Klasyfikacja satysfakcji pasażerów linii lotniczych

Modele uczenia maszynowego

Jakub Mikołajczak, Patryk Piasecki

11 czerwca 2026

Cel analizy

Problem

Przewidujemy, czy pasażer będzie zadowolony z podróży na podstawie cech demograficznych, operacyjnych i ocen jakości usług.

- Zadanie: klasyfikacja binarna zmiennej `satisfaction`.
- Klasa pozytywna: `satisfied`.
- Wynik modelu: prawdopodobieństwo przynależności do klasy lub decyzja klasyfikacyjna.
- Sens praktyczny: wskazać, które modele najlepiej przewidują satysfakcję i które cechy są najważniejsze dla klasyfikacji.

Zmienne wykorzystane w modelach

Cechy pasażera i podróży

- `Gender`, `Customer Type`, `Age`
- `Type of Travel`, `Class`
- `Flight Distance`
- `Departure Delay`, `Arrival Delay`

Oceny usług

- WiFi, booking online, online boarding
- seat comfort, leg room, cleanliness
- food and drink, entertainment
- check-in, baggage, on-board i inflight service

Dlaczego te zmienne?

Do modelu trafiają wszystkie predyktory po usunięciu pól technicznych i zmiennej celu. Nie dobieramy ich ręcznie pod wynik, tylko pozwalamy modelom ocenić ich użyteczność.

Interpretacja

Zmienne obejmują zarówno cechy obiektywne, jak i subiektywne oceny doświadczenia pasażera.

Dane i obowiązujący podział

Etap	Liczba obserwacji
Połączony zbiór	129 880
Zbiór treningowy	103 904
Zbiór testowy	25 976

Klasa	Kod
`satisfied`	1
`neutral/dissatisfied`	0

Interpretacja

Połączony zbiór został ponownie podzielony na część treningową i testową. Parametr `stratify=y` zachowuje podobny udział pasażerów zadowolonych i niezadowolonych w obu próbach, a `random_state=42` zapewnia powtarzalność podziału.

Czyszczenie danych i przeciek informacji

- Usunięto kolumny techniczne: `id`, `Unnamed: 0` oraz pomocnicze oznaczenia pliku źródłowego.
- Braki danych występowały tylko w `Arrival Delay in Minutes`: 393 obserwacje.
- Imputacja została umieszczona w potoku przetwarzania (`Pipeline`), czyli nie uzupełniamy braków ręcznie przed podziałem danych.
- Mediana dla zmiennej liczbowej jest wyznaczana na zbiorze treningowym, a następnie stosowana do zbioru testowego.

Dlaczego to ważne?

Zbiór testowy ma symulować nowe dane. Gdyby mediana, skalowanie albo kodowanie były dopasowane na całym zbiorze, model pośrednio korzystałby z informacji testowej. To jest przeciek informacji (*data leakage*) i mogłoby sztucznie zawyżyć jakość predykcji.

Przetwarzanie wstępne zmiennych

Przetwarzanie wstępne zamienia dane surowe na postać możliwą do użycia przez modele: uzupełnia braki, skaluje zmienne liczbowe i zamienia kategorie tekstowe na zmienne zero-jedynkowe.

Modele drzewiaste

- imputacja medianą i dominantą,
- `OneHotEncoder(...)` dla kategorii,
- brak skalowania, bo drzewa używają podziałów typu `Age < 35`.

Logit i MLP

- imputacja braków,
- standaryzacja zmiennych liczbowych,
- kodowanie kategorii na zmienne 0/1.

`OneHotEncoder` zamienia np. `Class = Eco/Eco Plus/Business` na zmienne typu `Class_Eco`, `Class_EcoPlus`, `Class_Business`.

Dlaczego te modele?

