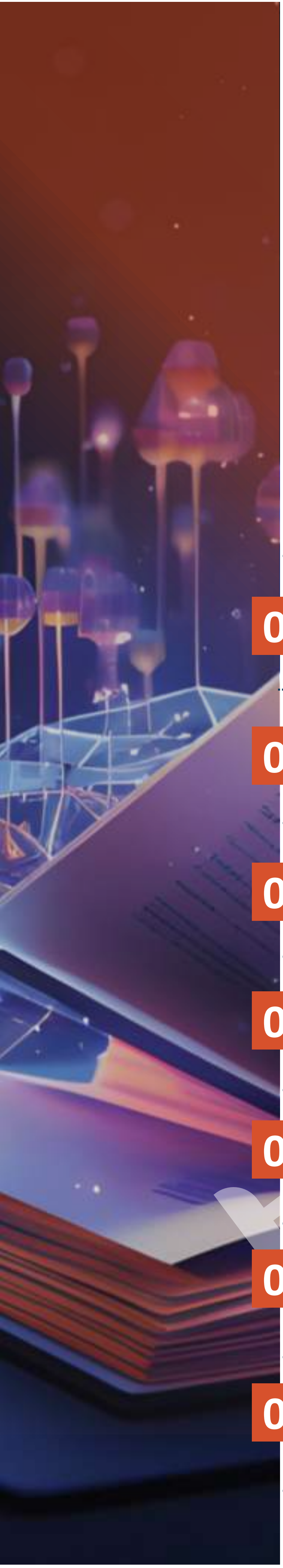
www.aileaders-project.eu

SCENARIO EXERCISE - The importance of data quality in AI-targeted marketing campaigns

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.



SCENARIO EXERCISE – The importance of data quality in AI-targeted marketing campaigns

01 Abstract

Type of OER

Scenario Exercise

Goal/Purpose

Raise awareness about the importance of data quality for implementing automated decision making algorithms (ADMS), particularly in the marketing sector.

Expected Learning Outcomes:

The student will be able to implement measures to address bias in customer behaviour predictions.

Keywords:

- Data Quality
- Biased Algorithms
- Unbalanced Data
- Marketing Campaigns
- ADMS

Suggested Methodological Approach:

Problem-Based Learning

02 Introduction

Context: predicting bank marketing success using machine learning

01 The dataset used in this project is the Bank Marketing Dataset from the UCI Machine Learning Repository. It contains details about clients contacted in a marketing campaign and whether they subscribed to a term deposit.

02 Dataset Description: Bank Marketing Dataset – UCI (bank.csv included in the OER materials).

03 Read carefully each variable and understand its meaning.

Key Problem Statement

01 How can we predict whether a client will subscribe to a term deposit based on their profile and past campaign interactions?

02 **Challenge**: The dataset is highly imbalanced, with far fewer clients subscribing to a deposit, leading to potential bias in the model.

