Predykcja anulacji rezerwacji hotelowych

Projekt Grupa 2

Problem biznesowy

Internetowe kanały rezerwacji hoteli radykalnie zmieniły możliwości rezerwacji i zachowania klientów. Znaczna liczba rezerwacji hotelowych jest nierealizowana z powodu anulowania lub niedojazdów. Typowe przyczyny anulowania obejmują zmianę planów, konflikty harmonogramów itp. Jest to często łatwiejsze dzięki możliwości zrobienia tego bezpłatnie lub po niskich kosztach, co jest korzystne dla gości hotelowych, ale jest to mniej pożądany i prawdopodobnie zmniejszający przychody czynnik dla hoteli.

Naszym celem jest predykcja potencjalnych anulacji rezerwacji hotelowych. Pozwoli to właścicielom zoptymalizować ceny/warunki rezerwacji.

Opis zmiennych

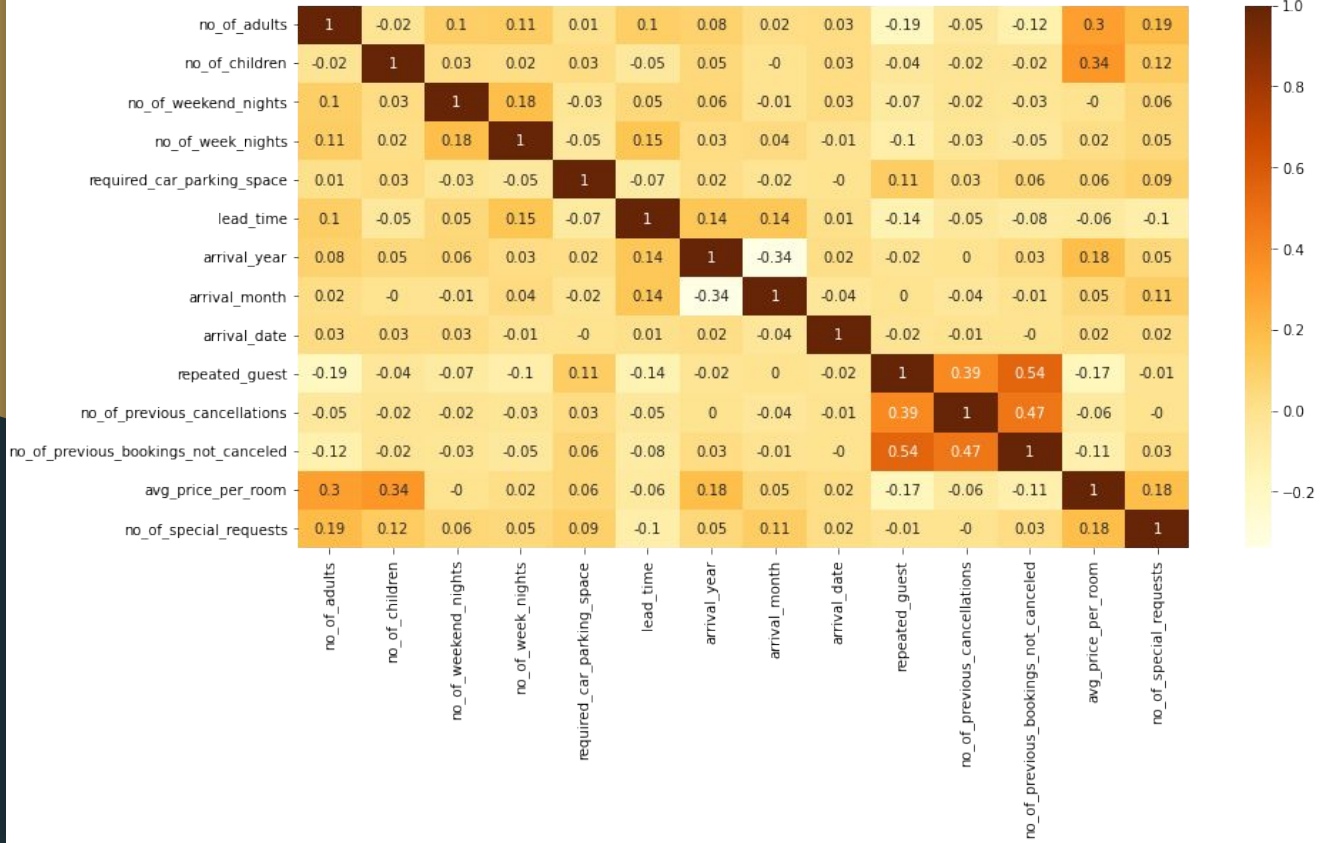
- **Booking_ID**: unikatowy numer identyfikacyjny rezerwacji
- **no_of_adults**: liczba dorosłych
- **no_of_children**: liczba dzieci
- **no_of_weekend_nights**: liczba nocy zarezerwowanych w weekend
- **no_of_week_nights**: liczba nocy zarezerwowanych w dni robocze
- **type_of_meal_plan**: rodzaj zarezerwowanego planu wyżywienia
- **required_car_parking_space**: informacja, czy klient zarezerwował parking (0 - nie, 1 - tak)
- **room_type_reserved**: rodzaj zarezerwowanego pokoju, wartości zostały zaszyfrowane przez INN Hotels
- **lead_time**: liczba dni między datą dokonania rezerwacji a pierwszym dniem pobytu
- **arrival_year**: rok pobytu
- **arrival_month**: miesiąc pobytu
- **arrival_date**: dzień miesiąca pobytu
- **market_segment_type**: oznaczenie segmentu rynku
- **repeated_guest**: informacja czy klient był wcześniej gościem hotelu (0 - nie, 1- tak)
- **no_of_previous_cancellations**: liczba wcześniej odwołanych rezerwacji przez klienta
