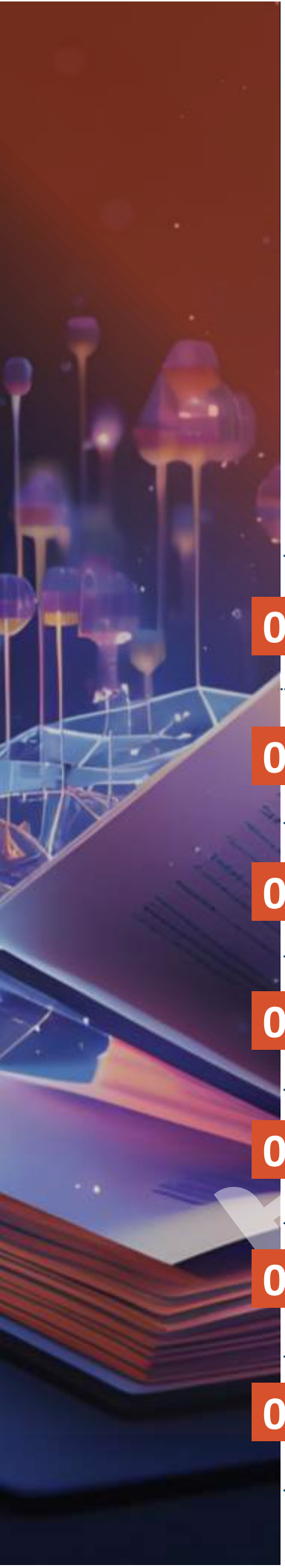
www.aileaders-project.eu

ĆWICZENIE SCENARIUSZOWE – Znaczenie jakości danych w kampaniach marketingowych opartych na sztucznej inteligencji

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.



ĆWICZENIE SCENARIUSZOWE – Znaczenie jakości danych w kampaniach marketingowych ukierunkowanych na sztuczną inteligencję kampaniach marketingowych

01 Streszczenie

Rodzaj OER

Ćwiczenie scenariuszowe

Cel/przeznaczenie

Podnoszenie świadomości na temat znaczenia jakości danych dla wdrażania algorytmów automatycznego podejmowania decyzji (ADMS), szczególnie w sektorze marketingowym.

Oczekiwane efekty kształcenia

Student będzie potrafił wdrożyć środki mające na celu wyeliminowanie błędów w prognozowaniu zachowań klientów.

Słowa kluczowe

- Jakość danych
- Algorytmy stronnicze
- Nierównowaga danych
- Kampanie marketingowe
- ADMS

Sugerowane Metodologiczne Podejście

Nauczanie oparte na problemach

02 Wprowadzenie

Kontekst: przewidywanie marketingu bankowego sukces przy użyciu uczenia maszynowego

01 Zbiór danych wykorzystany w tym projekcie to zbiór danych dotyczących marketingu bankowego z repozytorium UCI Machine Learning Repository. Zawiera on szczegółowe informacje o klientach, z którymi skontaktowano się w ramach kampanii marketingowej, oraz o tym, czy wykupili oni lokatę terminową.

02 Opis zbioru danych: Bank Marketing Dataset – UCI (plik bank.csv zawarty w materiałach OER).

03 Należy uważnie przeczytać każdą zmienną i zrozumieć jej znaczenie.

Kluczowe sformułowanie problemu

01 Jak możemy przewidzieć, czy klient wykupi lokatę terminową na podstawie jego profilu i dotychczasowych interakcji w ramach kampanii?

02 **Wyzwanie**: Zbiór danych jest bardzo niezrównoważony, znacznie mniej klientów wykupuje lokaty, co może prowadzić do potencjalnego błędu systematycznego w modelu.