03 Tools Presentation

Tools used in this scenario exercise

01 Python:

The most used programming language in Data Science

02 A python editor:

Google Colab or Jupyter Notebook

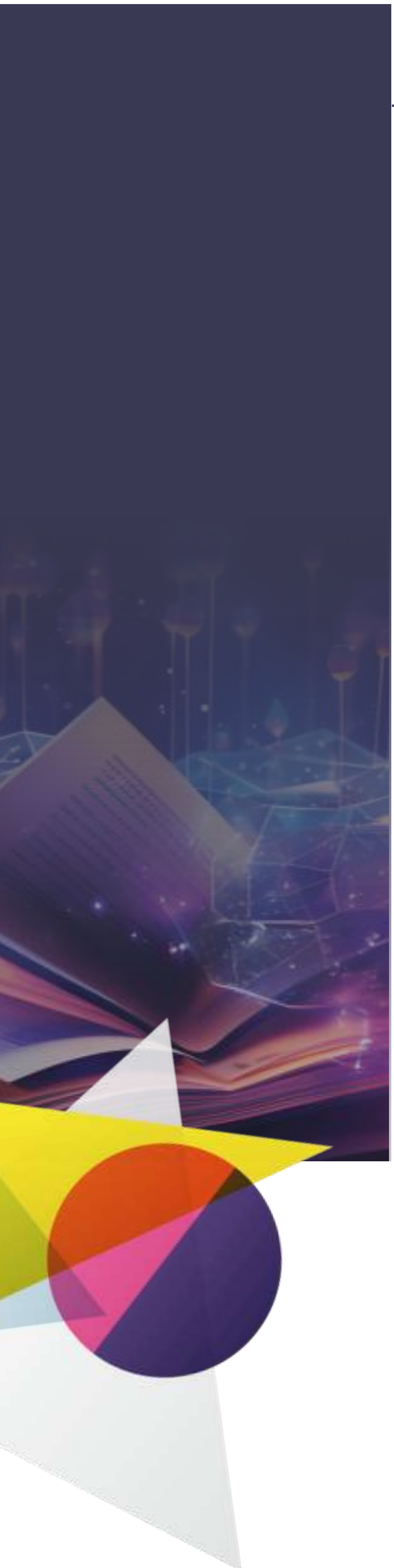
03 Libraries and Frameworks:

- Pandas: data manipulation and cleaning
- Scikit-learn: model training, evaluation, and data preprocessing
- XGBoost: optimized gradient boosting framework
- imbalanced-learn (SMOTE): to address class imbalance issues
- Matplotlib/Seaborn: visualization and data insights

04 Techniques Applied:

- Data Cleaning & Feature Engineering
- Handling Class Imbalance (SMOTE, Undersampling, Class Weights)
- Model Training (RandomForest, XGBoost)
- Performance Evaluation (Recall, Precision, and F1-score)

04 Hands-on Activities



01 Data Preprocessing

- Removed variables such as duration, campaign, and pdays that are only known after contacting clients to avoid data leakage.
- Encoded categorical variables and scaled numerical data for better model performance.

02 Handling Class Imbalance

Applied:

- Class Weights in RandomForest to penalize errors on the minority class.
- SMOTE (Synthetic Minority Over-sampling Technique) to artificially increase samples in the minority class.
- Undersampling to reduce the size of the majority class.

03 Model Training

Tested multiple models:

- RandomForest for baseline prediction.
- XGBoost for improved performance, optimized with Early Stopping and Feature Selection to reduce training time.

04 Evaluation

Evaluated results using

- Recall for minority class detection.
- F1-Score for balanced accuracy between precision and recall.

05 Conclusion

Key Insights

01 Bias in Data matters:

The imbalance in the dataset led initial models to ignore the minority class (clients subscribing to deposits).

02 Balancing Techniques are key:

Undersampling, Class Weights, and SMOTE improved recall significantly, albeit with trade-offs in overall accuracy.

03 XGBoost with Feature Selection:

By reducing the number of features and adding early stopping, XGBoost improved performance without compromising efficiency.

Key Lessons

01 Ethical AI Design:

Ethical AI design requires thoughtful dataset preparation, fair evaluation metrics, and awareness of potential bias in outcomes.

06 References

- Bank Marketing Dataset - UCI Machine Learning Repository: https://archive.ics.uci.edu/dataset/222/bank+marketing
- Scikit-learn Documentation: https://scikit-learn.org/
- XGBoost Documentation: https://xgboost.readthedocs.io/en/stable/
- Imbalanced-learn Documentation: https://imbalanced-learn.org/

