Zastosowanie modelowania matematycznego w bankowości

semestr letni 2024/25

Lista 2

Termin oddania sprawozdania: 22.04.2025 (do końca dnia).

🚀 Budowa modelu XGBoost na danych bankowych.

Twoim zadaniem jest zbudowanie modelu klasyfik1acyjnego XGBoost do przewidywania, czy klient zdecyduje się na lokatę terminową na podstawie cech demograficznych i informacji o kontakcie marketingowym.

Kroki do wykonania:

1. Wczytaj załączone do listy dane do Pythona i zapoznaj się z ich strukturą.
    - plik *dane_list2.csv* zawiera dane,
    - kolumna y jest zmienną docelową (wartości: *yes* – klient zgodził się na lokatę, *no* – klient odmówił).
2. Przygotuj dane do modelowania:
    - stwórz kolumnę jednoznacznie identyfikującą pojedynczą obserwację,
    - zamień kolumnę y na wartości binarne (1 = *yes*, 0 = *no*),
    - zamień zmienne kategoryczne na wartości numeryczne,
    - podziel dane na zbiór treningowy i testowy (np. 80/20).
3. Przeprowadź eksploracyjną analizę danych. Wytłumacz znaczenie danych, sprawdź rozkłady, sprawdź statystyki opisowe, poziom korelacji, występowanie duplikatów, pustych wartości i obserwacji odstających (*metoda Tukey fence dla chętnych), zależność ze zmienną objaśnianą, wyjaśnij co uznajemy za obserwację w próbach.
4. Zbuduj model XGBoost:
    a. w pierwszej kolejności zbuduj model na domyślnych hiperparametrach,
    b. ustal w jaki sposób będziesz weryfikować jakość budowanych modeli (miara GINI, miary jakości zmiennych),
    c. przeprowadź optymalizację hiperparametrów (np. za pomocą optymalizacji Bayesa),
    d. zweryfikuj wpływ użycia zmiennych typu *dummy* (0-1) na wyniki modelu,
    e. przeprowadź optymalizację liczby zmiennych wykorzystanych w modelu - zweryfikuj wpływ wykluczenia najmniej istotnych zmiennych na moc predykcyjną (np. możesz przygotować wykres, na którym przedstawiony będzie przyrost mocy GINI po uwzględnieniu w modelu kolejnych zmiennych od najsilniejszej do najsłabszej na zbiorze TRAIN i na tej podstawie ocenić czy wykluczenie danej zmiennej jest sensowne),
    f. wybierz najlepszy Twoim zdaniem model (model champion) i uzasadnij swoją decyzję,
    g. przeprowadź analizę jakości wybranego przez Ciebie modelu champion na zbiorach TRAIN oraz TEST (np. moc modelu – ogólna oraz na wybranych podpopulacjach, jakość zmiennych - feature importance, permutation importance, SHAP). Czy wyniki na zbiorze TEST różnią się od wyników na zbiorze TRAIN?,
    h. *wykorzystaj metodę walidacji krzyżowej (cross-validation) na dowolnym etapie budowy modelu - podpunkt dla chętnych,
    i. *ustal optymalny według Ciebie poziom odcięcia (cut-off) – podpunkt dla chętnych.

🚀 Opis zbioru danych – marketing bankowy

Dane dotyczą kampanii marketingowych portugalskiej instytucji bankowej. Kampanie marketingowe były prowadzone telefonicznie. Często konieczne było wykonanie więcej niż jednego kontaktu z tym samym klientem, aby ocenić, czy produkt (lokata terminowa) zostanie subskrybowany, czy nie. Zbiór danych zawiera wszystkie przykłady uporządkowane chronologicznie (od maja 2008 do listopada 2010).

Celem jest przewidzenie, czy klient zdecyduje się na subskrypcję lokaty terminowej (zmienna y).

Liczba rekordów (instancji): 45 211.

Liczba atrybutów: 16 + zmienna wyjściowa (czy klient założy lokatę).

Brakujące wartości: brak

Opis atrybutów

Zmienne wejściowe (input variables)

Dane klienta banku

1. age – wiek (numeryczna),
2. job – rodzaj pracy (kategoryczna: "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services"),
3. marital – stan cywilny (kategoryczna: "married" – zamężny/żonaty, "divorced" – rozwiedziony/wdowiec, "single" – singiel),
4. education – poziom wykształcenia (kategoryczna: "unknown", "secondary", "primary", "tertiary"),
5. default – czy klient ma zaległości w spłacie kredytu? (binarna: "yes", "no"),
6. balance – średnie roczne saldo na koncie klienta (w euro) (numeryczna),
7. housing – czy klient ma kredyt hipoteczny? (binarna: "yes", "no"),
8. loan – czy klient ma kredyt gotówkowy? (binarna: "yes", "no"),

Informacje o ostatnim kontakcie w bieżącej kampanii

9. contact – typ komunikacji (kategoryczna: "unknown", "telephone", "cellular"),
10. day – dzień ostatniego kontaktu w miesiącu (numeryczna),
11. month – miesiąc ostatniego kontaktu (kategoryczna: "jan", "feb", "mar", ..., "nov", "dec"),
12. duration – czas trwania ostatniego kontaktu (w sekundach) (numeryczna),

Inne zmienne

13. campaign – liczba kontaktów wykonanych w tej kampanii dla danego klienta (numeryczna, obejmuje ostatni kontakt),
14. pdays – liczba dni od ostatniego kontaktu klienta w poprzedniej kampanii (numeryczna, -1 oznacza, że klient nie był wcześniej kontaktowany),
15. previous – liczba kontaktów wykonanych przed tą kampanią dla danego klienta (numeryczna),
16. poutcome – wynik poprzedniej kampanii marketingowej (kategoryczna: "unknown", "other", "failure", "success").

Zmienna wyjściowa (target variable)

y – czy klient zdecydował się na subskrypcję lokaty terminowej? (binarna: "yes", "no").