Klasyfikacja pomijania piosenek podczas odsłuchu na platformie Spotify na podstawie danych o okolicznościach odsłuchu

Kacper Janus
Informatyka Stosowana, Politechnika Wrocławska
272701@student.pwr.edu.pl

Maj 2024

Spis treści

1 Wprowadzenie

1.1 Istota i cel problemu

Wszystkie duże platformy streamingowe zbierają dane użytkowników w celu przewidywania ich zachowań. Spotify nie jest wyjątkiem - szwedzki serwis tworzy zaawansowane modele zachowań użytkowników na podstawie ich interakcji z systemem i wykorzystuje przewidywane cechy do rekomendacji piosenek i generowania personalizowanych playlist [1]. Skuteczność przewidywania zachowania użytkowników ma bezpośredni wpływ na zarobki serwisu poprzez lepsze dopasowanie polecanych treści.

Pominięcie słuchanej piosenki jest istotnym przykładem interakcji użytkownika z systemem, którą można wykorzystać do modelowania jego zachowania. Takie działanie może być interpretowane jako bezpośredni sygnał zwrotny (ang. *feedback*) wskazujący na niezadowolenie, brak zainteresowania lub bezpośrednią chęć zmiany w aktualnym doświadczeniu korzystania z usługi Spotify. Platformy streamingowe wykorzystują informację o pomijaniu piosenek (oraz matematyczne oszacowania prawdopodobieństwa pominięcia piosenki) jako jeden z czynników do tworzenia modelu użytkownika.

Spotify klasyfikuje pojedynczy utwór jako pominięty, gdy był on słuchany przez krócej aniżeli 30 sekund. Jeśli utwór zostanie odtworzony ponownie będzie to liczone jako kolejne odtworzenie po ponownym przesłuchaniu przez 30 sekund. Zatem jeżeli użytkownik kliknie przycisk *Natępny utwór* pod koniec utworu, to o ile słuchał go przez ponad 30 sekund, nadal będzie to liczone jako odtworzenie (bez pominięcia!)

Niniejsza praca ma na celu wyekstrahowanie z danych użytkownika tych cech, które mają największy wpływ na pomijanie piosenek. Następnie, za pomocą metod uczenia maszynowego, analizuję, czy i w jaki sposób różne cechy interakcji, takie jak moment dnia, preferencje dotyczące albumów lub wykonawców oraz wcześniejsze nawyki pomijania utworów, pozwalają przewidywać pominięcie aktualnie odtwarzanej piosenki. Jako model uczenia maszynowego wybrałem drzewa decyzyjne, ponieważ w problemach klasyfikacji, gdzie mamy do czynienia z niezrównoważonymi danymi (jak w naszym przypadku, gdzie utwory pominięte mogą mieć inny udział w zbiorze danych od utworów niepominiętych), drzewa decyzyjne oferują opcję ważenia klas, co pozwala na lepsze radzenie sobie z problemem niezrównoważonych danych. Ponadto, drzewa decyzyjne są jednymi z najbardziej interpretowalnych algorytmów uczenia maszynowego [2]. Ich struktura przypomina hierarchię warunków decyzyjnych, co pozwala na łatwe śledzenie i zrozumienie procesu podejmowania decyzji przez model. W kontekście analizy zachowań użytkowników, interpretowalność jest kluczowa, ponieważ pozwala na identyfikację czynników wpływających na decyzje o pomijaniu utworów.

1.2 Dane

W pracy wykorzystujano obszerny zbiór danych obejmujący historię strumieniowania piosenek. Dane są dostępne dla użytkowników Spotify na wniosek i obejmują szczegółowy zapis historii odtwarzania na ich kontach. Informacje dostępne w otrzymanym zbiorze są znacznie bardziej rozbudowane niż te dostępne przy wykorzystaniu publicznego API Spotify lub te zebrane przez zewnętrzne serwisy agregujące historię odtwarzania, jak np. *Last.fm*.

Dane dotyczące historii odtwarzania są wczytywane z plików *JSON*, a następnie zapisywane do formatu *CSV*. Zbiór danych obejmuje 192545 piosenek słuchanych pomiędzy 03.02.2018 i 21.03.2024.W praktyce jedynie piosenki odtwarzane od 14.10.2022 posiadały oznaczenie pominięcia, co po przetworzeniu danych zmniejszyło rozmiar zbioru do 56601 rekordów.

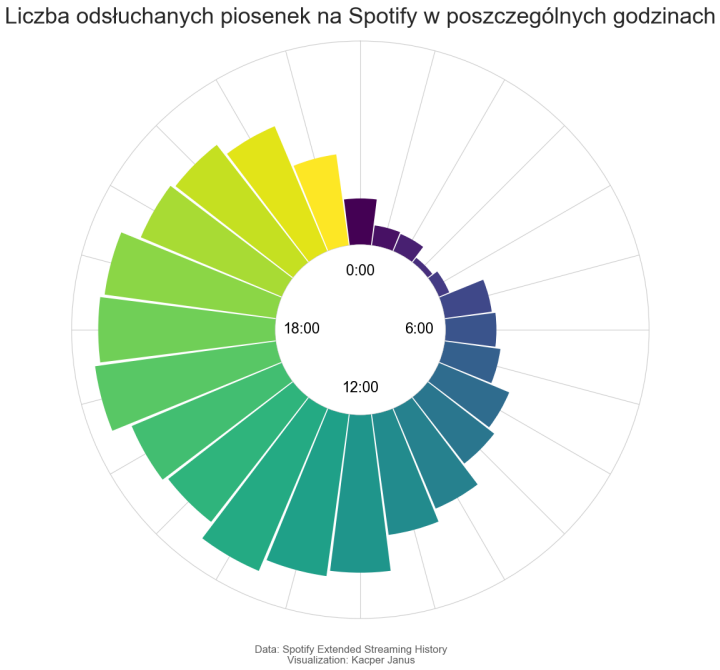
Każdy wiersz reprezentuje jedną odtworzoną piosenkę i zawiera następujące informacje potencjalnie istotne dla zadania przewidywania pomijania piosenek:

