

**run\_logistic\_regression** – klasyczna regresja logistyczna trenowana na pełnym zbiorze danych z opcjonalną inżynierią cech i kalibracją. Pełny zbiór, bez foldów, wysoka skuteczność na zbiorze treningowym

**run\_logistic\_regression\_kfold** – regresja logistyczna z walidacją K-Fold (stratyfikowaną) i tuningiem parametrów. Walidacja K-Fold - bardziej wiarygodna estymacja jakości, mniejszy overfit

**run\_logistic\_regression\_incremental** – regresja logistyczna uczona w trybie przyrostowym (incremental learning) z wykorzystaniem SGD, również z kalibracją i analizą cech na poziomie batcha i globalnie. Wersja batchowa, przyrostowa, nieco niższe wartości przy pierwszych batchach, poprawia się globalnie

## **Regresja logistyczna (run\_logistic\_regression i run\_logistic\_regression\_kfold)**

- **Model:** LogisticRegression (solver saga)
- **Cel:** klasyfikacja binarna (satisfaction = 0 lub 1)
- **Parametry:**
  - C – odwrotność regularyzacji; wyższa wartość → mniejsza regularyzacja
  - l1\_ratio – stosunek L1/L2 w regularizacji mieszanej (ElasticNet)
  - class\_weight – balansowanie klas (None lub "balanced")
  - max\_iter=10000 – zapewnienie zbieżności przy dużej liczbie cech
- **Kalibracja:**
  - CalibratedClassifierCV z metodą isotonic, CV=3
  - Cel: poprawa predykcji prawdopodobieństw (bardziej zgodnych z rzeczywistym rozkładem)
- **Inżynieria cech:**
  - **Pairwise linear:** tworzenie nowych cech jako sumy i różnice wybranych par cech (f1\_plus\_f2, f1\_minus\_f2)

- **Pairwise polynomial:** interakcje między cechami (PolynomialFeatures, stopień 2)
- **Top 10 features:** selekcja cech na podstawie współczynników regresji

## **Incremental learning (run\_logistic\_regression\_incremental)**

- Model: SGDClassifier z loss= log\_loss odpowiada regresji logistycznej
- Uczenie: batchowe, z użyciem partial\_fit
- Kalibracja: globalna po całym batchowym uczeniu, opcjonalnie CalibratedClassifierCV
- Inżynieria cech: analogiczna do klasycznych funkcji (pairwise\_poly, pairwise\_linear)
- Wyjaśnialność: obliczana po każdym batchu oraz globalnie
- Zapis modelu: joblib dla modelu bazowego, kalibrowanego i danych już wykorzystanych

Modele klasyczne (full data) mają wyższe wyniki na zbiorze treningowym, ale K-Fold pokazuje realną generalizację.

Incremental learning daje możliwość dalszego uczenia przy nowych danych, kosztem początkowej dokładności batchowej.

## **Wyjaśnialność modeli**

- Wszystkie modele posiadają interpretację współczynników regresji:
  - Wysokie wartości bezwzględne - cecha istotna dla klasyfikacji
  - Dodatnie wartości - zwiększą prawdopodobieństwo satisfaction=1
  - Ujemne wartości - zmniejszą prawdopodobieństwo satisfaction=1
- Top 10 pozytywne i negatywne cechy są wizualizowane w formie barplotów
- Incremental learning dodatkowo wylicza top cechy po każdym batchu, co umożliwia monitorowanie wpływu nowych danych

## **Wnioski:**

- Pozwala zrozumieć, które zmienne realnie wpływają na satysfakcję
- Analiza interakcji i pairwise polynomial daje lepsze dopasowanie modelu i głębszą interpretację
- Top 10 cech umożliwia prezentację wyników

## **Regularyzacja i tuning**

- C - regularyzacja L2/L1 (elastic net)
- l1\_ratio - udział L1 w regularyzacji mieszanej
- class\_weight - balans klas
- random\_search - szybka eksploracja parametrów

## **Efekt:**

- Modele są stabilne i nie przeuczają się przy dużej liczbie cech (szczególnie po pairwise polynomial)
- Incremental learning wymaga batch size = np. 25 - pozwala na stopniową naukę

## Podsumowanie i zalety trzech metod

Funkcja	Zalety	Ograniczenia
run_logistic_regression	szybkie dopasowanie pełnego modelu, wyjaśnialność, prosty w użyciu	brak walidacji, możliwy overfit
run_logistic_regression_kfold	wiarygodna ocena modelu, top features i interakcje, tuning parametrów	dłuższy czas treningu, duża liczba cech spowalnia poly
run_logistic_regression_incremental	umożliwia przyrostowe uczenie, batchowa analiza cech, zachowanie wiedzy	wymaga właściwego skalera i poly dla nowych danych, wyniki początkowe batchowe są niższe

## Ogólne wnioski:

- Predykcja zmiennej celu, analiza wpływu zmiennych, wyjaśnialność modeli.
- Incremental learning stanowi wartość dodaną dla rzeczywistych scenariuszy, gdzie dane napływają stopniowo..

## Przypuszczalne wnioski biznesowe

- Najważniejsze cechy wpływające na satysfakcję to Average Service Score, czas oczekiwania i kluczowe interakcje między cechami.
- Modele klasyczne dają szybki wgląd w predykcję satysfakcji, natomiast incremental learning pozwala monitorować zmiany wpływu cech wraz z nowymi danymi.
- Analiza top cech i interakcji pozwala przygotować rekomendacje dla zespołu operacyjnego lub marketingowego (np. optymalizacja czasu obsługi, priorytetyzacja interakcji).