Studium Licencjackie

Kierunek: Metody ilościowe w ekonomii i systemy informacyjne

Imię i nazwisko autora: Jakub Kołpa
Nr albumu: 109434

Identyfikacja czynników wpływających na popularność utworów na platformie streamingowej Spotify w Unii Europejskiej

Praca licencjacka
pod kierunkiem naukowym
dr hab. Małgorzaty Knauff
Katedra Ekonomii Ilościowej

Warszawa 2023

Spis treści

Wstęp

Spotify to cyfrowa platforma do odtwarzania muzyki, podcastów i filmów, która zapewnia natychmiastowy dostęp do milionów utworów i innych treści twórców z całego świata. Jest to jeden z największych serwisów streamingowych na świecie, z ponad 500 milionami aktywnych użytkowników w ponad 180 krajach[1]. Serwis ten oferuje swoim użytkownikom dostęp do ogromnej biblioteki utworów muzycznych, która stale się powiększa. W 2022 roku globalne przychody z rynku muzycznego odnotowały wzrost o 9,0%. Przychody cyfrowe na tym rynku stale rosły w ciągu ostatnich kilku lat, a streaming nadal był jego dominującą częścią stanowiąc 67,0% globalnych przychodów (po wzroście o 11,5% w 2022 roku). Do końca 2022 roku na świecie było 589 mln użytkowników posiadających płatne konta na serwisach streamingowych[2]. Z tego powodu badanie popularności utworów na Spotify jest nie tylko kwestią artystyczną, ale także ma znaczenie ekonomiczne.

Identyfikacja czynników mających wpływ na sukces odnoszony przez utwory muzyczne jest ważnym aspektem z punktu widzenia ekonomii, ponieważ popularność przekłada się na zyski dla artystów i wytwórni muzycznych. Im popularniejszy jest dany utwór, tym więcej razy będzie odtwarzany, co przekłada się na wyższe dochody twórców. Spotify jest jednym z największych serwisów streamingowych, więc sukces na tej platformie prawdopodobnie zagwarantuje znaczne korzyści finansowe[3]. Takie badanie może pomóc w zrozumieniu, co przyciąga słuchaczy oraz ułatwić wybór działań mających na celu zwiększenie popularności utworów.

Kolejnym argumentem potwierdzającym ekonomiczną istotność analizowanego problemu jest możliwy wpływ zwiększenia rozpoznawalności na zainteresowanie sponsorów, co może zapewnić twórcom dodatkowe źródło dochodów. Coraz więcej firm chce wykorzystywać muzykę jako narzędzie marketingowe lub po prostu jako korzystną inwestycję. Analiza czynników wpływających na sukces danego utworu czy artysty może pozwolić więc na podejmowanie skuteczniejszych decyzji marketingowych czy wybór najlepszego terminu wypuszczenia nowego albumu[4].

Wpływ na rynek pracy dla muzyków i osób związanych z przemysłem muzycznym jest kolejnym istotnym aspektem. Im popularniejszy jest utwór danego twórcy, tym większe szanse

[1] Spotify, *About Spotify*, https://newsroom.spotify.com/company-info/ (dostęp 23.04.2023)
[2] International Federation of the Phonographic Industry, *Industry data*, https://www.ifpi.org/our-industry/industry-data/ (dostęp 23.04.2023)
[3] C. Araujo, M. Cristo, R. Giusti, *Predicting Music Popularity on Streaming Platforms*, Federal University of Amazonas, Manaus 2019, s. 142
[4] ibidem, s. 142

na zwiększenie popytu na koncerty i inne wydarzenia z jego udziałem. To z kolei może przyciągać większą liczbę fanów i zwiększyć zainteresowanie organizatorów koncertów i festiwali muzycznych. Tak więc zrozumienie determinant sukcesu osiąganego przez daną piosenkę może wpłynąć nie tylko na samych artystów i wytwórnie, ale także cały przemysł muzyczny. Taka analiza umożliwia lepsze zrozumienie rynku i podejmowanie bardziej trafnych decyzji biznesowych. Dzięki temu można zoptymalizować procesy produkcyjne, co z kolei może przyczynić się do ogólnego rozwoju wszystkich podmiotów z tym przemysłem związanych.

Celem niniejszej pracy jest identyfikacja czynników wpływających na popularność utworów na platformie streamingowej Spotify. Analiza obejmuje rynki krajów należących do Unii Europejskiej, a badane utwory miały swoją premierę w latach 2000-2022. Dane zostały zaczerpnięte z platformy Spotify for Developers, a konkretnie Spotify Web API. Zawierają one głównie zmienne opisujące cechy dźwiękowe utworów, ale pojawiają się również determinanty takie jak rok wydania czy liczba obserwatorów artysty. Do analizy skorzystano z metod uczenia maszynowego: regresji logistycznej, regresji logistycznej z wykorzystaniem algorytmu *stepwise*, drzewa klasyfikacyjnego oraz lasu losowego. Zmienne objaśniające podzielono na dwa zbiory w celu dokładniejszej analizy problemu. Wszystkie analizy zostały wykonane z użyciem języka programowania R oraz darmowego oprogramowania RStudio.

W rozdziale pierwszym opisano literaturę podejmującą temat identyfikacji determinant oraz prognozowania popularności utworów muzycznych. Przedstawiono różne podejścia autorów do tego tematu, a także porównano rezultaty ich analiz. Wspomniane prace prezentują odmienne sposoby mierzenia popularności, ale również badają inne czynniki wpływające na rozpoznawalność. W większości artykułów analizowano wpływ cech dźwiękowych w celu porównania wyników niniejszej pracy z powstałymi wcześniej badaniami o podobnej tematyce. Pojawiają się jednak również publikacje badające oddziaływanie innych zmiennych m.in. wpływu społecznego czy wzmianek na platformach społecznościowych, aby pokazać złożoność tematu oraz mnogość możliwych determinant osiągnięcia sukcesu przez utwór muzyczny. Przedstawiono również proces pozyskania danych oraz szczegółowy opis zmiennych wraz z transformacją zmiennej zależnej. Opisano również działanie zastosowanych metod uczenia maszynowego.

W rozdziale drugim znajdują się wyniki analizy. W pierwszym podrozdziale opisano proces eksploracji danych. Zaprezentowano strukturę badanych zmiennych, wprowadzone do bazy danych zmiany oraz pokazano zależności występujące w analizowanym zbiorze

danych. W drugim podrozdziale przedstawiono wyniki estymacji modeli oraz przeprowadzono ich ocenę.

Zakończenie skupia się na podsumowaniu wyników analizy. Obejmuje ono dyskusję na temat istotności zmiennych, ale również wybór najlepiej klasyfikujących modeli.