- **ts**: Pole przedstawia oznaczenie czasowe określające czas zatrzymania odtwarzania utworu według czasu UTC. Znacznik wskazuje najpierw rok, miesiąc i dzień, a następnie godzinę.
- **platform**: Platforma, na której odtwarzano utwór (np. Google Chromecast, iOS 13.5.1 (iPhone10,6)).
- **ms_played**: Liczba milisekund, przez które odtwarzano utwór.
- **master_metadata_track_name**: Nazwa utworu.
- **master_metadata_album_artist_name**: Nazwa artysty lub zespołu.
- **master_metadata_album_album_name**: Nazwa albumu, na którym znajduje się odtwarzany utwór.
- **reason_start**: Powód rozpoczęcia odtwarzania utworu (np. remote).
- **shuffle**: Czy utwór był odtwarzany w trybie losowym
- **skipped**: Czy utwór został pominięty.
- **offline**: Czy utwór był odtwarzany w trybie offline.
- **incognito_mode**: Czy utwór był odtwarzany w trybie incognito.

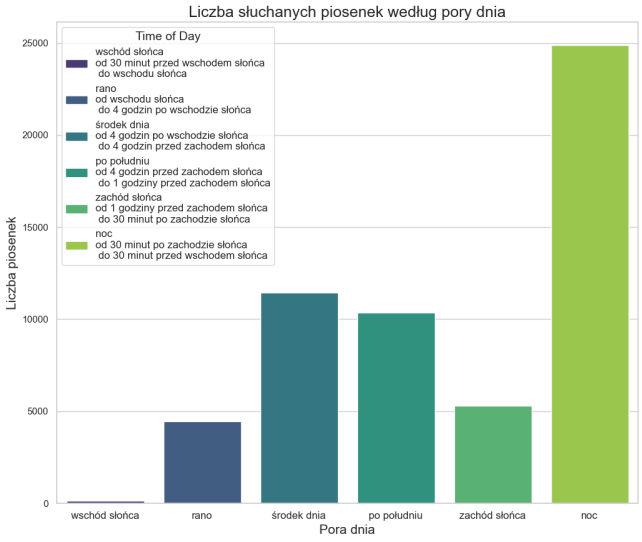
Warto zauważyć, że chociaż takie informacje jak nazwa artysty, utworu i albumu są dostępne dla modelu, to cechy odsłuchiwanej muzyki takie jak rok wydania czy gatunek już nie. Model będzie zatem klasyfikował pomijanie utworów na podstawie stosunku użytkownika wobec danego utworu, a nie inherentnych cech muzycznych utworu.

1.3 Przetwarzanie wstępne danych

Ponieważ dane nie są bezpośrednio przygotowane do trenowania na nich modelu drzewa decyzyjnego, to musiały zostać poddane wstępnej obróbce. Ze zbioru należało usunąć rekordy, dla których wartość pola *skipped* wynosiła null, ponieważ nie dostarczały wartościowych informacji do nauki przyszłego modelu. Godzina odtworzenia utworu została wyekstrahowana z pola *timestamp*, a z wykorzystaniem biblioteki Astral [3] dla odpowiedniej strefy czasowej obliczona została pora dnia, w której słuchano utworu. Wyspecjalizowanie czasowych okoliczności odsłuchu było potencjalnie istotne dla wykorzystania głębokiego uczenia wzmocnionego (*DRL*) (ang. *Deep Reinforcement Learning*) do zadania klasyfikacji pominięć w zależności od godziny i pory dnia.



Rysunek 1: Okrągły wykres kolumnowy reprezentujący zróżnicowanie liczby odsłuchów w zależności od godziny.



Rysunek 2: Histogram liczby odsłuchów według pory dnia

W celu przygotowania danych pod modele wykorzystujące *DRL* na podstawie danych w zbiorze dodano do niego kolumny zawierające liczbę odtworzeń danego utworu, liczbę odtworzeń wykonawcy utworu, wskaźnik pomijania danego artysty do tej pory (przyjmuje wartość z zakresu [0, 1]) oraz informację o tym, czy poprzednio słuchana piosenka została pominięta.