Model	Rodzina	Rola w projekcie
Logit	model liniowy	interpretacja i punkt odniesienia
Random Forest	bagging drzew	mocny benchmark drzewiasty
HistGradientBoosting	boosting drzew	główny kandydat predykcyjny
MLP	sieć neuronowa	nieliniowy benchmark parametryczny

Interpretacja

Modele różnią się nie tylko jakością predykcji, ale też tym, co można z nich wyjaśnić. Dlatego Logit traktujemy interpretacyjnie, a modele nieliniowe głównie predykcyjnie.

Logit jako model interpretacyjny

- Regresja logistyczna modeluje logarytm szans przynależności do klasy pozytywnej, czyli `satisfied`.
- Po zastosowaniu funkcji wykładniczej do współczynnika, `exp(coef)`, otrzymujemy iloraz szans.
- Iloraz szans większy od 1 oznacza wzrost szans klasy `satisfied`; wartość mniejsza od 1 oznacza ich spadek.
- Model jest czytelny interpretacyjnie, ale zakłada liniową zależność w logarytmie szans.

Najsilniejsze współczynniki Logitu

Cecha	coef	Iloraz szans	Interpretacja
Personal travel	-1.687	0.185	ok. 81.5% niższe szanse klasy `satisfied`; interpretacja kategorii jest pomocnicza
Disloyal customer	-1.330	0.264	ok. 73.6% niższe szanse klasy `satisfied`
Business travel	1.053	2.868	ok. 2.87 razy większe szanse klasy `satisfied`
Online boarding	0.818	2.267	wzrost oceny o 1 punkt wiąże się z ok. 2.27 razy większymi szansami
Loyal customer	0.697	2.008	ok. 2.01 razy większe szanse klasy `satisfied`
Inflight wifi	0.527	1.694	wzrost oceny o 1 punkt wiąże się z ok. 1.69 razy większymi szansami
Eco Plus	-0.519	0.595	ok. 40.5% niższe szanse klasy `satisfied`
Check-in service	0.418	1.519	wzrost oceny o 1 punkt wiąże się z ok. 1.52 razy większymi szansami

Tabela nie oznacza ręcznego wyboru zmiennych do Logitu. Model otrzymał wszystkie predyktory po usunięciu pól technicznych i zmiennej celu; pokazano tu zmienne o największych wartościach bezwzględnych współczynników po przetworzeniu danych. Nie raportujemy p-value: użyta implementacja `sklearn` służy predykcji.

Ograniczenia Logitu

- Relacje między ocenami usług a satysfakcją mogą być nieliniowe.
- Interakcje, np. typ podróży $\times$ klasa biletu, nie są automatycznie uchwycone.
- Współliniowość zmiennych oznacza, że niektóre predyktory niosą podobną informację.
- Przykład: opóźnienie odlotu i opóźnienie przylotu są silnie powiązane, więc trudno jednoznacznie przypisać wpływ do jednej zmiennej.

Rola w projekcie

Logit jest słabszy od modeli nieliniowych, ale nadal bardzo dobry predykcyjnie i użyteczny interpretacyjnie.