Rozdział I

Literatura i metodologia

1.1 Przegląd literatury

Identyfikowanie czynników wpływających na popularność utworów muzycznych oraz jej przewidywanie przyciąga uwagę wielu badaczy, a także jest tematem wielu badań naukowych i artykułów. Istnieje wiele interpretacji wyżej wymienionego problemu. Przyczyniają się do tego różne sposoby mierzenia popularności, ale również wyboru badanych determinant. W swoim artykule Myra Interiano wspólnie z innymi autorami porusza temat predykcji osiągnięcia sukcesu przez utwór muzyczny oraz trendów zaobserwowanych na przestrzeni lat w tym zakresie[5]. Analizowali oni ponad 500 000 utworów z lat 1985-2015 definiując popularność utworu jako pojawienie się w rankingach największych hitów. Sprawdzili zależność osiągnięcia tego sukcesu od cech dźwiękowych i akustycznych tych kompozycji oraz ocenili możliwości predykcyjne zbudowanych modeli. Badanie wykazało, że popularne utwory różnią się od przeciętnych – są weselsze, bardziej imprezowe i taneczne. Znaczenie miał również rok wydania – im bardziej zbliżano się do roku badanego tym skuteczniejsze okazywały się modele, co pokazuje kształtujący się trend. Zastosowano również dwa podejścia do budowy modelu: pierwszy opierał się tylko na cechach dźwiękowych, a do drugiego dodano jeszcze zmienną wskazującą na to czy dany artysta jest wyjątkowo popularny (pojawił się na pierwszym miejscu listy przebojów w niedalekiej przeszłości). Zmienna dodana w drugim podejściu okazała się być wyraźnie istotna, znacząco poprawiając ocenę jakości tego modelu. Z tego powodu w niniejszej pracy postanowiono również zastosować dwa zbiory zmiennych objaśniających – pierwszy z nich zawierający zmienną określającą liczbę obserwatorów danego twórcy, aby uwzględnić opisany w artykule Myry Interiano et al. wpływ dużej popularności wykonawcy.

Rutger Nijkamp w swojej pracy również bada problem odnoszonego sukcesu przez utwory muzyczne i czynników na to wpływających[6]. Stawia on hipotezy dotyczące relacji popularności piosenek z predyktorami dotyczącymi cech dźwiękowych utworów, aby następnie je badać. Według autora pomiędzy zmienną zależną a zmiennymi: *acousticness, duration_ms, liveness* oraz *speechiness* powinna występować zależność ujemna, a pomiędzy

[5] M. Interiano, K. Kazemi, L. Wang, J. Yang, Z. Yu, N. Komarova, *Musical trends and predictability of success in contemporary songs in and out of the top charts*, The Royal Society Publishing, Irvine 2018, s. 1-16

[6] R. Nijkamp, *Prediction of product success: explaining song popularity by audio features from Spotify data*, 11th IBA Bachelor Thesis Conference, Enschede 2018, s. 1-9

zmienną objaśnianą a zmiennymi: *danceability*, *energy*, *loudness*, *mode* oraz *valence* zależność dodatnia. Oprócz tego przewiduje on, że zmienna *tempo* nie będzie istotna. W niniejszej pracy sprawdzono czy rezultaty wykonanej analizy będą zgodne z tymi przypuszczeniami.

Agha Haider Raza oraz Krishnadas Nanath w swoim artykule starają się znaleźć wzór na prognozowanie czy dany utwór będzie popularny przed wypuszczeniem go na rynek używając przy tym metod uczenia maszynowego[7]. Brali oni pod uwagę nie tylko cechy dźwiękowe, ale również przeprowadzili analizę sentymentu tekstów. Zastosowane przez nich do predykcji modele to regresja logistyczna, drzewo decyzyjne, las losowy oraz naiwny klasyfikator Bayesa. Najbardziej znaczącymi zmiennymi w powyższym artykule okazały się być taneczność utworu oraz jego tempo. Istotność zmiennej *tempo* jest niezgodna z przypuszczeniami Rutgera Nijkampa opisanymi w poprzednim akapicie. Użyte przez autorów modele osiągnęły dokładność niewiele wyższą od 50%, co wskazuje na ich słabą efektywność – jest ona zbliżona do rzutu monetą. Badacze dochodzą do konkluzji, że przy aktualnie dostępnych danych i metodach predykcji nie jest możliwe przewidywanie sukcesu utworów przed ich wypuszczeniem. Zwracają uwagę na to, że warto dodać inne zmienne takie jak gatunek utworu lub informacje dostępne po udostępnieniu kompozycji do szerszego grona odbiorców m.in. wykorzystane strategie marketingowe czy platformy, na których dany utwór został umieszczony.

W pracy pod tytułem „Analiza danych wpływających na popularność produktów na międzynarodowym rynku muzycznym" Anna Duda oraz Izabela Jonek-Kowalska starały się również opracować pewien wzór, sposób na stworzenie hitu muzycznego[8]. W badaniu tym testowanymi możliwymi determinantami były cechy dźwiękowe utworu, tak jak w większości poruszanych w niniejszej pracy artykułów. Z przeprowadzonej analizy trendów wynikało, że najistotniejszymi atrybutami piosenek zwiększającymi ich popularność są tempo, energiczność i taneczność. Jest to zgodne z przypuszczeniami autorów podobnych prac ponownie z wyjątkiem zmiennej *tempo*, która według artykułu Rutgera Nijkampa powinna być nieistotna. Autorki zastosowały dwa podejścia do wyrażania popularności. W pierwszym z nich wartość zmiennej objaśnianej *popularity* to wartość pobrana z Spotify Web API obliczona przez Spotify. Zbudowany model regresji wielorakiej wskazał istotność zmiennej *speechiness*. Ważne jest jednak, że model ten wyjaśnia tylko 6,25% zmiennej zależnej. W drugim podejściu

[7] A. H. Raza, K. Nanath, *Predicting a Hit Song with Machine Learning: Is there an apriori secret formula?*, „2020 International Conference on Data Science, Artificial Intelligence and Business Analytics (DATABIA)", Medan 2020, s. 111-116

[8] A. Duda, I. Jonek-Kowalska, *Analiza danych wpływających na popularność produktów na międzynarodowym rynku muzycznym*, "Management and Quality – Zarządzanie i jakość" 2022, vol. 4, no. 3, s. 31-48

popularność wyrażona była w liczbie tygodni, podczas których utwór znajdował się na liście Billboard The Hot 100. W tym modelu istotne okazały się być głośność oraz obecność publiczności w nagraniu.

Wiele pozycji naukowych porusza temat wpływu na popularność utworów czynników innych niż cechy dźwiękowe, często trudniejszych do zmierzenia. W artykule o tytule „Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market" autorzy poruszają problematykę wpływu konformizmu na popularność utworów[9]. Przeprowadzono badanie empiryczne na grupie 14 341 uczestników. Każdy z nich był losowo przydzielany do jednej z dwóch grup – niezależnej lub zależnej społecznie. Grupy różniły się jedynie dostępnością informacji o wyborach poprzednich badanych. Uczestnicy pierwszej z nich mieli za zadanie wybrać czego chcieliby posłuchać jedynie na podstawie nazw zespołów, artystów oraz ich piosenek. Następnie po ich przesłuchaniu dawali oni ocenę w skali 1-5 oraz mieli możliwość pobrania słuchanego utworu. Członkowie drugiej grupy mogli zobaczyć ile razy dana kompozycja została pobrana przez poprzednie osoby w jednym z ośmiu „światów", do których zostali na początku przydzieleni. Rezultatem tego eksperymentu było wykrycie dużego wpływu konformizmu na popularność utworów. W niniejszej pracy wpływ ten jest uwzględniony w postaci zmiennej *followers* – liczba obserwatorów może silnie wpłynąć na decyzje o słuchaniu danego utworu m.in. poprzez chęć podążania za większą grupą.

Inną determinantą pojawiającą się w artykułach naukowych poruszających tematykę prognozowania popularności utworów muzycznych są posty na Twitterze[10]. Autor uzasadnia wybór takiego czynnika przeszłymi sukcesami w predykcji różnych zjawisk w innych obszarach m.in. przewidywanie wyników wyborów lub cen akcji na giełdzie. Praca ta analizuje dane uzyskane z Twittera powiązane z utworami i artystami, którzy znaleźli się wśród dziesięciu najpopularniejszych na listach Billboard Hot 100. Przeprowadzono analizę sentymentu tekstów oraz liczby wzmianek w celu przewidywania tych list w przyszłości. Zbadano ponad milion tweetów, wrzuconych pomiędzy październikiem i listopadem 2018 roku. Analiza wykazała umiarkowaną korelację między liczbą wzmianek a jej wynikami na listach przebojów, ale nie było znaczącej zależności między uwagą poświęcaną danemu artyście a sukcesem utworu. Badanie pozwala stwierdzić, że analiza postów na portalach społecznościowych ma istotny wpływ na popularność piosenek.