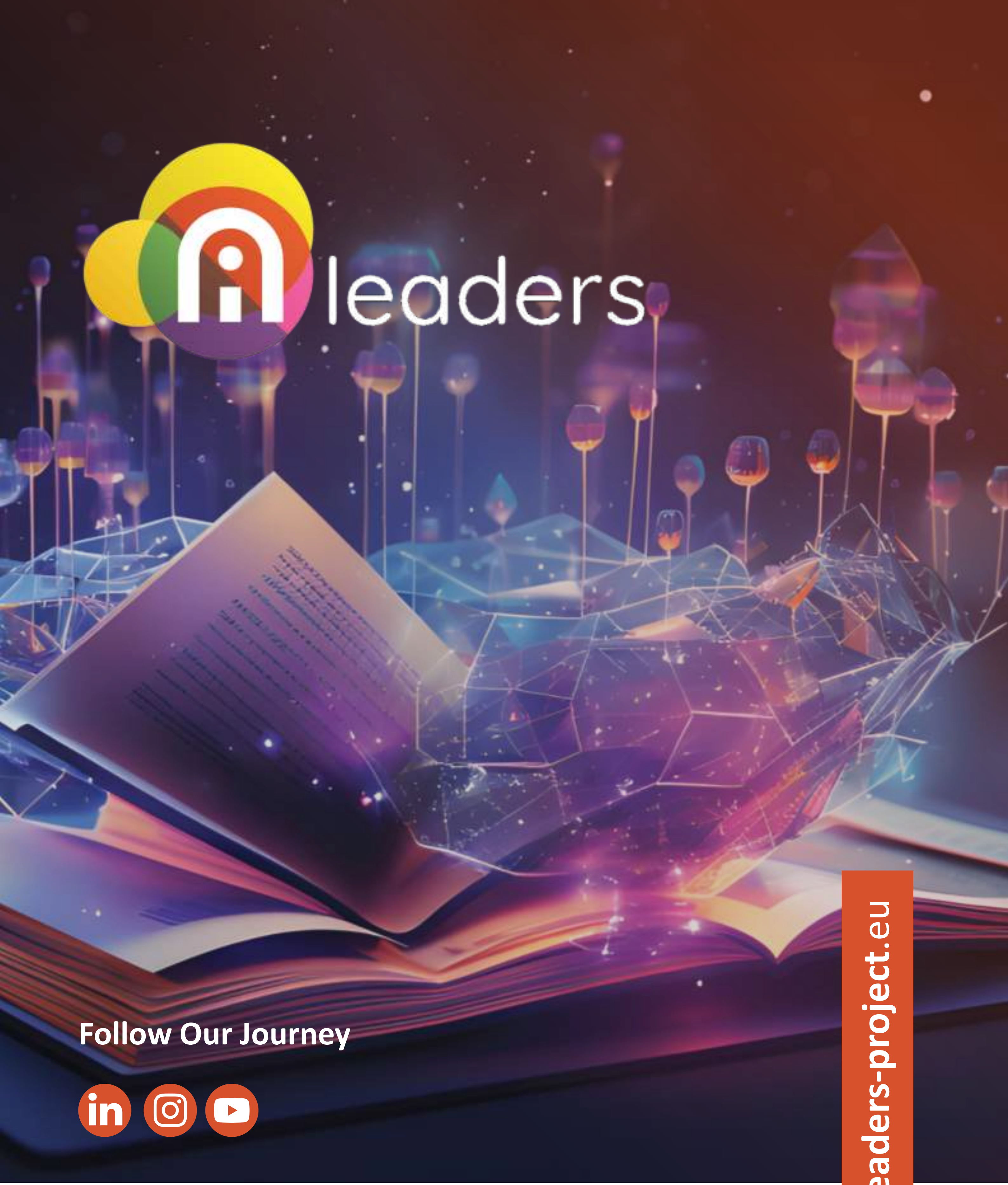
07 Complementary Material

Jupyter Notebook (IPYNB) File

A detailed Python Notebook with commented code that walks through each step of the process is included.

The notebook contains:

- Data cleaning and preprocessing steps.
- Feature engineering and variable selection logic.
- Implementation of different balancing techniques (SMOTE, Class Weights, and Undersampling).
- Model training with RandomForest and XGBoost.
- Evaluation metrics and insights from the results.



www.aileaders-project.eu

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.