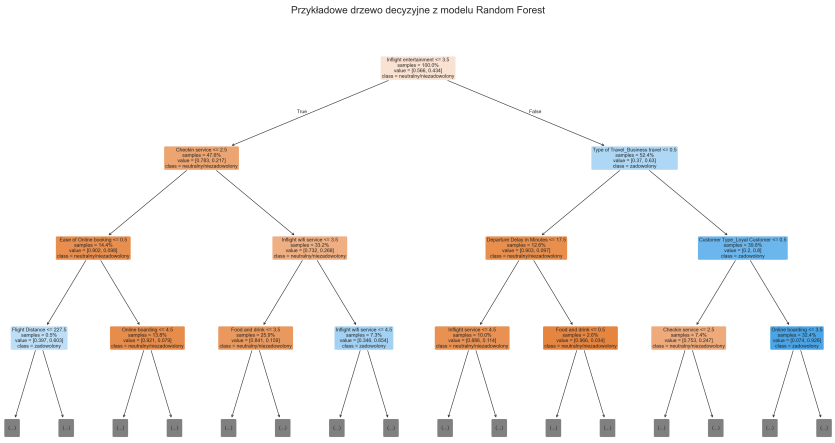
Random Forest jako model baggingowy

- Użyto `n_estimators=200`, `max_depth=15`, `min_samples_leaf=10`, `random_state=42`.
- Nie zastosowano `class_weight`; klasy są umiarkowanie niezbilansowane, ale nie skrajnie rzadkie.
- Las losowy uśrednia wiele drzew, co redukuje wariancję i dobrze obsługuje nieliniowości.
- W projekcie osiągnął bardzo wysoką jakość, ale był słabszy od HistGradientBoosting pod względem ROC AUC, accuracy i F1.

Interpretacja

Ważność cech dla Random Forest można analizować przez permutation importance, ale ta metoda nie daje bezpośredniego znaku wpływu zmiennej. Mierzy tylko spadek jakości po przetasowaniu cechy; nie rozróżnia, czy większa wartość cechy podnosi czy obniża prawdopodobieństwo klasy.

Przykładowe drzewo z Random Forest



Drzewo pokazuje kolejne reguły podziału obserwacji. W pełnym lesie decyzja jest uśrednieniem wielu takich drzew, dlatego pojedyncze drzewo jest ilustracją mechanizmu, a nie pełną interpretacją modelu.

HistGradientBoosting jako histogramowy gradient boosting

Model buduje kolejne drzewa sekwencyjnie. Każde następne drzewo poprawia błędy poprzedniego modelu. Wariant histogramowy przyspiesza obliczenia przez grupowanie wartości cech w przedziały.

Parametry użyte w notebooku

`learning_rate=0.08`, `max_iter=180`, `max_leaf_nodes=31`, `l2_regularization=0.01`, `min_samples_leaf=20`, `random_state=42`.

Interpretacja wyniku

To najlepszy model predykcyjny w projekcie: uzyskał najwyższe ROC AUC, F1 i accuracy. Najlepiej porządkował pasażerów według prawdopodobieństwa satysfakcji i najlepiej klasyfikował przy przyjętym progu.

MLP – prosta sieć neuronowa jako benchmark

- Zapis `(64, 32)` oznacza dwie warstwy ukryte: pierwsza ma 64 neurony, druga 32 neurony.
- Sieć uczy się nieliniowych kombinacji cech, ale nie daje prostej interpretacji pojedynczych współczynników.
- MLP wymaga standaryzacji, bo zmienne o dużej skali mogą dominować gradient i spowalniać albo destabilizować uczenie wag.
- Na danych tabelarycznych progi, interakcje i zmienne kategoryczne często są bardzo dobrze obsługiwane przez drzewa oraz boosting.
- Prosta sieć neuronowa nie ma tu przewagi architektonicznej znanej np. z obrazów lub tekstu, gdzie struktura danych jest silnie przestrzenna albo sekwencyjna.
- W tej prezentacji MLP traktujemy jako model praktycznie nieinterpretowalny i oceniamy go głównie predykcyjnie.

Metryki klasyfikacji

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

- Klasą pozytywną jest `satisfied`.
- F1 jest średnią harmoniczną precision i recall.
- F1 jest wysokie tylko wtedy, gdy zarówno precision, jak i recall są wysokie.
- ROC AUC mierzy zdolność rankingową modelu.

Uwaga

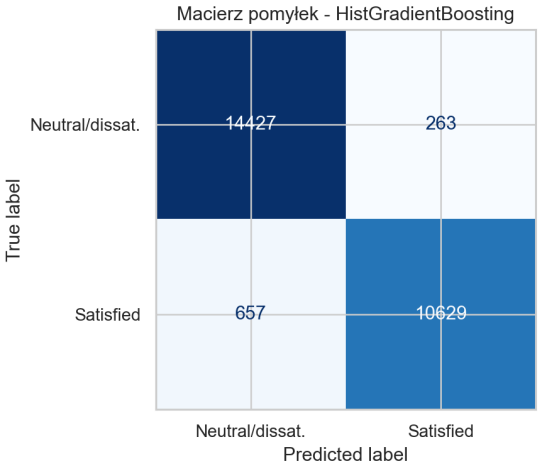
ROC jest krzywą, a AUC jej liczbowym podsumowaniem. Dlatego w tabeli wyników raportujemy ROC AUC.