Dane w kolumnie odpowiadającej platformie zostały odpowiednio zredagowane. Rekordy takie jak:

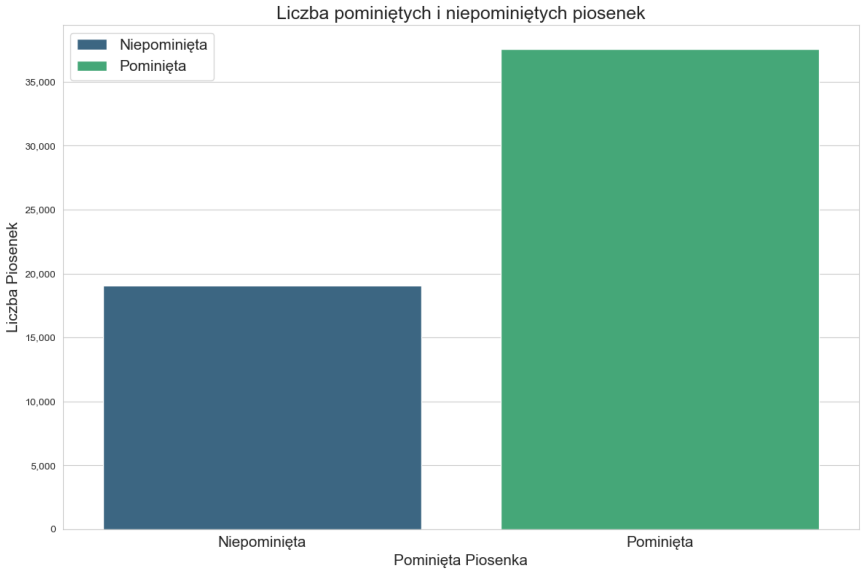
- iOS 13.1.3 (iPhone10,6)
- iOS 14.7.1 (iPhone10,6)
- iOS 14.6 (iPhone10,6)

zostały zredukowane do oznaczenia "iPhone". Założono, że wersja systemu operacyjnego nie wpływa na zachowanie użytkownika, a istotna jest jedynie informacja o tym, czy wykorzystywana jest platforma mobilna czy komputer stacjonarny.

Cechy kategoryczne wykorzystane w modelu predykcyjnym takie jak pora dnia, platforma czy nazwa albumu zostały zakodowane w formacie one-hot. Wszystkie cechy numeryczne są standaryzowane tak, aby miały rozkład z wartością średnią równą 0 i odchyleniem standardowym równym 1. Do kodowania i standaryzacji wykorzystane zostały narzędzia *StandardScaler* i *OneHotEncoder* z biblioteki sklearn.preprocessing.

1.4 Analiza eksploracyjna

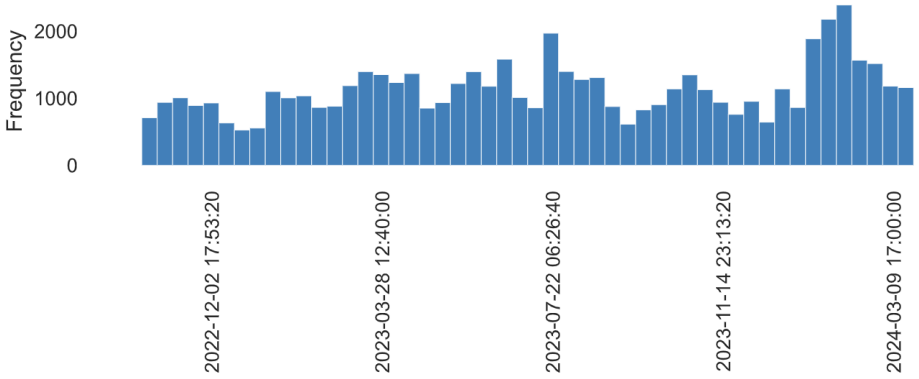
1. **Pominięte utwory (*skipped*)**



Rysunek 3: Histogram pomijania utworów

Około 66.4% utworów w zbiorze danych jest pomijana.

