Free Questions for Databricks-Certified-Professional-Data-Scientist by certsinside

Shared by Jensen on 24-05-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

What are the key outcomes of the successful analytical projects?

Options:

A- Code of the model

B- Technical specifications

C- Presentations for the Analysts

D- Presentation for Project Sponsors

When your analytical project successfully completed they come up with the following at the end of the projects. Presentations- You will be having presentations like for the all the stakeholders, generally these presentation will help seniors executives to make better decisions. Similarly you would be creating presentations for the other teams like analysts various visuals you would be creating like ROC Curves, Heat Maps, and Bar Charts etc.

Whatever tools you have used like SAS, R, or Python then accordingly code was developed and you will get that code as one of the outcome. Also you would have created a technical specifications for implementing the codes.

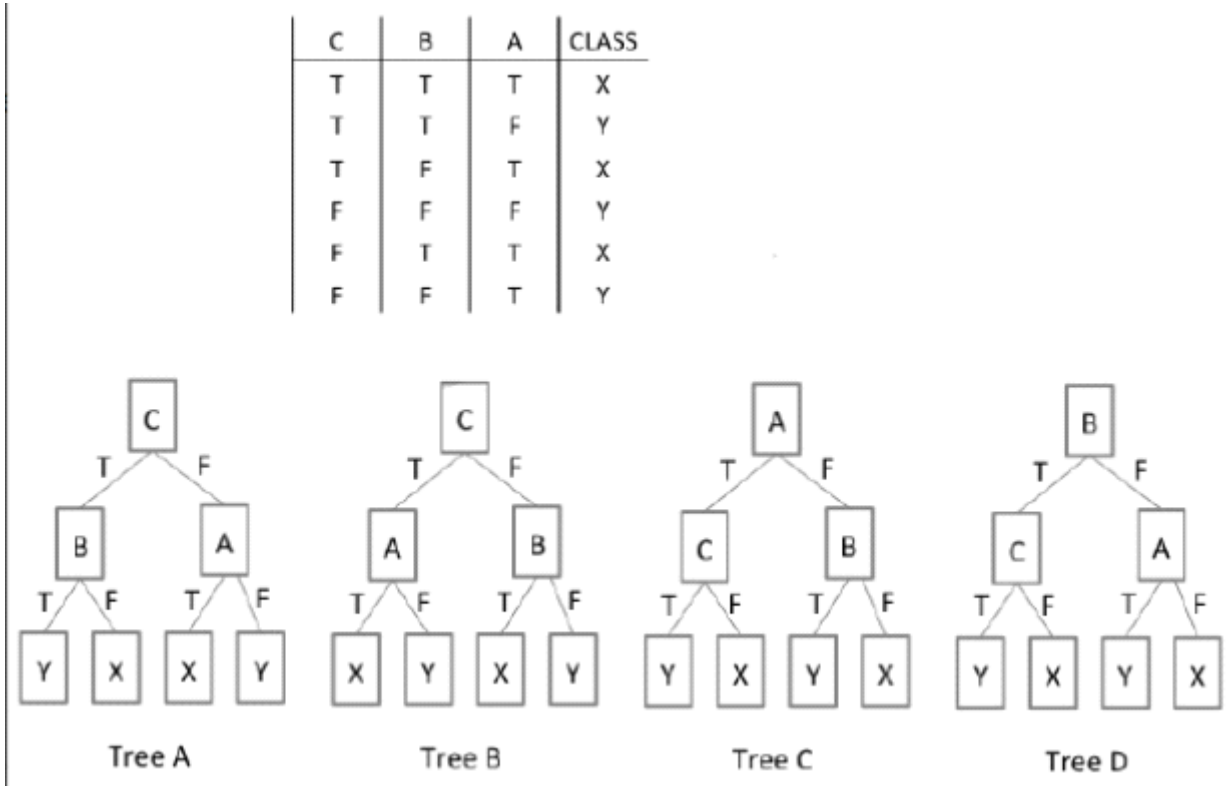
Answer:

A, B, C, D

Question 2

Question Type: MultipleChoice

Refer to the Exhibit.



C	B	A	CLASS
T	T	T	X
T	T	F	Y
T	F	T	X
F	F	F	Y
F	T	T	X
F	F	T	Y

In the Exhibit, the table shows the values for the input Boolean attributes "A", "B", and "C". It also shows the values for the output attribute "class". Which decision tree is valid for the data?

Options:

A- Tree A

B- Tree B

C- Tree C

D- Tree D

Answer:

B

Question 3

Question Type: **MultipleChoice**

Which of the following steps you will be using in the discovery phase?

Options:

A- What all are the data sources for the project?

B- Analyze the Raw data and its format and structure.

C- What all tools are required, in the project?

D- What is the network capacity required

E- What Unix server capacity required?