Główne wyniki

Model	Accuracy	Precision	Recall	F1	ROC AUC	Czas
HistGradientBoosting	0.9646	0.9759	0.9418	0.9585	0.9952	6.35 s
MLP	0.9580	0.9680	0.9343	0.9508	0.9933	27.37 s
Random Forest	0.9548	0.9571	0.9382	0.9475	0.9924	40.56 s
Logit	0.8741	0.8699	0.8352	0.8522	0.9272	1.13 s

HGB ma najlepszy wynik ogólny. MLP jest bardzo dobry, ale słabszy od boostingu i trudniejszy do interpretacji. Random Forest pozostaje mocnym benchmarkiem, a Logit mimo niższego wyniku ROC AUC nadal jest wartościowy interpretacyjnie. Precision 0.9759 dla HGB oznacza, że predykcje klasy `satisfied` są bardzo trafne, a recall 0.9418 oznacza, że model wykrywa większość zadowolonych pasażerów.

Macierz pomyłek najlepszego modelu



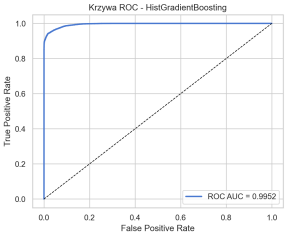
- Model poprawnie klasyfikuje 14 427 obserwacji klasy 0.
- Błędnie oznacza 263 osoby jako `satisfied`.
- Błędnie pomija 657 rzeczywiście zadowolonych pasażerów.
- Accuracy na próbie testowej: 0.9646.

Raport klasyfikacji: HistGradientBoosting

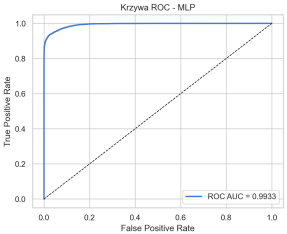
Klasa	Precision	Recall	F1	Support
neutral/dissat.	0.9564	0.9821	0.9691	14 690
satisfied	0.9759	0.9418	0.9585	11 286
accuracy			0.9646	25 976
macro avg	0.9661	0.9619	0.9638	25 976
weighted avg	0.9649	0.9646	0.9645	25 976

Accuracy jest jedną miarą dla całego modelu, dlatego nie ma osobnych wartości precision i recall. Dla klasy `neutral/dissat.` recall 0.9821 oznacza, że model poprawnie wykrywa 98.21% takich pasażerów. Dla klasy `satisfied` precision 0.9759 oznacza, że gdy model przewiduje satysfakcję, zwykle jest to trafne. Macro average to prosta średnia wyników obu klas, więc każda klasa ma tę samą wagę. Weighted average waży wynik liczebnością klas, dlatego większa klasa mocniej wpływa na średnią.

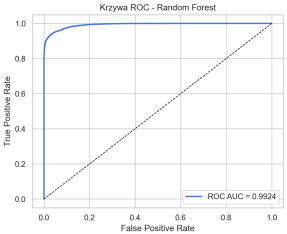
Krzywe ROC



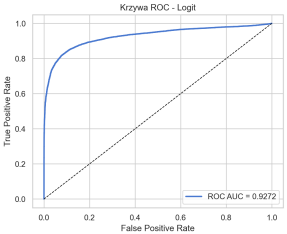
HistGradientBoosting



MLP



Random Forest



Logit

Logit jest wyraźnie słabszy od modeli nieliniowych, ale sam w sobie nadal osiąga bardzo dobry ROC AUC. Różnica sugeruje, że modele nieliniowe lepiej wykorzystują interakcje i efekty progowe w danych.