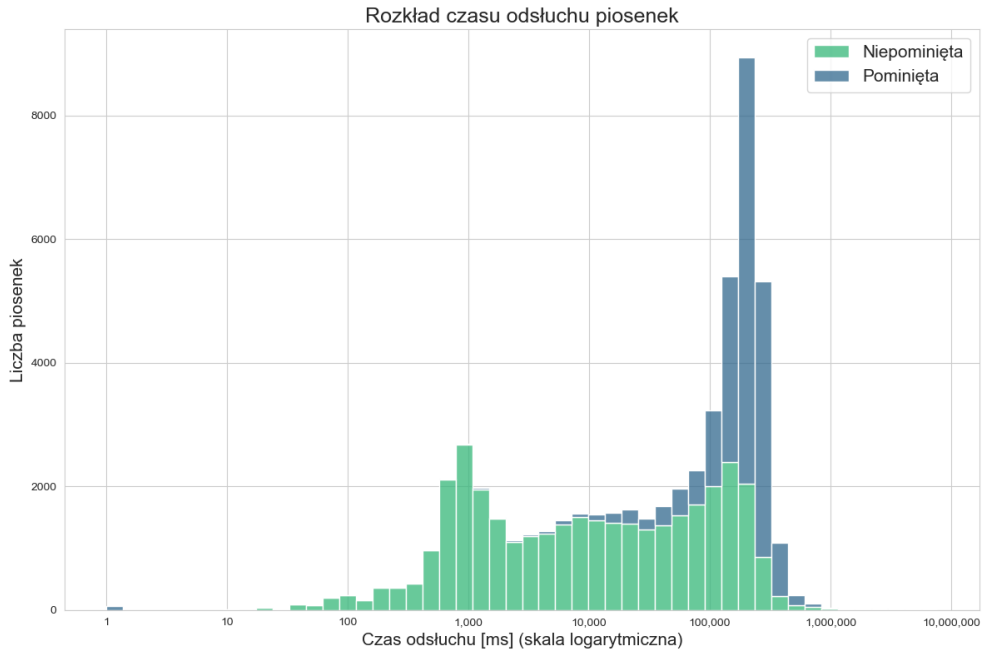
2. **Data odsłuchu (*ts*)**



Rysunek 4: Histogram daty odsłuchu

Liczba odtwarzanych utworów jest względnie stała w okresie zbierania danych.

3. **Długość odtwarzania (*ms_played*)**



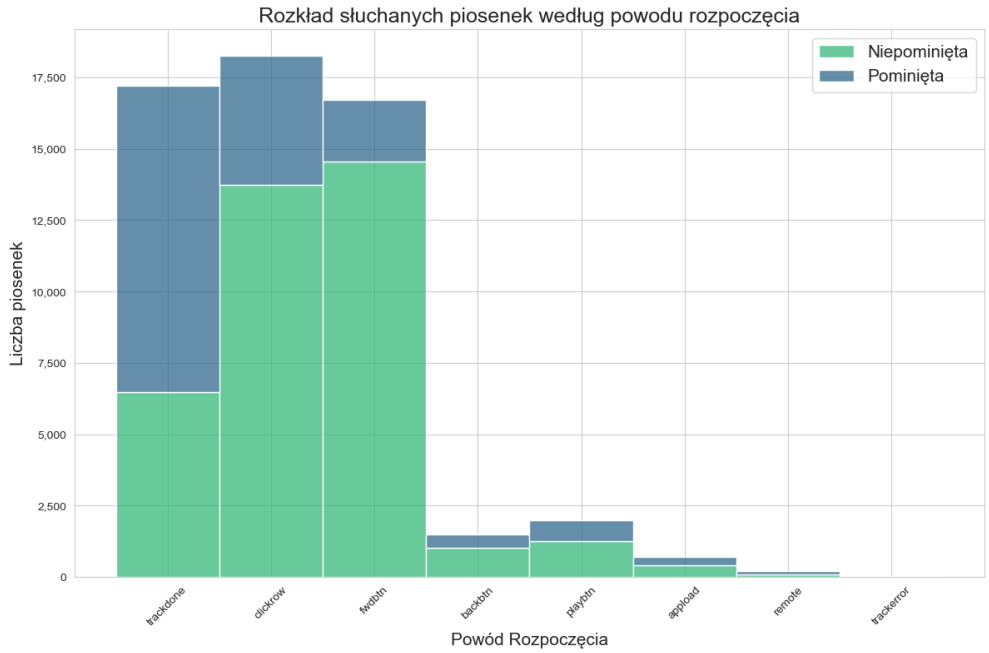
Rysunek 5: Histogram odsłuchów w zależności od długości odtwarzania utworu. Zastosowana skala logarytmiczna w osi odciętych.

Duża liczba utworów jest pomijana w trakcie pierwszych kilku sekund. Może to wskazywać na serię masowego pomijania. Gdy przesłuchana została znaczna częśćć piosenki, to bardziej prawdopodobne jest dokończenie jej.

Metryka	Wartość
Minimum	0
Maksimum	7737896
Średnia arytmetyczna	101844.32
Odchylenie standardowe	156520.79
Wartości zerowe (%)	3,7%
Unikatowe (%)	44,7%

Tabela 1: Statystyki długości odtwarzania.

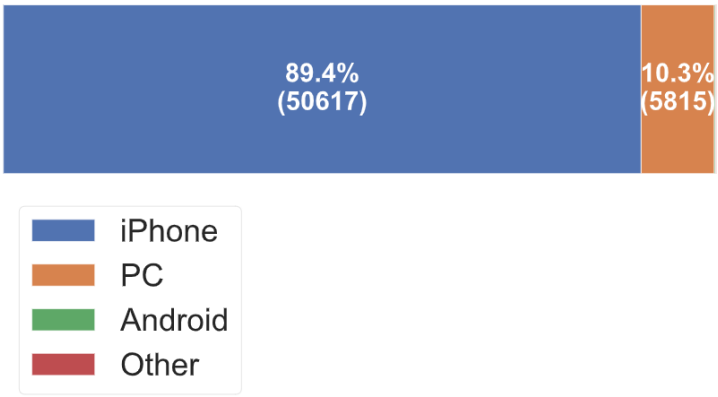
4. **Powód rozpoczęcia (*reason_ start*)**



Rysunek 6: Histogram odsłuchów w zależności od powodu rozpoczęcia odtwarzania utworu.