[9] M. J. Salganik, P. S. Dodds,, D. J. Watts, *Experimental study of inequality and unpredictability in an artificial cultural market*, "Science" 2006, vol. 311, s. 854-856

[10] E. Tsiara, C. Tjortjis, *Using Twitter to Predict Chart Position for Songs,* w: "Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology", vol. 583, I. Maglogiannis, L. Iliadis, E. Pimenidis (eds.), Springer, Neos Marmaras 2020, s. 62-72

Na podstawie przedstawionej literatury możliwe jest postawienie kilku hipotez, których prawdziwość zostanie następnie sprawdzona podczas analizy:

Hipoteza 1: Zmienna określająca liczbę osób obserwujących danego twórcę będzie istotna, a jej wpływ na zmienną objaśnianą będzie pozytywny.

Hipoteza 2: Taneczność, tempo oraz energiczność utworu muzycznego będą istotnymi czynnikami wpływającymi na wzrost popularności.

Hipoteza 3: Zmienna *speechiness* będzie negatywnie skorelowana ze zmienną zależną.

1.2 Pozyskanie danych

Dane użyte do badania czynników wpływających na popularność utworów zostały zaczerpnięte z platformy Spotify for Developers, a konkretnie Spotify Web API[11]. Strona ta umożliwia tworzenie aplikacji, które mogą wchodzić w interakcje z serwisem streamingowym Spotify. W opisywanym badaniu użyto funkcji tej platformy w celu uzyskania informacji o utworach i ich charakterystykach, artystach je wykonujących oraz albumach, na których się znajdują, aby stworzyć bazę danych zawierającą potrzebne zmienne do przeprowadzenia analizy. Do identyfikacji czynników wpływających na rozpatrywany problem użyto danych z rynków Unii Europejskiej. Badane utwory powstały w latach od 2000 do 2022 roku. Finalna baza danych zawarta w pliku *Song_popularity_dataset.xlsx* została utworzona przy użyciu biblioteki *spotifyr* w dniu 28 marca 2023 roku w następujących krokach:

1. Korzystając z modułu *Search for Item*[12] wyciągnięto numery identyfikacyjne utworów muzycznych z rynków w krajach należących do Unii Europejskiej z lat 2000-2022. Udało się uzyskać 5476 unikalnych identyfikatorów. Z powodów operacyjnych ograniczono bazę do 2000 obserwacji.
2. Korzystając z modułu *Get Saveral Tracks*[13], *Get Saveral Artists*[14] oraz *Get Saveral Albums*[15] pobrano podstawowe informacje o utworach, artystach je wykonujących i albumach, na których się znajdują na podstawie zaciągniętych w poprzednim kroku numerów identyfikacyjnych.

[11] Spotify for Developers, *Web API Documentation*, https://developer.spotify.com/documentation/web-api (dostęp 23.04.2023)
[12] Spotify for Developers, *Web API Documentation*, https://developer.spotify.com/documentation/web-api/reference/search (dostęp 23.04.2023)
[13] Spotify for Developers, *Web API Documentation*, https://developer.spotify.com/documentation/web-api/reference/get-several-tracks (dostęp 23.04.2023)
[14] Spotify for Developers, *Web API Documentation*, https://developer.spotify.com/documentation/web-api/reference/get-multiple-artists (dostęp 23.04.2023)
[15] Spotify for Developers, *Web API Documentation*, https://developer.spotify.com/documentation/web-api/reference/get-multiple-albums (dostęp 23.04.2023)

3. Korzystając z modułu *Get Tracks' Audio Features*[16] uzyskano dane na temat cech dźwiękowych badanych utworów, takich jak tempo czy tonacja. Z powodu braku informacji o takich charakterystykach dla dwóch utworów, usunięto je z wszystkich plików przed ostatnim krokiem.

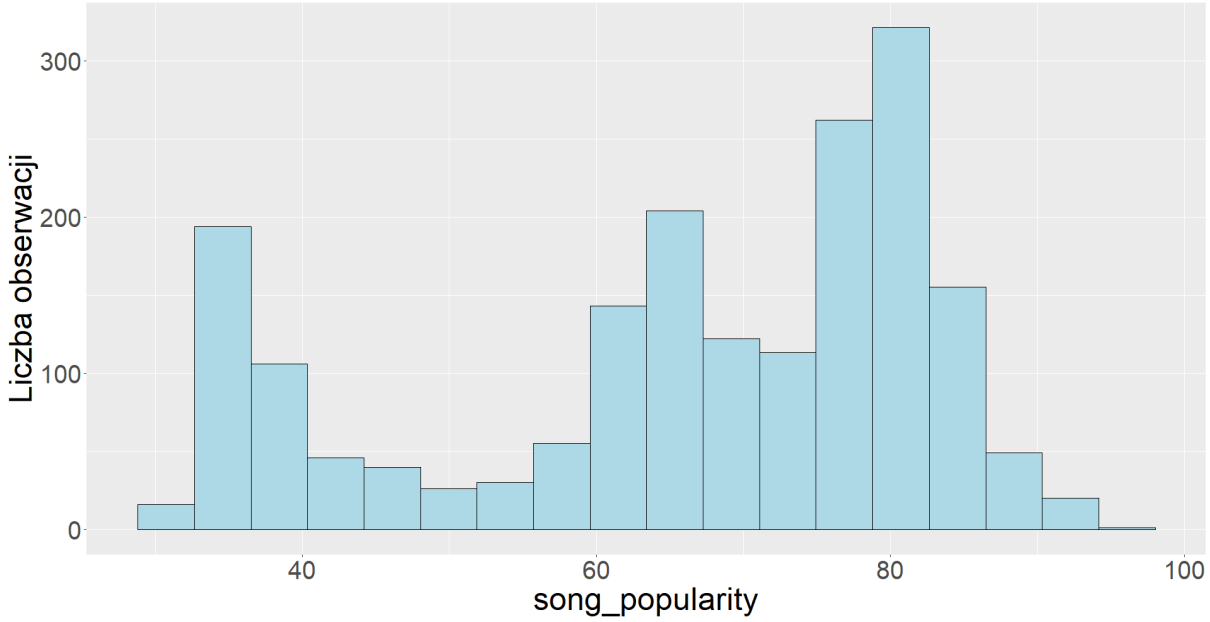
4. Połączenie plików utworzonych w kroku drugim i trzecim oraz usunięcie zduplikowanych kolumn w celu utworzenia finalnej bazy zawierającej 1998 obserwacji i zapisania jej do pliku używanego do dalszej analizy.

1.3 Opis zmiennych

Zmienną objaśnianą w użytych modelach jest zmienna *song_popularity* zaczerpnięta z wymienionego powyżej modułu *Get Saveral Tracks*, gdzie przyjmuje nazwę *popularity*. Jest to wskaźnik popularności utworu obliczany przez algorytm Spotify. Przyjmuje on wartości od 0 do 100, gdzie 100 oznacza największą popularność. Opiera się w dużej mierze na całkowitej liczbie odtworzeń utworu oraz na tym, jak niedawno miały one miejsce. Oznacza to, że utwory, które są odtwarzane często obecnie będą miały większą popularność od tych odtwarzanych wiele razy w przeszłości. Celem badania jest identyfikacja czynników determinujących czy dany utwór muzyczny jest rozpoznawalny. Z tego powodu zmienna *song_popularity* została przetransformowana na zmienną binarną, gdzie 1 oznacza, że kompozycja jest popularna (popularność większa od 70). Próg podziału został wybrany na podstawie histogramu zmiennej *song_popularity* przedstawionego na rysunku 1 oraz jej statystyk opisowych (mediana = 70).

