 of 0.10, the model is not valid ?a different analytical model should be attempted

Correct Answer: A

QUESTION 9

While having a discussion with your colleague, this person mentions that they want to perform K-means clustering on

text file data stored in HDFS.

Which tool would you recommend to this colleague?

A. Mahout

B. HBase

C. Scribe

D. Sqoop

Correct Answer: A

QUESTION 10

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

QUESTION 11

Refer to the exhibit.

	FREE HOUSING	HOME OWNER	RENTER	TOTAL
BAD CREDIT	54	476	270	800
GOOD CREDIT	75	1245	460	1780
TOTAL	129	1721	730	2580

Click on the calculator icon in the upper left corner. You are given a list of pre-defined association rules:

A. RENTER => BAD CREDIT

B. RENTER => GOOD CREDIT

C. HOME OWNER => BAD CREDIT

D. HOME OWNER => GOOD CREDIT

E. FREE HOUSING => BAD CREDIT

F. FREE HOUSING => GOOD CREDIT For your next analysis, you must limit your dataset based on rules with confidence greater than 60%. Which of the rules will be kept in the analysis?

A. Rules B and D

B. Rules A and F

C. Rules C and E

D. Rules D and E

Correct Answer: A

QUESTION 12

You are testing two new weight-gain formulas for puppies. The test gives the results: Control group: 1% weight gain Formula A. 3% weight gain

Formula B. 4% weight gain A one-way ANOVA returns a p-value = 0.027 What can you conclude?

A. Either Formula A or Formula B is effective at promoting weight gain.

B. Formula B is more effective at promoting weight gain than Formula A.

C. Formula A and Formula B are both effective at promoting weight gain.

D. Formula A and Formula B are about equally effective at promoting weight gain.

Correct Answer: A