AI Bootcamp

Tuning Models and Sampling Data

Module 14 Day 3

Class Objectives

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

1. Perform hyperparameter tuning.
2. Use random and synthetic resampling to address class imbalance.
3. Apply new techniques to the Bank Marketing dataset.
4. Demonstrate the application of an existing model to new data.

Instructor **Demonstration**

Review Bank Marketing Model

Any questions relating to:

1. Defining target column
2. Selecting and implementing metrics
3. Filling missing values
3. Encoding categorical data

Activity:

Hyperparameter Tuning

In this activity, you learn about resampling, hyperparameter tuning or optimizing, using sklearn's **GridSearchCV** and **RandomizedSearchCV**, and how to differentiate between the two.

Suggested Time:
15 Minutes

Hyperparameter tuning is the process of adjusting model parameters to improve performance.

GridSearchCV is a speedy function to obtain optimal values for the hyperparameter of a model. It loops through predefined hyperparameters and fits the model on the training set.

RandomizedSearchCV: Instead of trying every combination of hyperparameter values like **GridSearchCV**, **RandomizedSearchCV** checks randomly selected values. This method uses less resources and time for similar results.

Instructor **Demonstration**

Resampling

Undersampling: Removing rows of the majority class from a dataset with the intention of balancing the number of instances of each class

Oversampling: Adding additional rows from the minority class to a dataset with the intention of balancing the number of instances of each class

Random resampling: Removing or adding *existing* rows of data from a dataset

Synthetic resampling: Using synthetic data to oversample or undersample

Resampling Methods

We'll learn the following techniques

1. Random Oversampling
2. Random Undersampling
3. Cluster Centroids
4. Synthetic Minority Oversampling Technique (SMOTE)
5. SMOTE and Edited Nearest Neighbors (SMOTEENN)

While these techniques have different effects on the data, they are all implemented with very similar code. **How can we choose which technique to use?**

Activity:

Bank Marketing Resampling

In this activity, you will use random sampling techniques to better predict customer responses to future telemarketing campaigns to increase their effectiveness.

Suggested Time:
15 Minutes

Time’s up!

Let’s review

Activity:

What Else?

In this activity, you will brainstorm potential improvements and techniques that have not been applied yet, grapple with the viability of achieving a perfect balanced accuracy score, identify internal and external factors and additional data that may improve model performance, and discuss the next steps for the Bank Marketing dataset.

Suggested Time:
20 Minutes

What methods have we learned that we have **NOT** yet applied to the data?

Given the dataset we have, if we had infinite time to work on our preprocessing and on our model, is a balanced accuracy of 100% achievable? **Why or why not?**

What balanced accuracy do you think **is** possible given unlimited time?

Is there any **additional data** the bank could collect that would help our model?

Activity:
Fourth Model

In this activity, you are encouraged to use any technique to be able to improve the model.

Suggested Time:
60 Minutes

Time’s up!

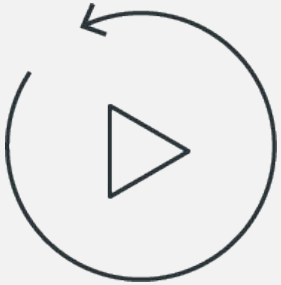
Let’s review

Activity:

New Data

In this activity, you will be given new data to apply your preprocessing steps to and make predictions on.

Suggested Time:
15 Minutes



Let’s **recap**

Review the Class Objective

In this lesson, you learned how to:

1. Perform hyperparameter tuning.
2. Use random and synthetic resampling to address class imbalance.
3. Apply new techniques to the Bank Marketing data.
4. Apply an existing model to new data.

Questions?

The End