[16] Spotify for Developers, *Web API Documentation*, https://developer.spotify.com/documentation/web-api/reference/get-several-audio-features (dostęp 23.04.2023)

Rysunek 1. Histogram zmiennej objaśnianej *song_popularity*



Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Zmienne objaśniające uwzględniane w modelowaniu wraz z opisem zawarte są w tabeli 1:

Tabela 1. Zmienne objaśniające wraz z opisem

Nazwa zmiennej	Opis zmiennej
Artist_popularity	Popularność artysty – wartość w przedziale od 0 do 100, która jest obliczana na podstawie popularności wszystkich utworów danego artysty.
Album_popularity	Popularność albumu – wartość w przedziale od 0 do 100.
Followers	Całkowita liczba osób obserwujących danego artystę.
Acousticness	Miara pewności w zakresie od 0 do 1 czy utwór jest akustyczny. 1 oznacza najwyższą pewność, że utwór jest akustyczny.
Danceability	Taneczność – opisuje jak bardzo utwór nadaje się do tańca w oparciu o kombinację elementów muzycznych, takich jak tempo, stabilność rytmu, siłę oraz rozłożenie akcentów w takcie, a także ogólną regularność utworu. Wartość 0 jest najmniej taneczna, a 1 najbardziej.
Duration_ms	Czas trwania ścieżki w milisekundach.

Energy	Energia – jest miarą od 0 do 1 i stanowi percepcyjną miarę intensywności i aktywności. Zazwyczaj utwory energetyczne są szybkie i głośne. Dla przykładu gatunki takie jak death metal mają wysoką energię, podczas gdy preludium Bacha uzyskuje niski wynik w tej kategorii. Cechy utworu wpływające na tę zmienną to zakres dynamiczny, postrzegana głośność, barwa i ogólna entropia.
Instrumentalness	Instrumentalność – przewiduje, czy utwór nie zawiera wokalu. Dźwięki takie jak *Ooh* i *aah* są w tym kontekście traktowane jako instrumentalne. Rap lub utwory mówione są wyraźnie *wokalne*. Im wartość instrumentalności jest bliższa 1, tym większe prawdopodobieństwo, że utwór nie zawiera treści wokalnych. Wartości powyżej 0,5 są przeznaczone do reprezentowania utworów instrumentalnych, ale pewność jest większa, gdy wartość zbliża się do 1.
Key	Tonacja, w której napisany jest utwór. Liczby całkowite odpowiadają tonacjom przy użyciu standardowej notacji Pitch Class[17]. Jeżeli nie wykryto żadnej tonacji, wartość zmiennej wynosi -1.
Liveness	Wykrywa obecność publiczności w nagraniu. Wyższe wartości liveness reprezentują zwiększone prawdopodobieństwo, że utwór został wykonany na żywo. Wartość powyżej 0,8 daje wysokie prawdopodobieństwo, że utwór jest wykonywany na żywo.
Loudness	Ogólna głośność utworu w decybelach (dB). Wartości głośności są uśredniane dla całego utworu i przydają się do porównywania względnej głośności utworów. Wartości zwykle mieszczą się w zakresie od -60 do 0 dB.
Mode	Modalność – wskazuje czy utwór jest w tonacji durowej czy mollowej. Skale durowe są reprezentowane przez 1, a mollowe przez 0.
Speechiness	Zmienna wykrywa obecność słów mówionych w utworze. Im bardziej nagranie przypomina wyłącznie mowę (m.in. talk show, książka audio, poezja), tym wartość atrybutu jest bliższa 1.

[17] M. Lavengood, *Pitch and Pitch Class*, https://viva.pressbooks.pub/openmusictheory/chapter/pitch-and-pitch-class/ (dostęp 23.04.2023)

	Wartości powyżej 0,66 opisują utwory, które prawdopodobnie są w całości wykonane ze słów mówionych. Wartości pomiędzy 0,33 a 0,66 opisują utwory, które mogą zawierać zarówno muzykę jak i mowę, w sekcjach lub warstwowo, włączając w to takie przypadki jak muzyka rap. Wartości poniżej 0,33 najprawdopodobniej reprezentują muzykę i inne utwory nie będące mową.
tempo	Ogólne szacunkowe tempo utworu w uderzeniach na minutę (BPM). W terminologii muzycznej, tempo jest prędkością danego utworu i wywodzi się bezpośrednio ze średniego czasu trwania uderzenia.
time_signature	Szacunkowa sygnatura czasowa (metrum) – określa ile uderzeń jest w każdym takcie. Sygnatura czasowa waha się od 3 do 7 wskazując na metrum od "3/4", do "7/4".
valence	Miara z zakresu od 0 do 1 opisująca pozytywność przekazywaną przez utwór. Utwory z wysoką wartością tej zmiennej brzmią bardziej pozytywnie (np. szczęśliwe, radosne, euforyczne), podczas gdy utwory z niską wartością brzmią bardziej negatywnie (np. smutne, przygnębione, złe).

Źródło: Opracowanie własne na podstawie Spotify Web API Documentation

1.4 Metodologia

Do badania determinant popularności utworów zastosowano cztery metody estymacji modelu:

1. Zmienna objaśniana jest dychotomiczna, więc pierwszą z metod jest binarna regresja logistyczna, która służy do przewidywania obecności danej cechy lub jej braku na podstawie wartości zmiennych objaśniających. Wzór opisujący regresję logistyczną to[18]:

$$P = \frac{e^{logit(p)}}{1 + e^{logit(p)}} = \frac{1}{1 + e^{-logit(p)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

[18] A. Kirpsza, *Zastosowanie regresji logistycznej w studiach nad Unią Europejską*, w: *Metody jakościowe i ilościowe w badaniu organizacji i działania Unii Europejskiej*, K. Ławniczak (red.), Wydział Dziennikarstwa i Nauk Politycznych Uniwersytetu Warszawskiego, Warszawa 2013, s. 11–34.

Otrzymane wyniki służą do oszacowania ilorazów szans czyli stosunku prawdopodobieństwa wystąpienia danego zdarzenia (p) do prawdopodobieństwa, że to zdarzenie nie wystąpi (1-p). Współczynniki regresji logistycznej są obliczane używając metody największej wiarygodności, która polega na stworzeniu funkcji wiarygodności umożliwiającej znalezienie wektora parametrów β, który daje największe prawdopodobieństwo uzyskania empirycznych wartości zmiennej objaśnianej. Funkcja wiarygodności jest zlogarytmizowana:

$$\ln(L) = \sum_{i=1}^{n} \ln(P_i) + (1 - y_i)\ln\,(1 - P_i)$$

Ważnymi założeniami regresji logistycznej oprócz dychotomiczności zmiennej zależnej jest brak współliniowości zmiennych niezależnych oraz odpowiednia liczebność próby.