03 Prezentacja narzędzi

Narzędzia wykorzystane w tym scenariuszu ćwiczeniowym

01 Python

Najczęściej używany język programowania w nauce o danych

02 Edytor języka Python

Google Colab lub Jupyter Notebook

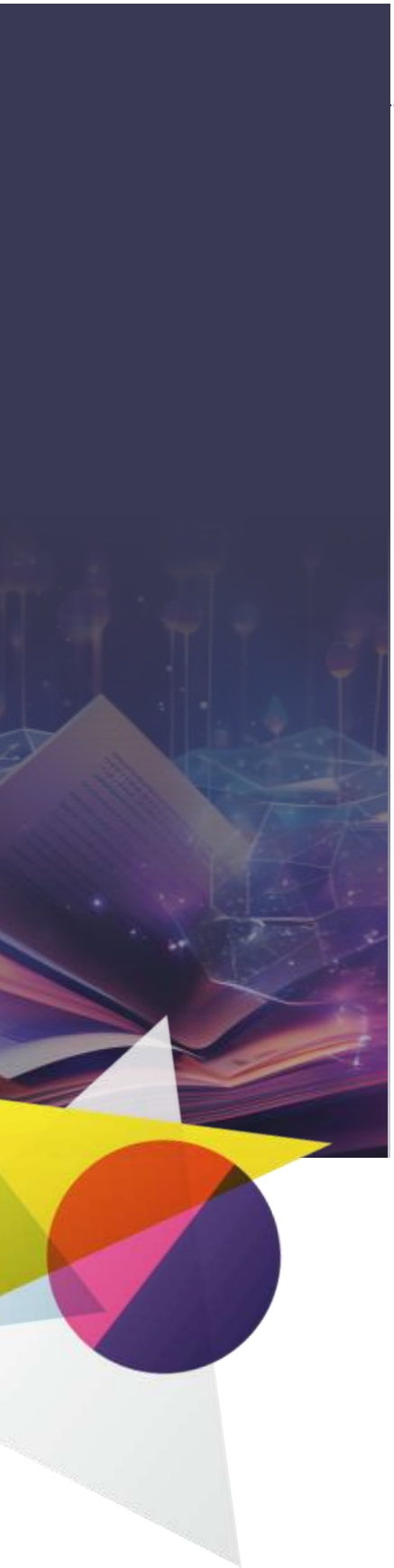
03 Biblioteki i frameworki

- Pandas: manipulacja danymi i czyszczenie danych
- Scikit-learn: szkolenie modeli, ocena i przetwarzanie wstępne danych
- XGBoost: zoptymalizowany framework do wzmacniania gradientowego
- imbalanced-learn (SMOTE): rozwiązanie problemów związanych z nierównowagą klas
- Matplotlib/Seaborn: wizualizacja i analiza danych

04 Zastosowane techniki

- Czyszczenie danych i inżynieria cech
- Postępowanie w przypadku nierównowagi klas (SMOTE, undersampling, wagi klas)
- Trening modelu (RandomForest, XGBoost)
- Ocena wydajności (przypomnienie, precyzja i wynik F1)

04 Ćwiczenia praktyczne



01 Wstępne przetwarzanie danych

- Usunięto zmienne, takie jak czas trwania, kampania i pdays, które są znane dopiero po skontaktowaniu się z klientami, aby uniknąć wycieku danych.
- Zakodowano zmienne kategoryczne i skalowano dane liczbowe w celu uzyskania lepszej wydajności modelu.

02 Postępowanie w przypadku nierównowagi klas

Zastosowano:

- Wagi klas w RandomForest, aby penalizować błędy w klasie mniejszościowej.
- SMOTE (Synthetic Minority Over-sampling Technique) w celu sztucznego zwiększenia liczby próbek w klasie mniejszościowej.
- Podpróbkowanie w celu zmniejszenia rozmiaru klasy większościowej.

03 Trening modelu

Przetestowano wiele modeli:

- RandomForest do prognozowania bazowego.
- XGBoost w celu poprawy wydajności, zoptymalizowany za pomocą funkcji Early Stopping i Feature Selection w celu skrócenia czasu szkolenia.

04 Ocena

Oceniono wyniki przy użyciu

- Recall do wykrywania klas mniejszościowych.
- Wynik F1 dla zrównoważonej dokładności między precyzją a przypomnieniem.

05 Wnioski

Kluczowe spostrzeżenia

01 Znaczenie stronniczości danych:

Nierównowaga w zbiorze danych spowodowała, że początkowe modele ignorowały klasę mniejszościową (klientów subskrybujących depozyty).

02 Kluczowe znaczenie mają techniki równoważenia:

Podpróbkowanie, wagi klas i SMOTE znacznie poprawiły przywołanie, choć kosztem ogólnej dokładności.

03 XGBoost z wyborem cech:

Dzięki zmniejszeniu liczby cech i dodaniu funkcji wczesnego zatrzymywania, XGBoost poprawił wydajność bez utraty efektywności.

Kluczowe wnioski

01 Etyczne projektowanie sztucznej inteligencji:

Etyczne projektowanie sztucznej inteligencji wymaga przemyślanego przygotowania zbioru danych, sprawiedliwych wskaźników oceny oraz świadomości potencjalnych błędów systematycznych w wynikach.

