Feature Engineering quiz

8 out of 8 correct

1. What is a common technique used for handling missing data in a dataset?

- ○ Dropping the rows with missing data
- ● Imputing the missing values with the mean value
- ○ Ignoring the missing data and proceeding with analysis
- ○ None of the above

Explanation: Mean imputation is a common technique used for handling missing data in a dataset.

2. Why is it important to handle missing data before performing any analysis or modeling?

- ○ It can cause bias in the results
- ○ It can lead to inaccurate predictions
- ○ It can affect the generalizability of the model
- ● All of the above

Explanation: Handling missing data is important because it can lead to bias in the results, inaccurate predictions, and affect the generalizability of the model.

3. What is imbalanced data?

- ○ A dataset with missing values
- ◉ **A dataset with unequal distribution of classes**
- ○ A dataset with outliers
- ○ A dataset with noisy data

4. What is a common technique used for handling imbalanced data in a dataset?

- ○ Up-sampling the minority class
- ○ Down-sampling the majority class
- ○ Using cost-sensitive learning algorithms
- ◉ **All of the above**

Explanation: Up-sampling, down-sampling, and using cost-sensitive learning algorithms are all common techniques used for handling imbalanced data in a dataset.

5. What is a common technique used for down-sampling a dataset?

- ◉ **Randomly selecting a subset of the majority class**

- ○ Creating a new class in the dataset
- ○ Removing data from the minority class
- ○ None of the above

Explanation: Down-sampling involves randomly selecting a subset of the majority class to balance the dataset.

6. What metric is commonly used to evaluate the performance of a machine learning model on an imbalanced dataset?

- ○ Accuracy
- ○ Precision
- ○ Recall
- ● All of the above

Explanation: Accuracy, precision, and recall are all commonly used metrics to evaluate the performance of a machine learning model on an imbalanced dataset.

7. You are working with a dataset that has missing values, and you decide to use mean imputation to fill in the missing values. What is a potential drawback to using mean imputation?

- ● It can result in inaccurate imputed values
- ○ It can be computationally expensive
- ○ It can introduce bias into the analysis
- ○ It can only be used for numerical data

Explanation: Mean imputation assumes that the missing values are

missing completely at random and that the values can be accurately estimated based on the mean of the observed values. However, this may not always be the case and can result in imputed values that are not accurate.

8. You are working on a project that requires predicting customer satisfaction. However, the dataset is imbalanced with a majority of customers being satisfied. What is a common technique you can use to down-sample the majority class and balance the dataset?

- (●) Random undersampling
- ( ) Random oversampling
- ( ) SMOTE
- ( ) ADASYN

Explanation: This technique involves randomly selecting instances from the majority class to match the number of instances in the minority class.

Submit