2. Regresja logistyczna z wykorzystaniem algorytmu *stepwise*[19], którego celem jest wybór najistotniejszych zmiennych do modelu. Ma on dwa rodzaje: *forward selection* i *backward selection*. Budowę modelu przy użyciu algorytmu *stepwise forward selection* zaczyna się od pustego modelu, aby następnie dobierać zmienne po kolei według największej istotności. Do oceny istotności zmiennych zwraca się uwagę na wartości *p-value* dla każdej z nich, a także na ocenę jakości modelu m.in. wskaźniki AIC, BIC czy wyniki F-test[20]. Decyzja o dołączeniu danej zmiennej do modelu następuje, gdy wystąpi znacząca poprawa jakości modelu. Algorytm kończy swoje działanie jeżeli nie jest możliwe dodanie kolejnej zmiennej przy jednoczesnej poprawie oceny jakości całej regresji. Tworzenie modelu z pomocą *stepwise backward selection* rozpoczyna się od modelu ze wszystkimi zmiennymi. Następnie usuwa się stopniowo predyktory o najmniejszej istotności jeżeli poprawia to wyniki regresji. Analogicznie do *forward stepwise selection* algorytm kończy swoje działanie, gdy model nie zawiera już zmiennych, których usunięcie powoduje poprawę jego jakości. W badaniu stosowany jest algorytm łączący obie metody, aby uzyskać najlepsze dopasowanie.
3. Drzewo klasyfikacyjne jest modelem uczenia maszynowego, który powstał w wyniku rekurencyjnego podziału badanego zbioru na *n* niezależnych podzbiorów.

[19] J. Osborne (red.), *Best practices in quantitative methods*, SAGE Publications, Inc., Thousand Oaks 2008, s. 377

[20] M. Wang, J. Wright, R. A. Buswell, A. Brownlee, *A comparison of approaches to stepwise regression for global sensitivity analysis used with evolutionary optimization*, "Proceedings of BS 2013: 13th Conference of the International Building Performance Simulation Association" 2013, s. 2552

Celem budowy takiego modelu jest uzyskanie jak najbardziej homogenicznych podzbiorów pod względem zmiennej objaśnianej[21]. Składa się ono z korzenia, z którego wychodzą gałęzie prowadzące do kolejnych węzłów. W każdym z węzłów przeprowadzany jest test obejmujący jeden z predyktorów, na którego podstawie następuje podział na kolejne podzbiory, reprezentowane przez gałęzie prowadzące do kolejnych węzłów poniżej. Na dole drzewa znajdują się węzły końcowe, czyli liście, przypisujące zmienną zależną do odpowiedniej klasy. Na każdym etapie do określania warunków umożliwiających najlepszy podział zbioru na jednorodne podzbiory analizuje się wszystkie predyktory, aby wybrać ten najskuteczniejszy. Kryterium wyboru testu w węźle jest minimalizacja entropii, czyli wartości oczekiwanej informacji. Tworzy się kolejne węzły aż do momentu spełnienia kryterium stopu – wtedy tworzony jest liść. Do takich kryteriów należy między innymi[22]:

- osiągnięcie ustalonej wcześniej dokładności podziałów,
- uzyskanie zakładanej ilości gałęzi,
- osiągnięcie ilości predyktorów mniejszej niż ustalony próg.

4. Las losowy jest algorytmem uczenia zespołowego, który polega na stworzeniu i połączeniu wielu różnych drzew klasyfikacyjnych. Każde z nich jest tworzone na losowej próbie n obserwacji ze zwracaniem ze zbioru uczącego. Zmienne używane w poszczególnych drzewach są również wybierane losowo. W każdym węźle losuje się m predyktorów (najczęściej $m \approx \sqrt{p}$, gdzie p to liczba wszystkich zmiennych), z których wybrany zostanie najlepszy podział. Na koniec obserwacje przypisywane są do klasy, w której występują najczęściej[23]. Metoda ta pozwala na zmniejszenie wariancji prognozy w porównaniu do pojedynczego drzewa klasyfikacyjnego.

Do oceny jakości wymienionych modeli zastosowano poniższe miary oraz wykresy[24]:

1. Dokładność (*accuracy*) jest obliczona poprzez podzielenie liczby dobrze zaklasyfikowanych obserwacji przez wielkość zbioru.

[21] M. Łapczyński, *Drzewa klasyfikacyjne w badaniach satysfakcji i lojalności klientów*, „StatSoft Polska" 2003, s. 93
[22] G. Wilczewski, *InTrees: Modularne podejście do Drzew Decyzyjnych*, Uniwersytet Mikołaja Kopernika, Toruń 2008, s. 16
[23] R. Górska, P. Staszkiewicz, *Zastosowanie algorytmu lasów losowych do prognozowania modyfikacji opinii biegłego rewidenta*, „Zarządzanie i finanse" 2017, nr 3, s. 342
[24] Z. Vujovic, *Classification Model Evaluation Metrics*, "International Journal of Advanced Computer Science and Application" 2021, vol. 12, s. 599-606

2. Czułość (*sensitivity*) jest obliczona dzieląc liczbę obserwacji poprawnie zaklasyfikowanych pozytywnie przez wszystkie należące do klasy pozytywnej.
3. Specyficzność (*specificity*) oblicza się poprzez podzielenie liczby obserwacji poprawnie zaklasyfikowanych do klasy negatywnej przez wszystkie pozytywne rekordy.
4. F1 (*F-score*) jest obliczony na podstawie miar precyzji i czułości. Precyzja (*precision*) to liczba poprawnie zaklasyfikowanych obserwacji pozytywnych przez całkowitą liczbę rekordów zaklasyfikowanych pozytywnie. Miarę F1 oblicza się ze wzoru:

$$F1 = \frac{2 * precyzja * czułość}{precyzja + czułość}$$

5. Krzywa ROC (*Receiver Operating Characteristics*) – to krzywa pokazująca zależność między skutecznością klasyfikacji obserwacji pozytywnych (czułość) a nieskutecznością klasyfikacji tych negatywnych (1 – specyficzność) na każdym z poziomów prawdopodobieństwa.
6. AUC (*Area Under Curve*) to obszar pod krzywą ROC określający jak dobra jest ta krzywa.

Rozdział II

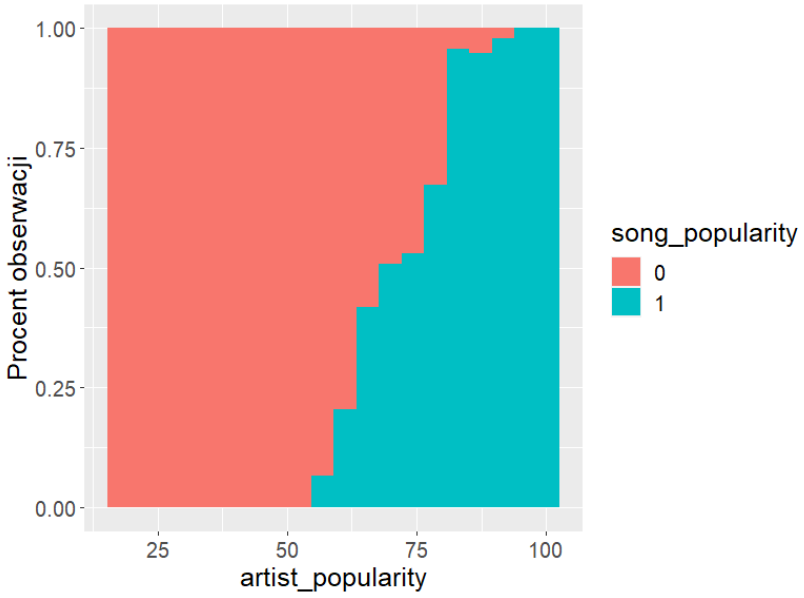
Wyniki analizy

2.1 Eksploracja danych

Utworzona baza danych zawierała zduplikowane utwory, więc pierwszym krokiem w przygotowaniu danych do modelowania było ich usunięcie. Po tej operacji baza zawierała 1903 obserwacje. Kolejnym etapem było utworzenie zmiennej year w celu zbadania wpływu roku, w którym dany utwór powstał na jego popularność. Taka decyzja spowodowana była sposobem obliczania wskaźnika popularity przez Spotify, a konkretnie faktem, że jest on zależny od tego jak niedawno miały miejsce odtworzenia piosenki. Następnym krokiem było zbadanie histogramów zmiennych oraz udziału poszczególnych klas zmiennej zależnej w celu wykrycia obserwacji odstających lub braków danych. Analiza pozwoliła wykryć potrzebę wprowadzenia kilku zmian:

