CHAPTER 4

BIG DATA PROJECTS

1. (A) feature design.

Explanation

Data exploration encompasses exploratory data analysis, feature selection, and feature engineering.

(Module 4.2, LOS 4.d)

Related Material

SchweserNotes - Book 1

2. (B) recall is the ratio of correctly predicted positive classes to all actual positive classes.

Explanation

Recall (also called sensitivity) is the ratio of correctly predicted positive classes to all actual positive classes. Precision is the ratio of correctly predicted positive classes to all predicted positive classes. Accuracy is the percentage of correctly predicted classes out of total predictions.

(Module 4.3, LOS 4.c)

Related Material

SchweserNotes - Book 1

3. (A) tokenization.

Explanation

Text is considered to be a collection of tokens, where a token is equivalent to a word. Tokenization is the process of splitting a given text into separate tokens. Bag-of-words (BOW) is a collection of a distinct set of tokens from all the texts in a sample dataset. Stemming is the process of converting inflected word forms into a base word.

(Module 4.1, LOS 4.g)

Related Material

SchweserNotes - Book 1

4. (C) An analyst adjusts daily stock index data from two countries for their different market holidays.

Explanation

Curation is ensuring the quality of data, for example by adjusting for bad or missing data. Word clouds are a visualization technique. Moving data from a storage medium to where they are needed is referred to as transfer.

(Module 4.1, LOS 4.a)

Related Material

SchweserNotes - Book 1

Freja Karlsson is a bond analyst with Storbank AB. Over the past several months, Karlsson has been working to develop her own machine learning (ML) model that she plans to use to predict default of the various bonds that she covers. The inputs to the model are various pieces of financial data that Karlsson has compiled from multiple sources.

After Karlsson has constructed the model using her knowledge of appropriate variables, Karlsson runs the model on the training set. Each firm's bonds are classified as predicted-to-default or predicted-not-to-default. When Karlsson's model predicts that a bond will default and the bond actually defaults, Karlsson considers this to be a true positive. Karlsson then evaluates the performance of her model using error analysis. The confusion matrix that results is shown in Exhibit 1.

N = 474		Actual bond Status	
		Bond Default	No Default
Model Prediction	Bond Default	307	31
	No Default	23	113

5. (C) 91%.

Explanation

Precision, the ratio of correctly predicted positive classes (true positives) to all predicted positive classes, is calculated as:

Precision (P) = TP /(TP + FP) = 307 / (307 + 31) = 0.9083 (91%)

In the context of this default classification, high precision would help us avoid the situation where a bond is incorrectly predicted to default when it actually is not going to default.

(Module 4.3, LOS 4.c)

Related Material

SchweserNotes - Book 1

6. (C) 93%.

Explanation

Recall that = TP / (TP + FN)

= 307 / (307 + 23)

= 0.9303

= 93%.

Recall is useful when the cost of a false negative is high, such as when we predict that a bond will not default but it actually will. In cases like this, high recall indicates that false negatives will be minimized.

(Module 4.3, LOS 4.c)

Related Material

SchweserNotes - Book 1

7. (C) 92%.

Explanation

The model's F1 score, which is the harmonic mean of precision and recall, is calculated as:

F1 score = (2 × P × R) / (P + R)

= (2 × 0.9083 × 0.9303) / (0.9083 + 0.9303)

= 0.9192(92%)

Like accuracy, F1 is a measure of overall performance measures that gives equal weight to FP and FN.

(Module 4.3, LOS 4.c)

Related Material

SchweserNotes - Book 1

8. (B) 89%.

Explanation

The model's accuracy is the percentage of correctly predicted classes out of total predictions.

Model accuracy is calculated as:

Accuracy = (TP + TN) / (TP + FP + TN + FN) = (TP + TN) / N

= (307 + 113) / (307 + 31 + 113 + 23) = (307 + 113) / (474).

= 0.8861 = 89%

(Module 4.3, LOS 4.c)

Related Material

SchweserNotes - Book 1

9. (C) Volume and velocity.

Explanation

Big Data may be characterized by its volume (the amount of data available), velocity (the speed at which data are communicated), and variety (degrees of structure in which data exist). "Terabyte" is a measure of volume. "Latency" refers to velocity.

(Module 4.1, LOS 4.a)

Related Material

SchweserNotes - Book 1

10. (C) veracity.

Explanation

Big data is defined as data with high volume, velocity, and variety. Big data often suffers from low veracity, because it can contain a high percentage of meaningless data.

(Module 4.1, LOS 4.a)

Related Material

SchweserNotes - Book 1

11. (C) The model treats true parameters as noise.

Explanation

Underfitting describes a machine learning model that is not complex enough to describe the data it is meant to analyze. An underfit model treats true parameters as noise and fails to identify the actual patterns and relationships. A model that is overfit (too complex) will tend to identify spurious relationships in the data. Labelling of input data is related to the use of supervised or unsupervised machine learning techniques.

(Module 4.3, LOS 4.f)

Related Material

SchweserNotes - Book 1