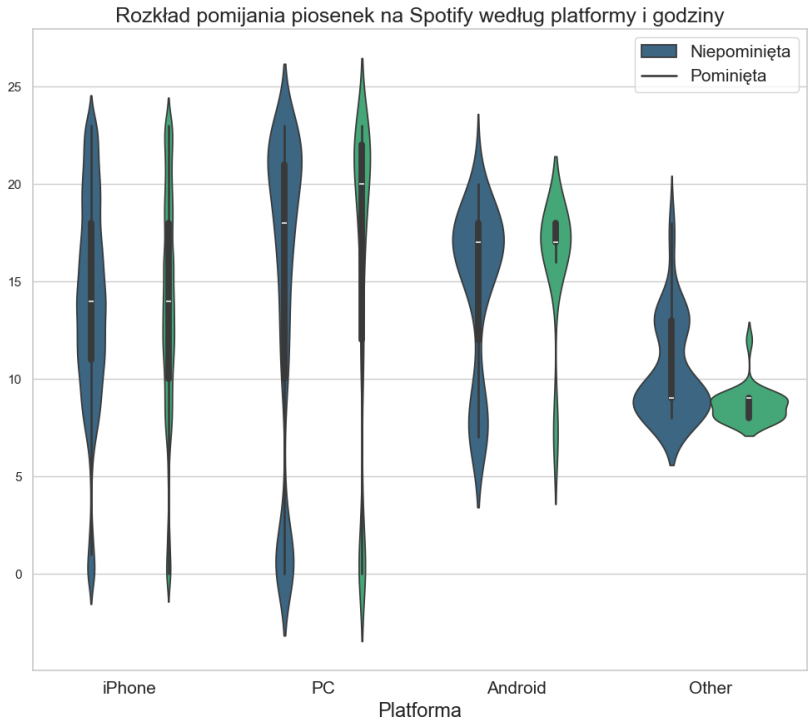
Najpopularniejszym powodem rozpoczęcia odtwarzania utworu jest wybranie go poprzez kliknięcie. W przypadku odtwarzania utworu przez wykorzystanie przycisku *"Następna piosenka"* piosenka najczęściej jest klasyfikowana jako pominięta, może to wskazywać na serię masowego pomijania

5. **Platforma (*platform*)**



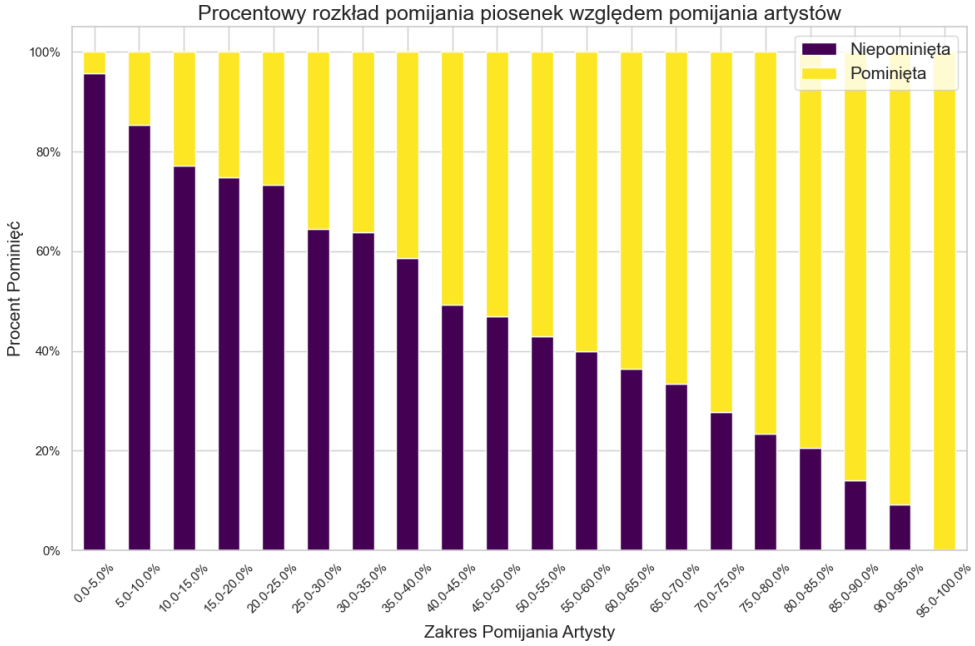
Rysunek 7: Udział odtworzeń z poszczególnych platform we zbiorze danych

Zdecydowanie najpopularniejszą platformą do odtwarzania muzyki jest urządzenie mobilne marki Apple. Ponad 5 tysięcy odtworzeń pochodzi z komputera stacjonarnego. Rekordów zawierających inną platformę jest łącznie 168.



Rysunek 8: Pominięcia utworów na poszczególnych platformach w zależności od godziny dnia