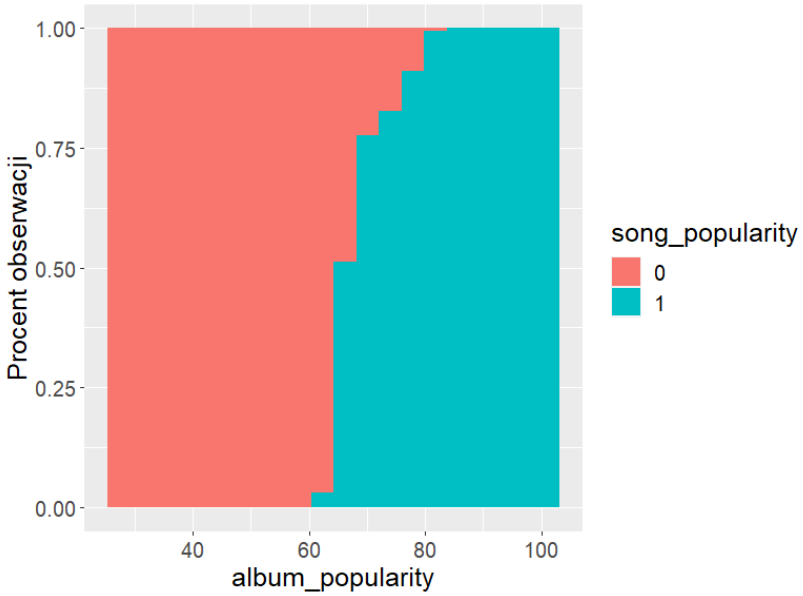
1. Pozbycie się zmiennych *artist_popularity* oraz *album_popularity*

Rysunek 2. Wykres udziału klas zmiennej zależnej na rozkładzie zmiennej *artist_popularity*



Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Rysunek 3. Wykres udziału klas zmiennej zależnej na rozkładzie zmiennej *album_popularity*

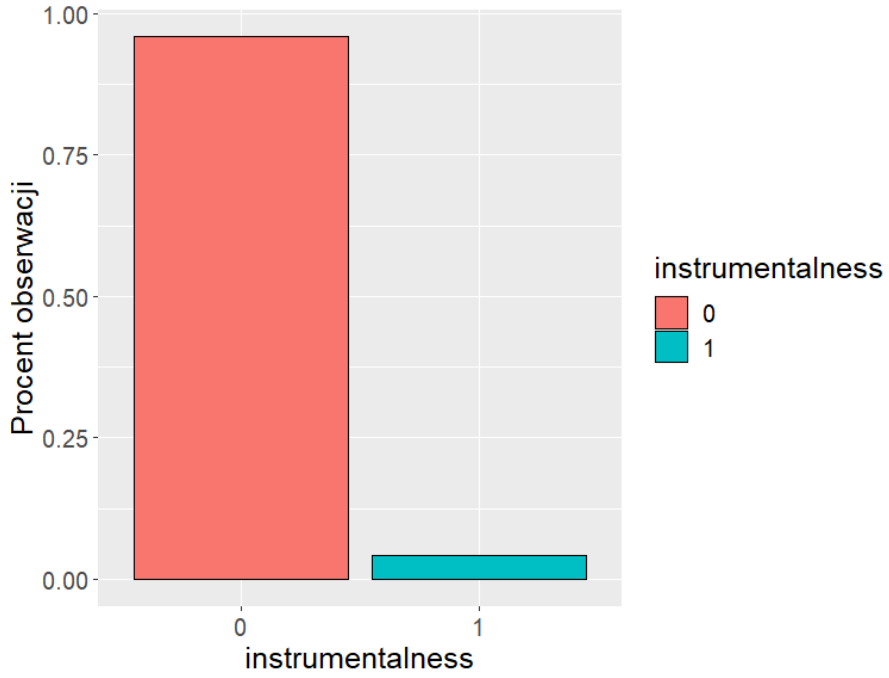


Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Rysunek 2 oraz rysunek 3 pokazują bardzo wysoką korelację obu zmiennych ze zmienną objaśnianą. Z tego powodu oba predyktory nie będą brane pod uwagę w modelach.

2. Zmienne *followers* oraz *tempo* przyjmowały wartości zero dla odpowiednio jednej i dwóch obserwacji, dlatego usunięto te rekordy z danych.
3. Usunięto zmienną *instrumentalness* z bazy z powodu występowania 899 obserwacji, dla których zmienna ta przyjmowała wartość 0. Próba jej przekształcenia w zmienną binarną określającą czy dany utwór jest wokalny czy instrumentalny potwierdziła brak jej zbilansowania, który widać na rysunku 4.

Rysunek 4. Rozkład zmiennej *instrumentalness* po transformacji

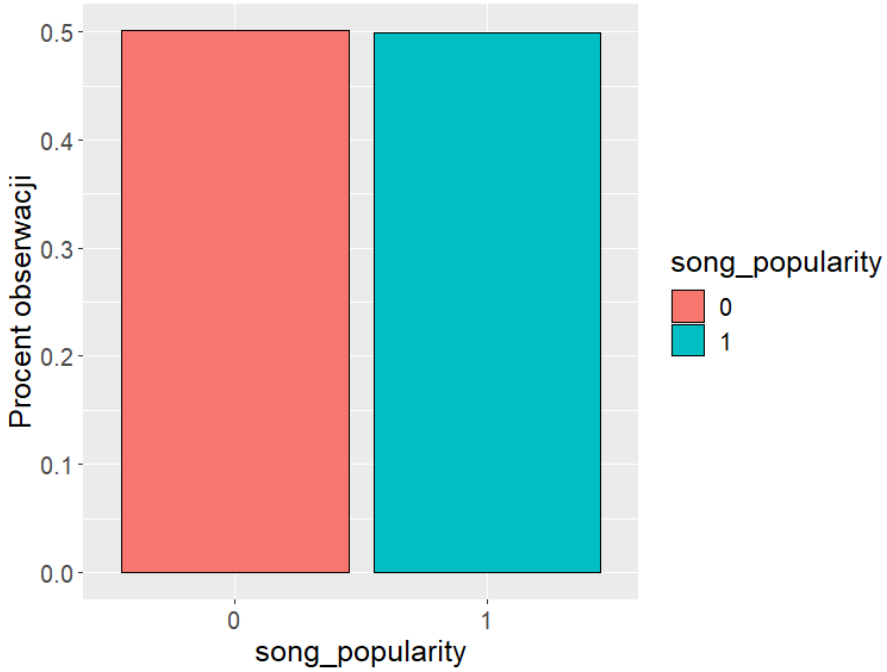


Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

4. Zmienna *key* nie będzie używana w modelu, ponieważ jej wartości względne są nieznaczące w modelu. Przykładowo tonacja C# mająca wartość 1 nie jest lepsza od tonacji C mającej wartość 0.
5. Analiza zmiennych *loudness*, *speechiness* oraz *liveness* wykazała odpowiednio jedną, dwie i trzy obserwacje odstające, więc zostały one usunięte z badanej bazy danych przed rozpoczęciem budowy modeli.
6. Zmienna *time_signature* nie będzie brana pod uwagę w modelach, ponieważ jest ona niezbilansowana – zdecydowana większość (1801) obserwacji przyjmuje wartość 4.
7. Zmienna *followers* została podzielona przez 1000000, ale także miara *duration*_ms (zamieniona na *duration_s)* została podzielona przez 1000 w celu łatwiejszej interpretacji parametrów modelu.