- **no_of_previous_bookings_not_canceled**: liczba wcześniej nie odwołanych rezerwacji przez klienta
- **avg_price_per_room**: średnia cena za dobę, w EUR
- **no_of_special_requests**: liczba specjalnych życzeń klienta
- **booking_status**: informacja o odwołaniu lub nieodwołaniu rezerwacji

Typy danych i unikatowe wartości

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36275 entries, 0 to 36274
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   Booking_ID                            36275 non-null  object 
 1   no_of_adults                          36275 non-null  int64  
 2   no_of_children                        36275 non-null  int64  
 3   no_of_weekend_nights                  36275 non-null  int64  
 4   no_of_week_nights                     36275 non-null  int64  
 5   type_of_meal_plan                     36275 non-null  object 
 6   required_car_parking_space            36275 non-null  int64  
 7   room_type_reserved                    36275 non-null  object 
 8   lead_time                             36275 non-null  int64  
 9   arrival_year                          36275 non-null  int64  
 10  arrival_month                         36275 non-null  int64  
 11  arrival_date                          36275 non-null  int64  
 12  market_segment_type                   36275 non-null  object 
 13  repeated_guest                        36275 non-null  int64  
 14  no_of_previous_cancellations          36275 non-null  int64  
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64  
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64  
 18  booking_status                        36275 non-null  object 
dtypes: float64(1), int64(13), object(5)
memory usage: 5.3+ MB