6. **Wskaźnik pomijania wykonawcy utworu (*artist_skip_rate_so_far*)**



Rysunek 9: Pominięcia utworów w zależności od wskaźnika pomijania wykonawcy

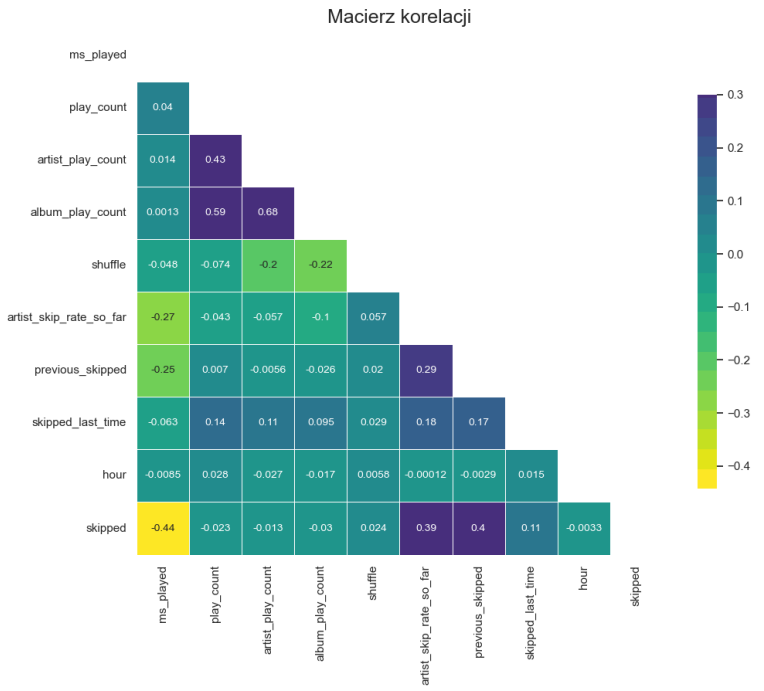
Metryka	Wartość
Minimum	0
Maksimum	1
Średnia arytmetyczna	0.65
Odchylenie standardowe	0.20
Wartości zerowe (%)	2,9%
Unikatowe (%)	42,7%

Tabela 2: Statystyki wskaźnika pomijania wykonawcy.

2 Korelacja danych

Korelacja między cechami została obliczona za pomocą współczynnika korelacji Pearsona (1) wyliczonego za pomocą fukncji biblioteki *pandas*.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$



Rysunek 10: Macierz korelacji zmiennych numerycznych

Z analizy macierzy korelacji wynika, że istnieje kilka interesujących zależności:

- **ms_played** jest silnie ujemnie skorelowane z cechą **skipped**. Oznacza to, że im dłużej utwór jest odtwarzany, tym mniejsze jest prawdopodobieństwo, że zostanie pominięty.
- **play_count** ma umiarkowanie ujemną korelację z cechą **skipped**. Wskazuje to, że utwory, które są częściej odtwarzane, są mniej prawdopodobne do pominięcia.
- **artist_skip_rate_so_far** ma dodatnią korelację z cechą **skipped**, co sugeruje, że jeśli użytkownik często pomija utwory danego artysty, istnieje większe prawdopodobieństwo, że pominie również bieżący utwór tego artysty.
- wprowadzona cecha **hour** wykazuje niską korelacje z kolumną **skipped**, co wskazuje, że zachowanie pomijania wbrew początkowym założeniom nie jest silnie skorelowane z godziną dnia.

3 Klasyfikacja

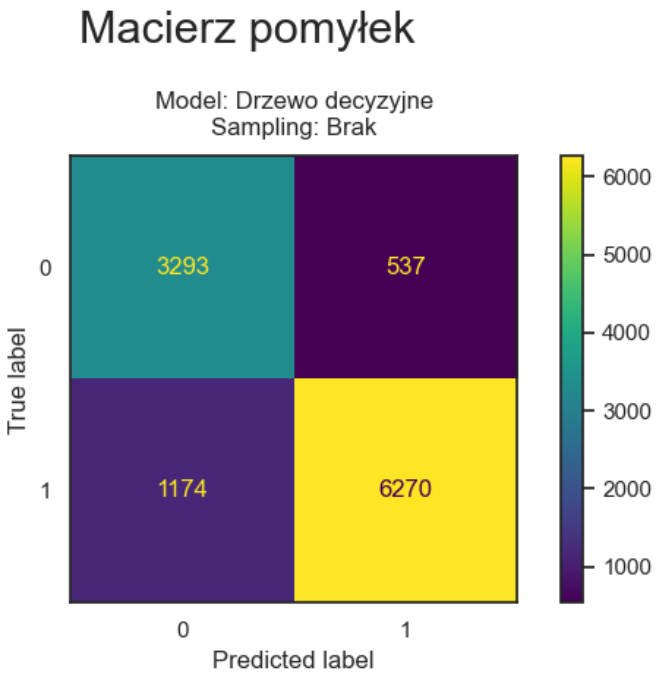
Do przewidywania pomijania utworów zastosowano drzewa decyzyjne. Dane zostały podzielone na zestaw treningowy i testowy, a następnie model został przetrenowany na zestawie treningowym. Ponieważ przewidywanie utworów niepominiętych okazało się istotnie trudniejsze dla modelu predykcyjnego, model drzewa decyzyjnego został zainicjowany z parametrem wagowym *class_weight={0: 5, 1: 1}* [4], aby nadać większe znaczenie mniej licznej klasie. W modelu wykorzystano następujące cechy:

- **Cechy kategoryczne**:
  - Oznaczenie czasowe (*ts*)
  - Pora dnia (*time_of_day*)
  - Nazwa artysty lub zespołu (*master_metadata_album_artist_name*)
  - Nazwa albumu (*master_metadata_album_album_name*)
  - Powód rozpoczęcia odtwarzania (*reason_start*)
  - Platforma (*platform*)
- **Cechy numeryczne**:
  - Liczba milisekund odtwarzania utworu (*ms_played*)
  - Liczba odtworzeń utworu (*play_count*)
  - Liczba odtworzeń artysty (*artist_play_count*)
  - Liczba odtworzeń albumu (*album_play_count*)
  - Tryb losowy (*shuffle*)
  - Wskaźnik pomijania artysty (*artist_skip_rate_so_far*)
  - Pominięcie poprzedniego utworu (*previous_skipped*)
  - Pominięcie ostatniego odtworzenia (*skipped_last_time*)

Wyniki dla każdego modelu zostały wyliczone z wykorzystaniem funkcji biblioteki *sklearn.metrics*.