Po wszystkich opisanych powyżej zmianach pozostało 1894 obserwacji. Obie klasy zmiennej objaśnianej posiadają zbliżoną liczbę rekordów, co potwierdza histogram zmiennej *song*_popularity widoczny na rysunku 5:

Rysunek 5. Rozkład zmiennej objaśnianej po przeprowadzonych zmianach w danych



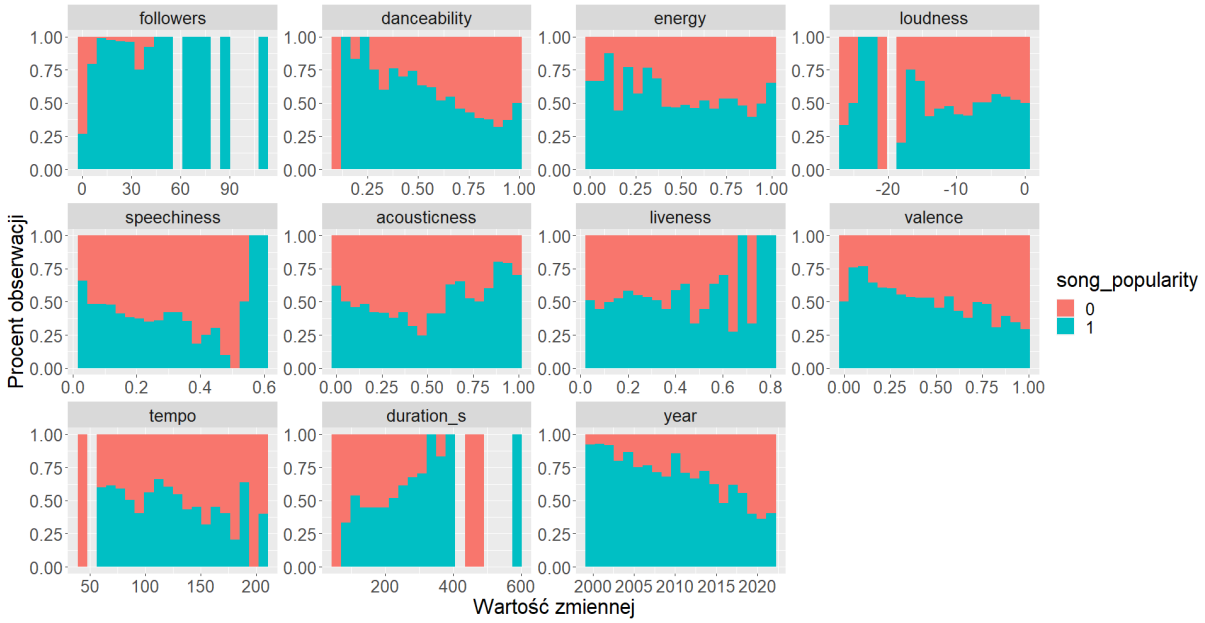
Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Zmienne używane w budowaniu modeli to: *followers, danceability, energy, loudness, mode, speechiness, acousticness, liveness, valence, tempo, duration_s, year*. Utworzono dwa zbiory danych do modelowania: pierwszy z nich zawierający zmienną *followers* oraz drugi bez niej. Powodem takiej decyzji jest wysoka korelacja popularności utworu z liczbą obserwatorów artysty je wykonującego. Uznano jednak, że warto zobaczyć jak zachowa się model zawierający ten predyktor, mając na uwadze wyniki osiągnięte w modelu zawartym w badaniu Myry Interiano et al., a konkretnie użyciu przez autorów zmiennej *superstar*[25].

Poniżej przedstawiono wykresy pokazujące udział klas zmiennej objaśnianej (poza binarną zmienną *mode*). Można na ich podstawie poczynić przypuszczenia dotyczące kierunku zależności poszczególnych predyktorów i zmiennej zależnej. Zmienne *followers* i *liveness* powinny mieć dodatnią relację ze zmienną objaśnianą w przeciwieństwie do zmiennych *danceability, energy, valence* czy *year*. Takie przewidywania są niezgodne z przypuszczeniami wspomnianych w rozdziale pierwszym badań. Taneczność, energiczność oraz pozytywność miały mieć odwrotny wpływ na zmienną zależną. Pozostałe relacje trudniej jest przewidzieć na podstawie wykresów widocznych na rysunku 6.

[25] M. Interiano, K. Kazemi, L. Wang, J. Yang, Z. Yu, N. Komarova, *Musical trends and predictability of success in contemporary songs in and out of the top charts*, The Royal Society Publishing, Irvine 2018, s. 15

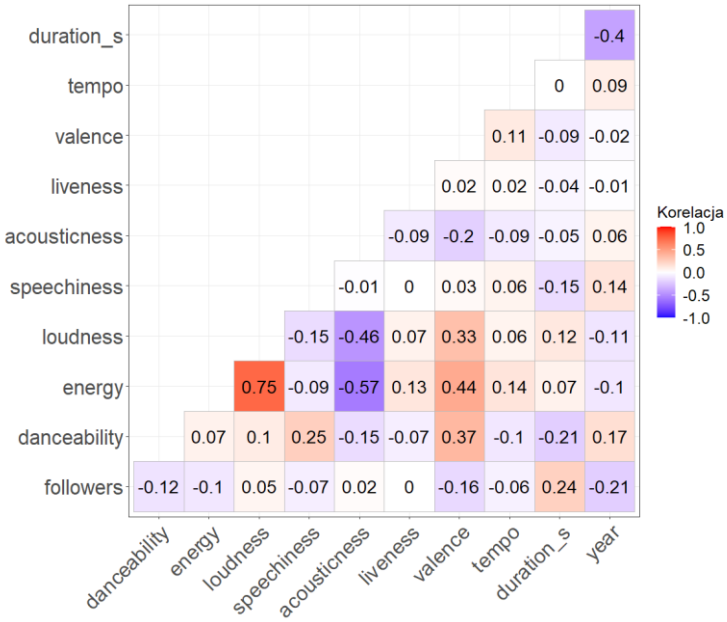
Rysunek 6. Wykresy udziału klas zmiennej zależnej na rozkładach zmiennych objaśniających



Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Na rysunku 7 znajduje się również korelogram zmiennych objaśniających (ponownie z wyłączeniem *mode*). Największa korelacja występuje pomiędzy zmiennymi *energy* oraz *loudness*, ponieważ energiczne utwory często cechują się również większą głośnością. Obie zmienne wykazują również niemałą ujemną korelację z zmienną *acousticness*. Prawdopodobnym powodem jest fakt, że piosenki akustyczne są zazwyczaj spokojniejsze od przeciętnego utworu.

Rysunek 7. Korelogram zmiennych objaśniających



Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

2.2 Wyniki modeli

Przed rozpoczęciem tworzenia modeli podzielono zbiór danych na zbiór uczący i testowy, z czego 75% obserwacji to zbiór treningowy, a 25% to zbiór testowy. Na zbiorze treningowym tworzono modele, a na zbiorze testowym dokonano ich oceny.