During the discovery phase you need to find how much resources are required as early as possible and for that even you can involve various stakeholders like Software engineering team, DBAs,
Network engineers, System administrators etc. for your requirement and these resources are already available or you need to procure them. Also, what would be source of the data? What all tools and software's are required to execute the same?

Answer:

A, B, C, D, E

Question 4

Question Type: **MultipleChoice**

Which of the following skills a data scientists required?

Options:

A- Web designing to represent best visuals of its results from algorithm.

B- He should be creative

C- Should possess good programming skills

D- Should be very good at mathematics and statistic

E- He should possess database administrative skills.

Yes a data scientists should have combination of skills like to solve the complex problem he should be creative as well as able to find new solutions and use of existing data. And solve the problem skills required are programming as currently we see SAS, R: Python, Spark, Java and SPSS even day by day new technologies are coming.

To apply various existing and new algorithm using Machine Learning, or Al it require good mathematics and statistics skills (Where the programmer feels, weaknesses). Another skill required is using visualization techniques like Qlik, Tableau etc

Answer:

B, C, D

Question 5

Question Type: **MultipleChoice**

Which of the following question statement falls under data science category?

Options:

A- What happened in last six months?

B- How many products have been sold in a last month?

C- Where is a problem for sales?

D- Which is the optimal scenario for selling this product?

E- What happens, if these scenario continues?

This question wants to check your understanding about Bl and Data Science. Bl was already existing and analytics team already using it. They need to improve and learn data science technique to solve some problems. If you check the option given in the question, it will confuse you. But if you have worked in Bl or as a Data Scientist then it is easy to answer. First 3 option can be easily answered using reporting solution, what sales happened in last six month, what was the problem etc.

But for the last two option you need to apply data science techniques like which all scenarios are optimal for product sales, you need to collect the data and applying various techniques for that. Hence, last two option can only be answered using Data Science technique

And for this you need to apply techniques like Optimization, predictive modeling, statistical analysis on structured and un-structured data.

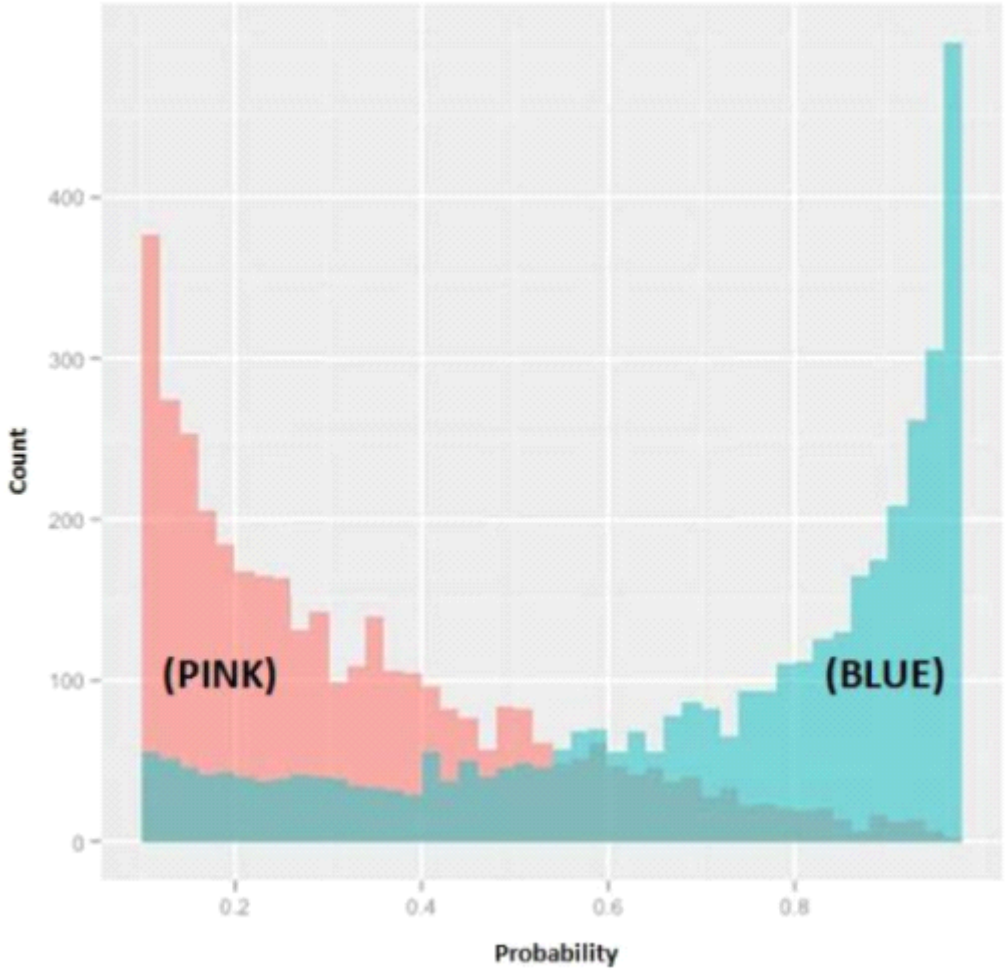
Answer:

D, E

Question 6

Question Type: MultipleChoice

Refer to Exhibit



In the exhibit, the x-axis represents the derived probability of a borrower defaulting on a loan. Also in the exhibit, the pink represents borrowers that are known to have not defaulted on their loan, and the blue represents borrowers that are known to have defaulted on

their loan. Which analytical method could produce the probabilities needed to build this exhibit?