```

```
Booking_ID                            36275
no_of_adults                              5
no_of_children                            6
no_of_weekend_nights                      8
no_of_week_nights                        18
type_of_meal_plan                         4
required_car_parking_space                2
room_type_reserved                        7
lead_time                               352
arrival_year                              2
arrival_month                            12
arrival_date                             31
market_segment_type                       5
repeated_guest                            2
no_of_previous_cancellations              9
no_of_previous_bookings_not_canceled     59
avg_price_per_room                     3930
no_of_special_requests                    6
booking_status                            2
dtype: int64
```

Macierz korelacji

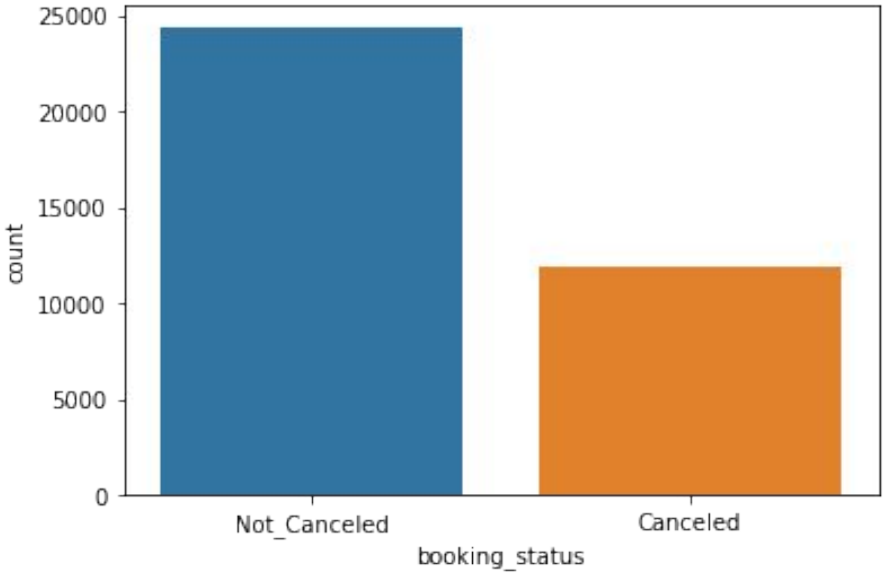


Brak silnej korelacji pomiędzy zmiennymi - cechy nie są współliniowe

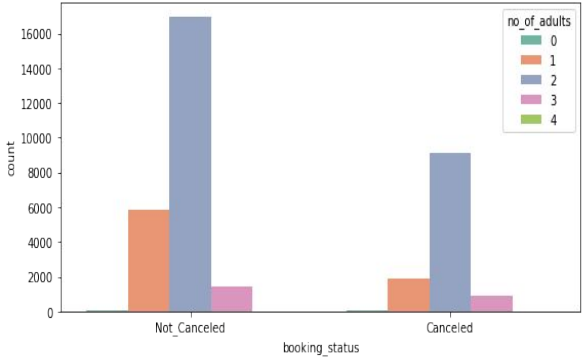
Istnieje zależność pomiędzy ceną za pokój a liczbą dorosłych i dzieci

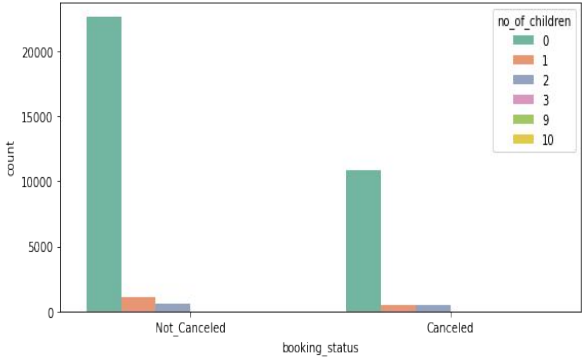
Rozkład statusu rezerwacji

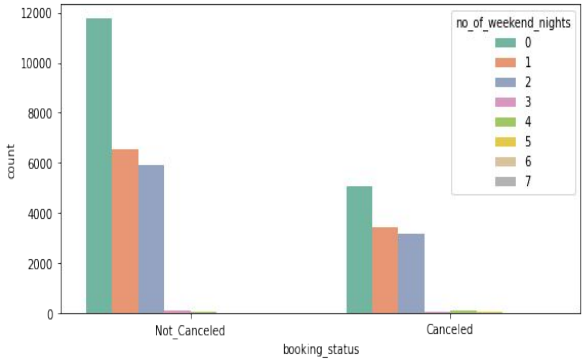
W zbiorze znajduje się dwukrotnie więcej nieodwołanych rezerwacji niż rezerwacji odwołanych.
Informacja szczególnie istotna przy podziale na zbiór treningowy i testowy.

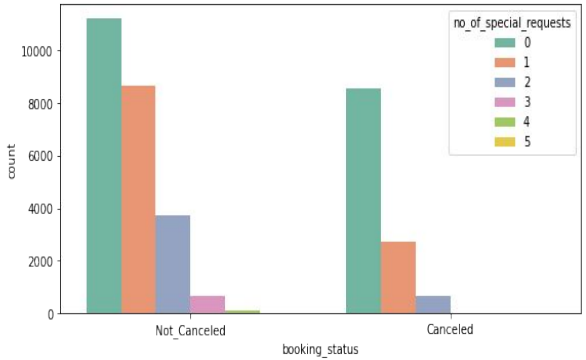


Analiza zmiennych wpływających na odwołanie rezerwacji







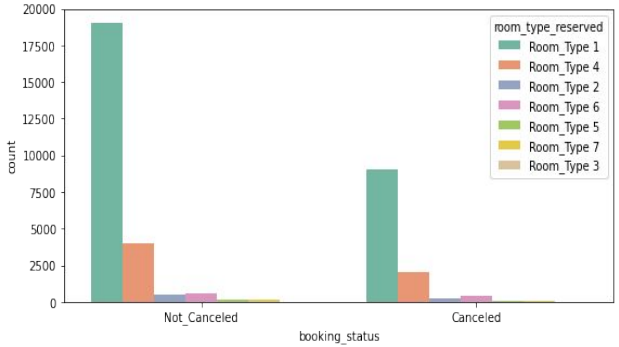


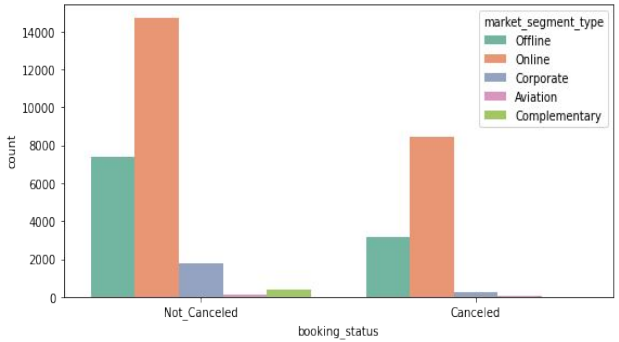
Najwięcej rezerwacji wykonywanych jest dla pary bez dzieci. Przy rezerwacji dla dwójki dzieci wzrasta prawdopodobieństwo anulacji.