3.1 Predykcja modelu dla zbioru danych

Model drzewa decyzyjnego osiągnął następujące wyniki dla pełnego zbioru danych, w którym utwory pominięte stanowią ponad 66% rekordów:



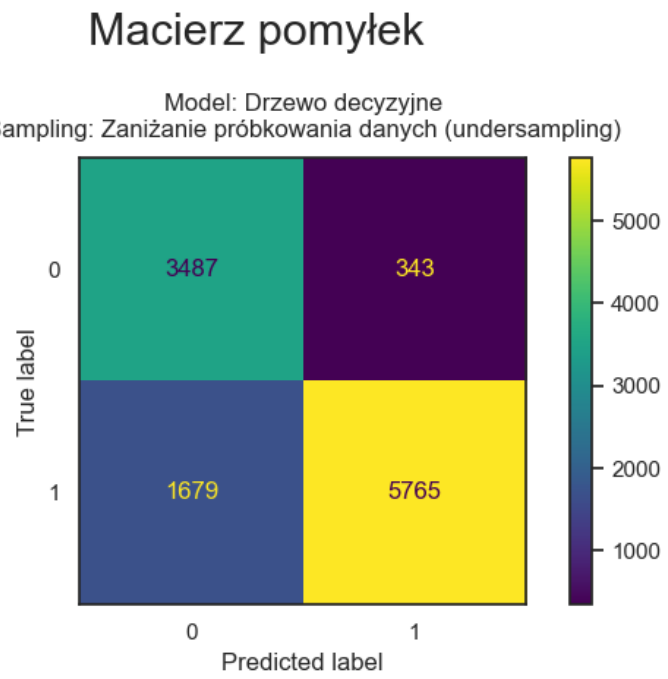
Rysunek 11: Macierz pomyłek dla zbioru bez próbkowania danych

	Precision	**Recall**	**F1-Score**
0	0.74	0.86	0.79
1	0.92	0.84	0.88
Accuracy	0.85		
Śr. arytm.	0.83	0.85	0.84
Śr. ważona	0.86	0.85	0.85

Tabela 3: Wyniki modelu drzewa decyzyjnego

3.2 Zaniżanie próbkowania danych (*Undersampling)*

Model drzewa decyzyjnego osiągnął następujące wyniki dla zbioru treningowego z zaniżeniem próbkowania, w którym utwory pominięte mają taki sam udział jak utwory niepominięte. Zaniżanie próbkowania odbyło się poprzez usunięcie losowo wybranych rekordów z pominiętymi piosenkami tak, aby zredukować ich liczność do liczności rekordów niepominiętych. Model został sprawdzony z tymi samymi parametrami (w tym *random_state=2* dla umożliwienia reprodukcji wyników obliczeń) na tych samych danych testowych.



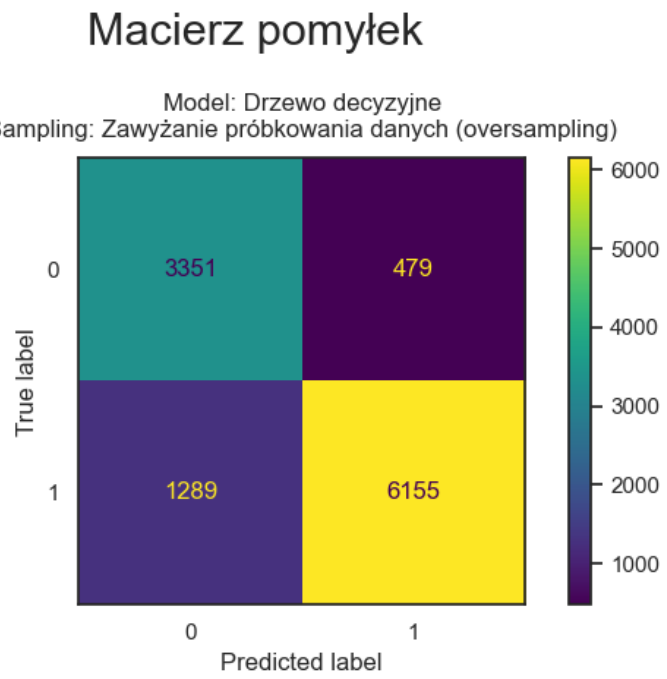
Rysunek 12: Macierz pomyłek dla zbioru z zaniżaniem próbkowania danych

	Precision	**Recall**	**F1-Score**
0	0.67	0.91	0.78
1	0.94	0.77	0.85
Accuracy	0.82		
Śr. arytm.	0.81	0.84	0.81
Śr. ważona	0.85	0.82	0.83

Tabela 4: Wyniki modelu drzewa decyzyjnego z zaniżaniem próbkowania danych

3.3 Zawyżanie próbkowania danych (*Oversampling*)

Model drzewa decyzyjnego osiągnął następujące wyniki dla zbioru treningowego z zawyżaniem próbkowania, w którym utwory pominięte mają taki sam udział jak utwory niepominięte. Zawyżanie próbkowania odbyło się poprzez zduplikowanie losowo wybranych rekordów z niepominiętymi piosenkami tak, aby zwiększyć ich liczność do liczności rekordów niepominiętych. Model został sprawdzony z tymi samymi parametrami (w tym *random_state=2* dla umożliwienia reprodukcji wyników obliczeń) na tych samych danych testowych.