Options:

A- Linear Regression

B- Logistic Regression

C- Discriminant Analysis

D- Association Rules

Answer:

B

Question 7

Question Type: **MultipleChoice**

Assume some output variable "y" is a linear combination of some independent input variables "A" plus some independent noise "e". The way the independent variables are combined is defined by a parameter vector B y=AB+e where X is an m x n matrix. B is a vector of n unknowns, and b is a vector of m values. Assuming that m is not equal to n and the columns of X are linearly independent, which expression correctly solves for B?

A. $b * (A^T * A)^{-1} * A^T$

B. $A^{-1} * b$

C. $(A^T * A)^{-1} * b$

D. $(A^T * A)^{-1} * A^T * b$

Options:

A- Option A

B- Option B

C- Option C

D- Option D

This is the standard solution of the normal equations for linear regression. Because A is not square, you cannot simply take its inverse.

Answer:

D

Question 8

Question Type: MultipleChoice

Clustering is a type of unsupervised learning with the following goals

Options:

A- Maximize a utility function

B- Find similarities in the training data

C- Not to maximize a utility function

D- 1 and 2

E- 2 and 3

type of unsupervised learning is called clustering. In this type of learning, The goal is not to maximize a utility function, but simply to find similarities in the training data.

The assumption is often that the clusters discovered will match reasonably well with an intuitive classification. For instance, clustering individuals based on demographics might result in a clustering of the wealthy in one group and the poor in another. Clustering can be useful when there is enough data to form clusters (though this turns out to be difficult at times) and especially when additional data about members of a cluster can be used to produce further results due to dependencies in the data.

Answer:

E

Question 9

Question Type: MultipleChoice

You are creating a model for the recommending the book at Amazon.com, so which of the following recommender system you will use you don't have cold start problem?

Options:

A- Naive Bayes classifier

B- Item-based collaborative filtering

C- User-based collaborative filtering

D- Content-based filtering

The cold start problem is most prevalent in recommender systems. Recommender systems form a specific type of information filtering (IF) technique that attempts to present information items (movies, music, books, news, images, web pages) that are likely of interest to the user. Typically, a recommender system compares the user's profile to some reference characteristics. These characteristics may be from the information item (the content-based approach) or the user's social environment (the collaborative filtering approach). In the content-based approach, the system must be capable of matching the characteristics of an item against relevant features in the user's profile. In order to do this, it must first construct a sufficiently-detailed model of the user's tastes and preferences through preference elicitation. This may be done either explicitly (by querying the user) or implicitly (by observing the user's behaviour). In both cases, the cold start problem would imply that the user has to dedicate an amount of effort using the system in its 'dumb' state - contributing to the construction of their user profile - before the system can start providing any intelligent recommendations.

Content-based filtering recommender systems use information about items or users to make recommendations, rather than user

preferences, so it will perform well with little user preference data. Item-based and user-based collaborative filtering makes predictions based on users' preferences for items, os they will typically perform poorly with little user preference data. Logistic regression is not recommender system technique.

Answer:

D

Question 10

Question Type: MultipleChoice

What is one modeling or descriptive statistical function in MADlib that is typically not provided in a standard relational database?

Options:

A- Expected value

B- Variance

C- Linear regression

D- Quantiles

Linear regression models a linear relationship of a scalar dependent variable y to one or more explanatory independent variables x to

build a model of coefficients.

Answer:

C

Question 11

Question Type: MultipleChoice

Which technique you would be using to solve the below problem statement? "What is the probability that individual customer will not repay the loan amount?"

Options:

A- Classification

B- Clustering

C- Linear Regression

D- Logistic Regression

E- Hypothesis testing

Answer:

D

Question 12

Question Type: MultipleChoice

Which of the following statement is true for the R square value in the regression model?

Options:

A- When R square =1 , all the residuals are equal to 0

B- When R square =0, all the residual are equal to 1

C- R square can be increased by adding more variables to the model.

D- R-squared never decreases upon adding more independent variables.

R square can be made high, it means when we add more variables R-square will increase. And R-square will never decreases if you add more independent variables. Higher R square value can have lower the residuals.

Answer:

A, C, D