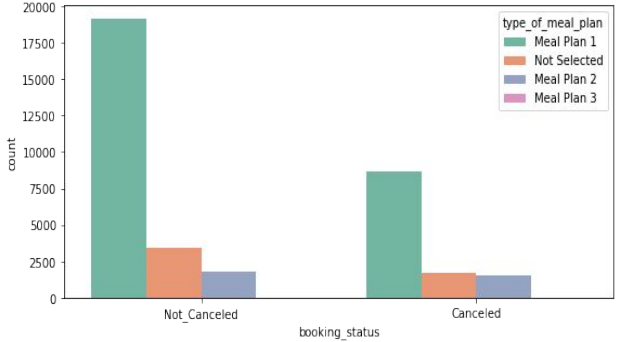
Proporcja dla odwołanych i nieodwołalnych rezerwacji weekendowych zostaje zachowana.

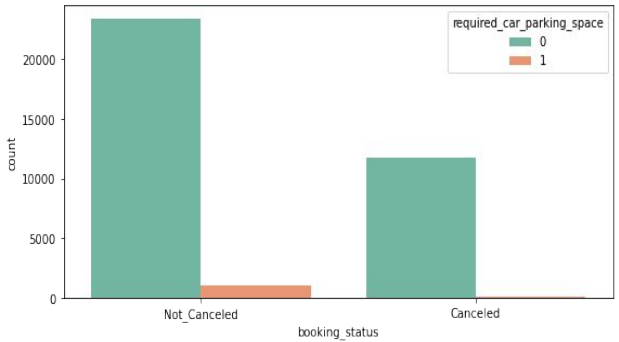
Im więcej spersonalizowanych ofert tym prawdopodobieństwo anulacji jest mniejsze.

Analiza zmiennych wpływających na odwołanie rezerwacji









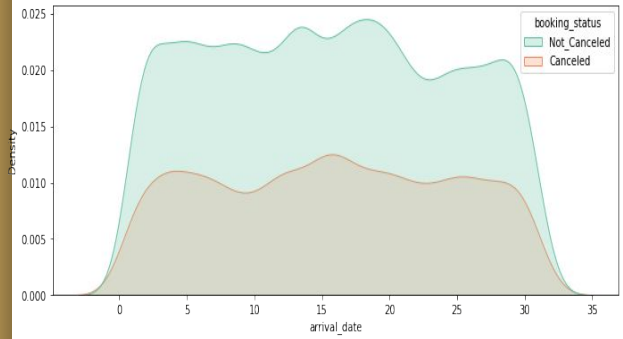
Najwięcej rezerwacji dla typu pokoju 1 natomiast szansa na anulacje rezerwacji wzrasta dla typu 6

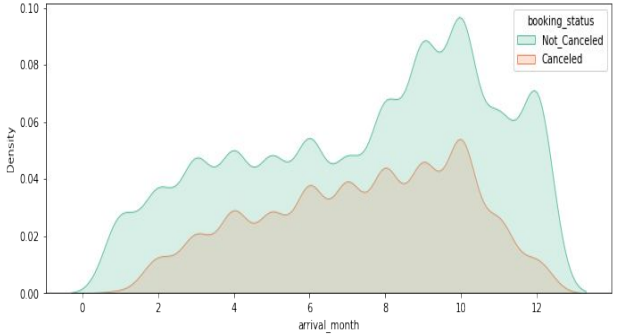
Dla rezerwacji korporacyjnych istnieje niewielka szansa na odwołanie rezerwacji - mały odsetek odwołanych rezerwacji

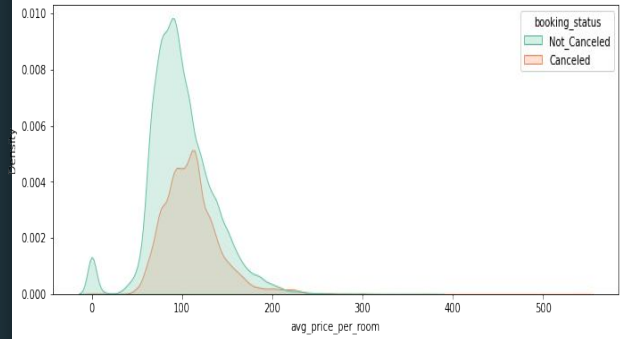
Przy wyborze drugiego planu wyżywienia wzrasta możliwość anulacji rezerwacji

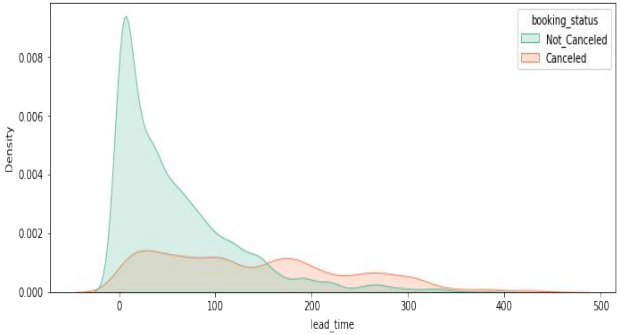
Częściej dokonywane są rezerwacje bez dostępności parkingu

Analiza zmiennych wpływających na odwołanie rezerwacji