Rysunek 13: Macierz pomyłek dla zbioru z zawyżaniem próbkowania danych

	Precision	**Recall**	**F1-Score**
0	0.72	0.87	0.79
1	0.93	0.83	0.87
Accuracy	0.84		
Śr. arytm.	0.82	0.85	0.83
Śr. ważona	0.86	0.84	0.85

Tabela 5: Wyniki modelu drzewa decyzyjnego z zawyżaniem próbkowania danych

4 Wnioski

Na podstawie przeprowadzonych analiz i wyników modeli można wyciągnąć kilka kluczowych wniosków dotyczących skuteczności modelu, wpływu poszczególnych cech na decyzje modelu oraz wpływu metod próbkowania danych na uzyskane rezultaty.

4.1 Skuteczność Modelu

Model drzewa decyzyjnego osiągnął wysoką precyzję (92%) w przewidywaniu utworów pominiętych (klasa 1) oraz dobrą czułość (86%) w wykrywaniu utworów niepominiętych (klasa 0). Oznacza to, że model jest efektywny w identyfikacji piosenek, które użytkownicy pomijają, co jest kluczowe dla zrozumienia zachowań słuchaczy i poprawy funkcjonalności rekomendacji na platformie Spotify. Platforma, mając dostęp do bardziej złożonych zbiorów danych, mogłaby skonstruować jeszcze bardziej precyzyjny model, który brałby pod uwagę gatunek muzyczny, rok wydania utworu, czy dane demograficzne użytkowników i rozpoznawał masowe serię pominiętych utworów, na przykład za pomocą algorytmu centroidów (*K-means clustering*).

4.2 Wpływ cech na decyzje modelu

Analiza wpływu poszczególnych cech na decyzje modelu drzewa decyzyjnego wykazała, że najistotniejsze były cechy związane z historią odsłuchów i preferencjami użytkownika. Cechy takie jak liczba milisekund odtwarzania utworu (*ms_played*), liczba odtworzeń utworu (*play_count*), oraz wskaźnik pomijania artysty (*artist_skip_rate_so_far*) miały największy wpływ na decyzje modelu. Utwory, które wcześniej były często odtwarzane były pomijane z mniejszym prawdopodobieństwem, co jest zgodne z intuicją, że użytkownicy rzadziej pomijają utwory, które im się podobają. Wskaźnik pomijania artysty pokazał, że jeżeli użytkownik często pomija utwory danego artysty, to istnieje wysokie prawdopodobieństwo, że pominie także bieżący utwór tego artysty.

4.3 Wpływ próbkowania danych na wyniki

Dla zbioru danych z zaniżonym próbkowaniem (*undersampling*), model osiągnął niższą precyzję (67%) w przewidywaniu utworów niepominiętych, ale wyższą czułość (91%). Oznacza to, że model jest bardziej efektywny w identyfikacji utworów niepominiętych kosztem większej liczby fałszywych alarmów. Zastosowanie undersamplingu pokazuje typowy problem *precision-recall tradeoff*, gdzie zwiększenie czułości (*recall*) odbywa się kosztem spadku precyzji (*precision*) [5].

Dla zbioru danych z zawyżonym próbkowaniem (*oversampling*), model osiągnął precyzję (72%) i czułość (87%) w przewidywaniu utworów niepominiętych. Zastosowanie oversamplingu pozwala na lepsze zrównoważenie precyzji i czułości, co jest istotne dla utrzymania równowagi między identyfikacją prawdziwych pozytywów i minimalizacją fałszywych negatywów

Model osiągnął najwyższą dokładność (*accuracy*) (85%) dla pełnego zbioru danych, co sugeruje, że podczas zmiany próbkowania danych może zachodzić utrata istotnych informacji lub wprowadzenie zniekształceń, które negatywnie wpływają na wydajność modelu.

Literatura

[1] Spotify. Understanding spotify recommendations. `https://www.spotify.com/us/safetyandprivacy/understanding-recommendations`, 2024. Dostęp: 2024-05-18.

[2] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On explaining decision trees, 2020.

[3] Astral Developers. Astral documentation. `https://astral.readthedocs.io/en/latest/`, 2024. Dostęp: 2024-05-14.

[4] Scikit learn developers. Decisiontreeclassifier. `https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html`, 2024. Dostęp: 2024-05-17.

[5] Michael Gordon and Manfred Kochen. Recall-precision trade-off: A derivation. *Journal of the American Society for Information Science*, 40(3):145–151, 1989.