06 Bibliografia

- Zbiór danych dotyczących marketingu bankowego – repozytorium UCI Machine Learning Repository: https://archive.ics.uci.edu/dataset/222/bank+marketing
- Dokumentacja Scikit-learn: https://scikit-learn.org/
- Dokumentacja XGBoost: https://xgboost.readthedocs.io/en/stable/
- Dokumentacja Imbalanced-learn: https://imbalanced-learn.org/

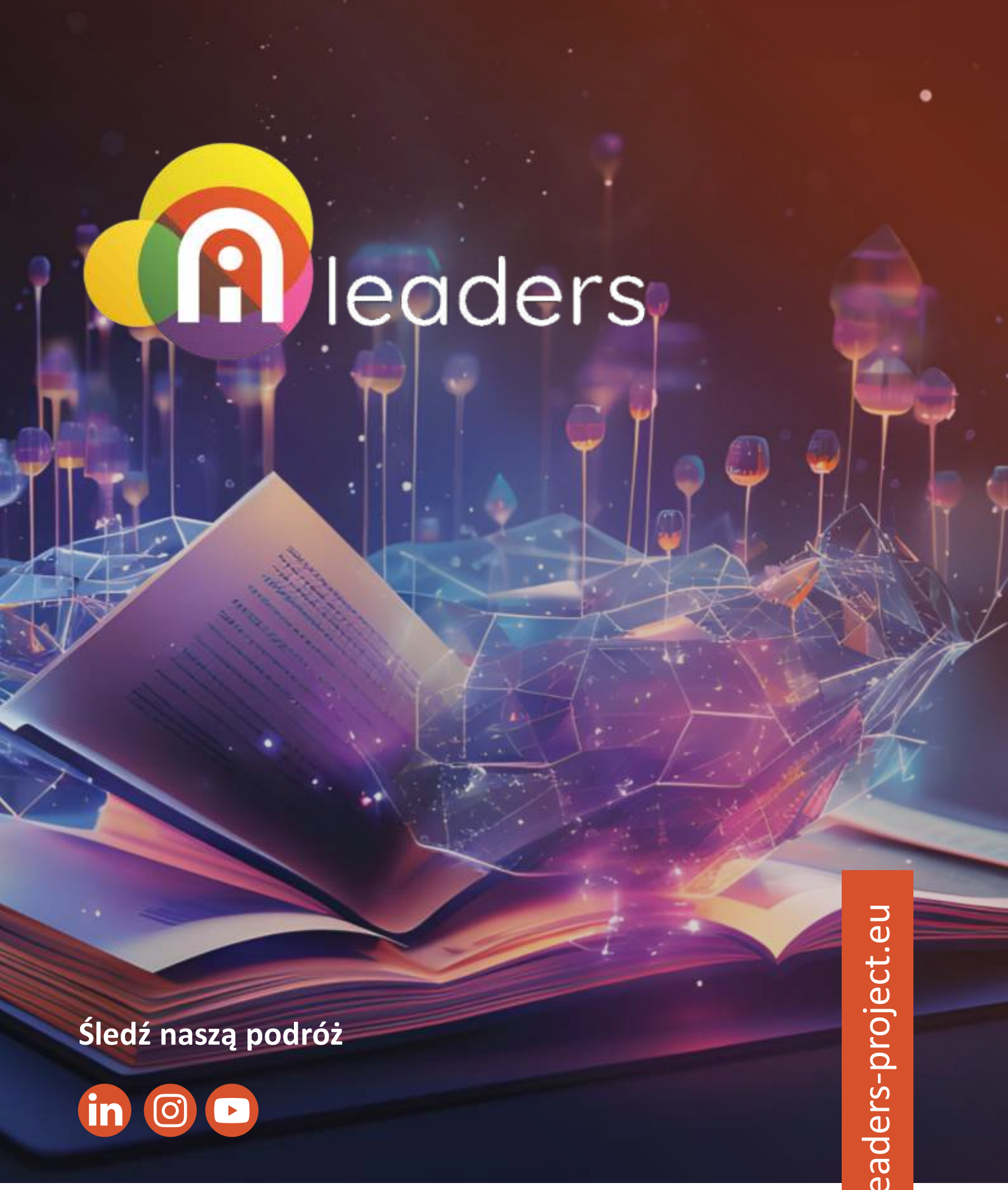
07 Materiały uzupełniające

Plik Jupyter Notebook (IPYNB)

Dołączono szczegółowy notatnik Python z komentowanym kodem, który przeprowadza użytkownika przez każdy etap procesu.

Notatnik zawiera

- Etapy czyszczenia i przetwarzania wstępnego danych.
- Inżynierię cech i logikę wyboru zmiennych.
- Wdrożenie różnych technik równoważenia (SMOTE, Class Weights i Undersampling).
- Trening modelu za pomocą RandomForest i XGBoost.
- Wskaźniki oceny i wnioski wynikające z rezultatów.



Śledź naszą podróż

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.