Rozkład rezerwacji anulowanych i nieanulowanch rozkłada się równomiernie w perspektywie dnia przyjazdu.
Najmniej rezerwacji nieanulowanych ma miejsce ok. 25 dnia miesiąca

W drugiej połowie roku realizowanych jest najwięcej rezerwacji. Ryzyko odwołania rezerwacji przypada na miesiące wakacyjne. Od listopada najwięcej jest rezerwacji nieanulowanych

Najczęstsze rezerwacje są w granicach cen 50-100 euro.

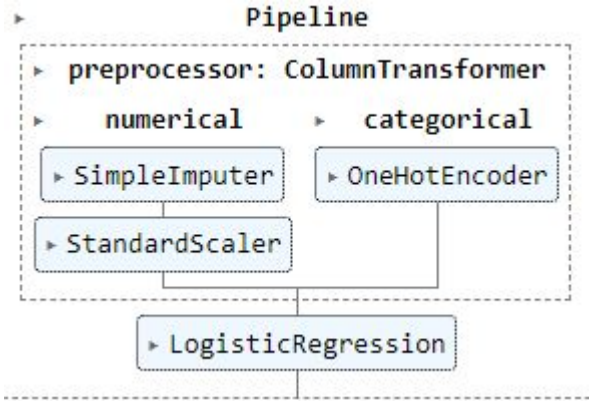
Im krótszy czas pomiędzy data rezerwacji, a data przybycia tym mniejsza szansa na anulacje rezerwacji
przy rezerwacji z wyprzedzeniem powyżej 150 jest większe prawdopodobieństwo anulacji rezerwacji niż jej zrealizowania

Zastosowane modele i wykorzystane zmienne:

1) Logistic Regression
2) Dummy Classifier
3) Random Forest Classifier

```
x_cols= ['no_of_adults', 'no_of_children', 'no_of_weekend_nights', 'no_of_week_nights', 'required_car_parking_space', 'lead_time',