Czy bardzo wysokie wyniki oznaczają przeuczenie?

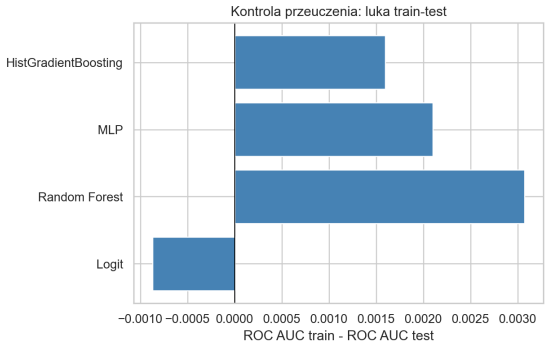
- ROC AUC = 0.9952 jest bardzo wysokim wynikiem, ale nie oznacza automatycznie przeuczenia.
- Może wynikać z silnego sygnału w danych, łatwego problemu klasyfikacyjnego albo przecieku informacji.
- Ryzyko przecieku ograniczamy przez dwa elementy: osobny zbiór testowy oraz potoki przetwarzania dopasowywane tylko na zbiorze treningowym.
- Uzupełnianie braków, skalowanie i kodowanie kategorii są na zbiorze testowym tylko stosowane, a nie ponownie dopasowywane.

Konsekwencja

Model nie wykorzystuje statystyk obliczonych na danych testowych, ale bardzo wysokie wyniki nadal wymagają kontroli train-test.

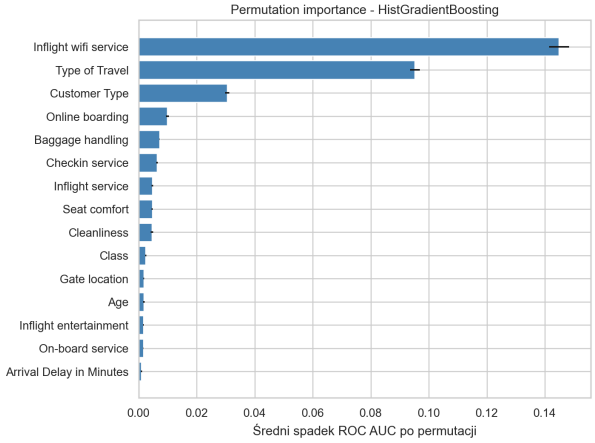
Kontrola przeuczenia

Model	ROC train	ROC test	Luka
HGB	0.9968	0.9952	0.0016
MLP	0.9954	0.9933	0.0021
RF	0.9954	0.9924	0.0031
Logit	0.9264	0.9272	-0.0009



Mała luka train-test sugeruje, że model nie dopasował się wyłącznie do próby treningowej. Nie dowodzi jednak, że będzie równie dobry na danych z innej linii lotniczej, okresu lub rynku. Pełna ocena wymagałaby walidacji zewnętrznej.

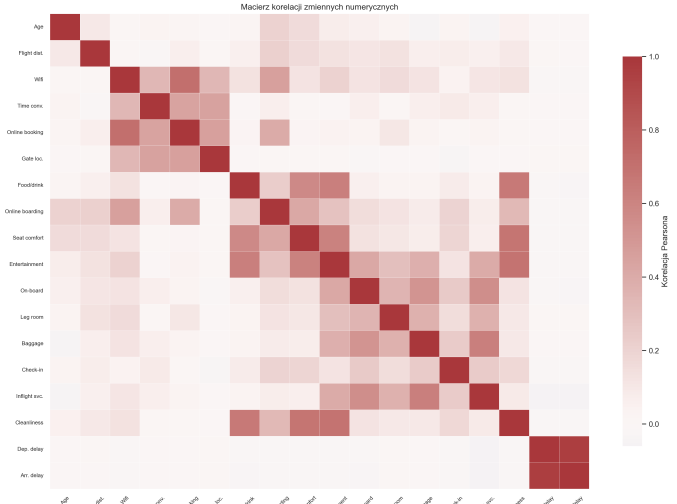
Ważność cech: HistGradientBoosting



Cecha	Spadek ROC AUC
Inflight wifi	0.1449
Type of Travel	0.0952
Customer Type	0.0305
Online boarding	0.0098
Baggage handling	0.0071
Checkin service	0.0063

