AI Bootcamp

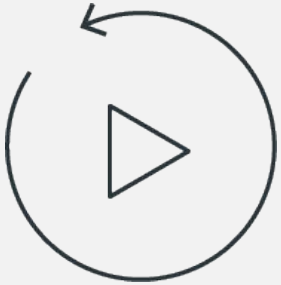
Advanced Preprocessing Techniques

Module 14 Day 2

Class Objectives

By the end of class, you will be able to:

1. Recognize and address data leakage in datasets.
2. Apply innovative methods to handle missing values in data.
3. Evaluate and select appropriate encoding strategies for categorical data.
4. Utilize **OneHotEncoder** and **OrdinalEncoder** for data transformation.
5. Ensure prevention of data leakage during train–test data splits.
6. Construct preprocessing functions to streamline data preparation.
7. Design and incorporate new features to enhance machine learning model performance.



Let’s **recap**

Recap

01 Day 1

Focused on metrics and target selection

02 Day 2

Focuses on advanced preprocessing and refining data

Day 1: Model Validation and Imbalanced Data

Day 2: Focus on advanced preprocessing and refining data

Instructor **Demonstration**

Bank Model Review

Instructor **Demonstration**

Understanding Data Leakage

Data leakage occurs when a model is trained using information that won't be available when making predictions. The model will lack performance in production.

Examples of Data Leakage

Target column left in X data

Model trained on `Row_number` column

X value scaling

eBay issue

Activity:

Spotting Data Leakage

In this activity, you will engage with a dataset and identify potential data leakage.

Suggested Time:
10 Minutes

Time’s up!

Let’s review

Activity:

Missing Values

In this activity, you will analyse a dataset, identify missing values, perform imputations, and begin preprocessing data.

Suggested Time:
20 Minutes

Imputation involves the use of alternative values in place of missing data.

Imputation

Imputation is an art and science

Contextual data

Is imputing necessary?

Double-check your model

Activity:

Choosing Encodings

In this activity, you will learn about the use of encoding data using **OneHotEncoder** and **OrdinalEncoder** from sklearn.

Suggested Time:
20 Minutes

Encoding involves the transformation of categorical variables from strings into numbers.

Nominal variables don’t have an inherent order. They are simply categories that can be distinguished from each other.

Ordinal variables have an inherent order and can be ranked from highest to lowest or vice versa.

Instructor **Demonstration**

Feature Engineering

Feature engineering refers to the conversion of raw observations into features.

Feature Engineering

01

Simple operations can reveal patterns and detect outliers.

02

Increase model efficiency

03

Don’t create noise

Activity:

Third Model

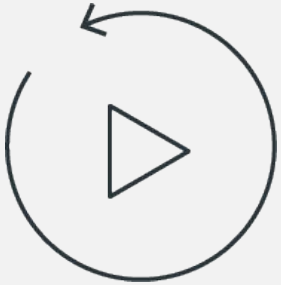
In this activity, you will practice your new skills on the Bank Marketing dataset.

Suggested Time:
40 Minutes

Time’s up!

Let’s review

Questions?



Let’s **recap**

Review the Class Objective

In this lesson, you learned how to:

1. Recognize and address data leakage in datasets.
2. Apply innovative methods to handle missing values in data.
3. Evaluate and select appropriate encoding strategies for categorical data.
4. Utilize **OneHotEncoder** and **OrdinalEncoder** for data transformation.
5. Ensure prevention of data leakage during train–test data splits.
6. Construct preprocessing functions to streamline data preparation.
7. Design and incorporate new features to enhance machine learning model performance.

Next

In the next lesson, you will delve deeper into hyperparameter tuning, explore resampling techniques, and work on building or improving a fourth model.

Questions?

The End