         'repeated_guest', 'no_of_previous_cancellations', 'no_of_previous_bookings_not_canceled', 'avg_price_per_room',
         'no_of_special_requests', 'type_of_meal_plan', 'room_type_reserved', 'market_segment_type']
y_col = 'booking_status'
```

Logistic Regression

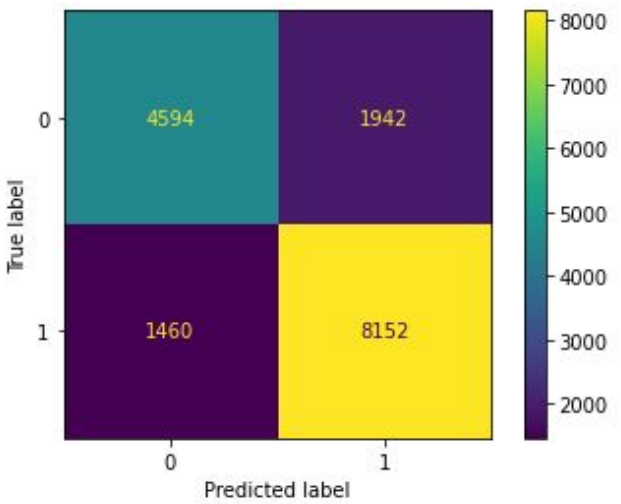


Metryki:

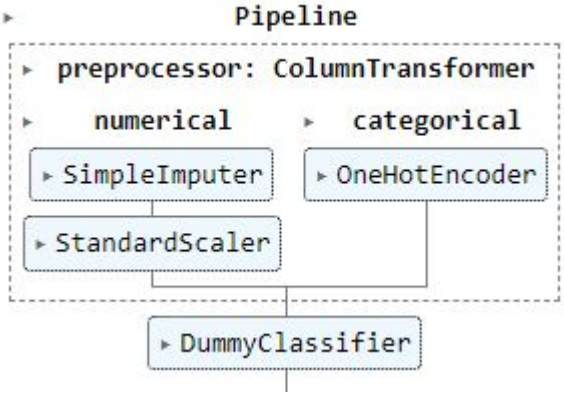
accuracy = 0.7893237552638097

F1 = 0.7297855440826052

confusion_matrix:



Dummy Classifier

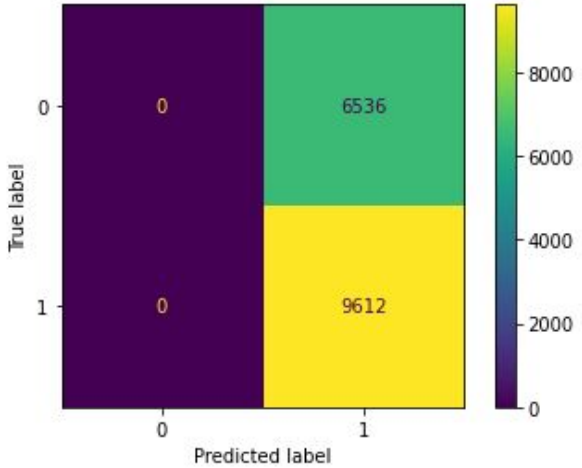


Metryki:

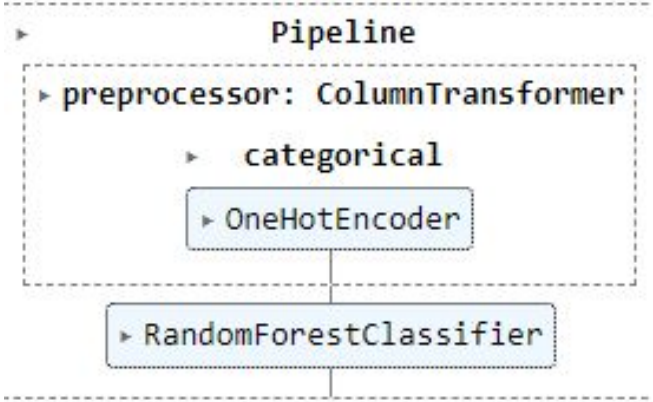
accuracy = 0.5952439930641565

F1 = 0.0

confusion_matrix:



Random Forest Classifier

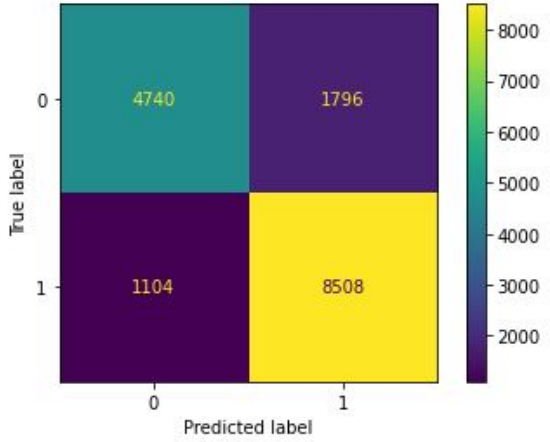


Metryki:

accuracy = 0.8204111964329948

F1 = 0.765751211631664

confusion_matrix:



Wniosek:

Uwzględniając zastosowane metryki nasuwa się wnioske iż model “Random Forest Classifier” jest najlepiej dopasowany