Permutation importance sprawdza, jak bardzo pogarsza się ROC AUC po losowym przemieszaniu wartości jednej cechy. Wysoka ważność WiFi oznacza, że model silnie używa tej zmiennej, ale nie jest to dowód przyczynowego wpływu WiFi na satysfakcję.

Heatmapa korelacji



Jak czytać heatmapę korelacji?

Interpretacja

Heatmapa pokazuje siłę liniowego skorelowania między zmiennymi liczbowymi. Ciemniejsze pola oznaczają silniejszą korelację dodatnią lub ujemną.

- Szczególnie interesują nas pary zmiennych silnie skorelowanych.
- Silne skorelowanie może oznaczać częściowe powielanie informacji.
- Jest to ważne przede wszystkim dla interpretacji współczynników Logitu.

Pary o wysokiej korelacji

Zmienna A	Zmienna B	Korelacja
Departure Delay	Arrival Delay	0.9653
Inflight wifi service	Ease of Online booking	0.7148

Jeżeli dwie zmienne są silnie skorelowane, modelowi trudniej rozdzielić, która z nich odpowiada za zmianę prawdopodobieństwa. Współczynniki mogą być wtedy mniej stabilne. Nie musi to jednak pogarszać predykcji, bo model nadal wykorzystuje wspólną informację zawartą w tych zmiennych.

Predykcja a interpretowalność

Model	Jakość predykcji	Interpretowalność	Rola
Logit	niższa	wysoka	model wyjaśniający
Random Forest	wysoka	średnia/niska	benchmark drzewiasty
HistGradientBoosting	najwyższa	niska/średnia	model predykcyjny
MLP	wysoka	niska	benchmark nieliniowy

Wniosek

Jeden model rzadko jednocześnie najlepiej przewiduje i najlepiej wyjaśnia. W projekcie Logit pełni rolę interpretacyjną, a HistGradientBoosting predykcyjną.

Decyzja modelowa i wnioski biznesowe

Decyzja modelowa

- Do predykcji wybieramy HistGradientBoosting.
- Do interpretacji warto równolegle używać Logitu i permutation importance.
- MLP nie daje przewagi nad boostingiem.
- Random Forest pozostaje mocnym benchmarkiem.
- Przed wdrożeniem potrzebne są kalibracja, próg decyzyjny i walidacja na nowych danych.

Wnioski biznesowe

- Kluczowe obszary: WiFi, online boarding, typ podróży i typ klienta.
- WiFi traktować jako element bazowego doświadczenia.
- Segmentować pasażerów biznesowych i prywatnych.
- Rozwijać program lojalnościowy także jako narzędzie poprawy doświadczenia.
- Usprawniać cyfrowe punkty styku, szczególnie online boarding..
- Model może wspierać wskazywanie pasażerów wymagających uwagi.

Wnioski i ograniczenia

Wnioski

- Najlepszy model: HistGradientBoosting.
- Logit jest słabszy od modeli nieliniowych, ale nadal bardzo dobry i interpretacyjnie użyteczny.
- Modele nieliniowe wskazują na interakcje i efekty progowe w danych.
- Małe luki train-test nie wskazują na silne przeuczenie.

Ograniczenia

- Dane przekrojowe nie dowodzą przyczynowości.
- Brakuje ceny biletu, trasy, przewoźnika, lotniska i okresu.
- Bardzo wysokie wyniki wymagają walidacji zewnętrznej.
- Wdrożenie wymaga kalibracji i analizy progu decyzyjnego.