2.2.1 Zbiór pierwszy

Regresja logistyczna

Model binarnej regresji logistycznej został zbudowany za pomocą wbudowanej w program R funkcji *glm*. Przyjmując poziom istotności $\alpha = 0{,}05$, w modelu występuje 6 zmiennych istotnych statystycznie: *followers, danceability, speechiness, tempo, duration_s* oraz *year*. Wzrost wartości zmiennej *followers* o milion obserwatorów powoduje wzrost prawdopodobieństwa przypisania do klasy pozytywnej o 36,6% ceteris paribus. Parametry strukturalne dla pozostałych zmiennych istotnych są ujemne, co oznacza, że ich wzrost powoduje spadek prawdopodobieństwa na to, że utwór będzie popularny. Wzrost zmiennej *danceability* o 1% zmniejsza prawdopodobieństwo przypisania obserwacji do klasy pozytywnej średnio o 0,893% jeżeli pozostałe zmienne pozostaną stałe[26]. Największy wpływ na zmniejszenie szansy wśród zmiennych ciągłych na osiągnięcie zakładanej popularności ma *speechiness* – zmniejsza ją średnio o 0,969% ceteris paribus przy wzroście o 1%. W modelu nie występuje współliniowość – wartości czynnika inflacji wariancji nie przekraczają 4. Wpływ liczby obserwatorów, ilości słów mówionych oraz długości utworu jest zgodny z przypuszczeniami oraz opisaną wyżej literaturą w przeciwieństwie do pozostałych istotnych statystycznie zmiennych. Wzrost taneczności i roku wydania powoduje spadek prawdopodobieństwa przypisania obserwacji do klasy pozytywnej w przeciwieństwie do hipotezy pierwszej oraz wyników podobnych badań. Wszystkie opisane wyniki znajdują się w tabeli 2.

Tabela 2. Wyniki estymacji regresji logistycznej na zbiorze pierwszym

Zmienna	Parametr strukturalny	Błąd standardowy	p-value	Istotność
followers	0,3119	0,0285	< 2e-16	***
danceability	-2,2334	0,5651	7,73e-05	***

[26] R. Nijkamp, *Prediction of product success: explaining song popularity by audio features from Spotify data*, 11th IBA Bachelor Thesis Conference, Enschede 2018, s. 8

energy	-0,5445	0,6857	0,4272	
loudness	0,0147	0,0374	0,6949	
mode	0,2655	0,1396	0,0570	
speechiness	-3,4738	0,7630	5,30e-06	***
acousticness	-0,6295	0,3628	0,0827	
liveness	0,5520	0,5645	0,3281	
valence	-0,6332	0,3588	0,0776	
tempo	-0,0074	0,0027	0,0055	**
duration_s	-0,0060	0,0018	0,0006	***
year	-0,0492	0,0171	0,0039	**

‘***’ p= 0,001 ‘**’ p = 0,01 ‘*’ p = 0,05

Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Regresja logistyczna z wykorzystaniem algorytmu *stepwise*

Wyniki estymacji wykonanej używając funkcji *stepAIC* z pakietu *MASS* wskazują na statystyczną istotność ośmiu zmiennych przy poziomie istotności $\alpha = 0,05$. Poza zmiennymi z poprzedniego modelu istotne są również zmienne *mode* i *valence*. Wzrost zmiennej *valence* o 1%, zmniejsza prawdopodobieństwo przypisania obserwacji do klasy pozytywnej średnio o 0,469% jeżeli pozostałe zmienne pozostaną stałe. Jeżeli badany utwór jest w tonacji durowej, zwiększa się szansa na osiągnięcie zakładanej popularności średnio o 30,4% ceteris paribus. W modelu nie występuje współliniowość – wartości czynnika inflacji wariancji są mniejsze niż 2. Wpływ zmiennej określającej tonację utworu jest zgodny z przypuszczeniami w przeciwieństwie do pozytywności. Wszystkie opisane wyniki znajdują się w tabeli 3.

Tabela 3. Wyniki estymacji regresji logistycznej z wykorzystaniem algorytmu *stepwise* na zbiorze pierwszym

Zmienna	Parametr strukturalny	Błąd standardowy	p-value	Istotność
followers	0,3140	0,0283	< 2e-16	***
danceability	-2,2092	0,5554	6,96e-05	***
mode	0,2772	0,1389	0,0461	*
speechiness	-3,4591	0,7582	5,06e-06	***
acousticness	-0,5110	0,2980	0,0864	
valence	-0,7216	0,3296	0,0286	*

tempo	-0,0076	0,0027	0,0045	**
duration_s	-0,0062	0,0017	0,0004	***
year	-0,0487	0,0170	0,0041	**

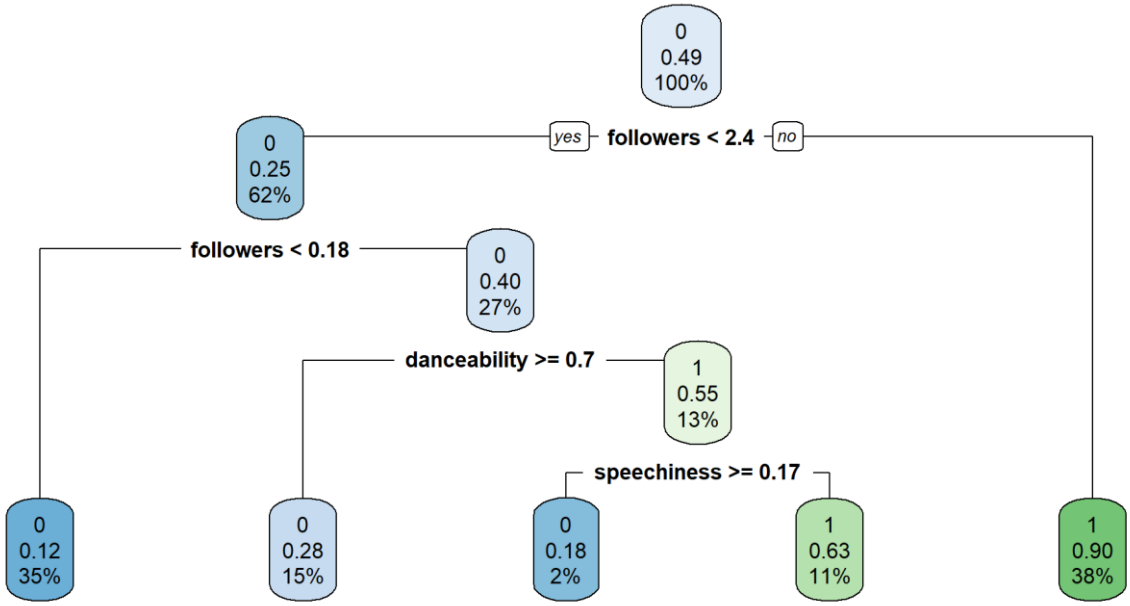
‘***’ p= 0,001 ‘**’ p = 0,01 ‘*’ p = 0,05

Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Drzewo klasyfikacyjne

Model drzewa klasyfikacyjnego powstał przy użyciu funkcji *rpart* z pakietu *rpart,* a wizualizacja została wykonana za pomocą funkcji *rpart.plot* z pakietu *rpart.plot*. Drzewo z rysunku 8 generuje 5 reguł decyzyjnych i ma głębokość 4. Pierwszy test sprawdza czy liczba obserwatorów artysty wykonującego dany utwór jest mniejsza niż 2,4 miliona osób, Jeżeli jest większa to obserwacja przypisana jest do klasy pozytywnej, Taka sytuacja występuje w 38% wszystkich obserwacji, W przeciwnym wypadku przeprowadzany jest kolejny test na zmiennej *followers*, Sprawdza on czy liczba obserwatorów nie przekracza 180 000 osób, Jeżeli taka jest sytuacja to utwór zaliczany jest jako niepopularny i dzieje się tak w 35% obserwacji, Jeżeli liczba ta przekracza 180 000 obserwatorów przeprowadzany jest test na zmiennej *danceability*, Utwór przypisywany jest do klasy negatywnej gdy jego „taneczność” jest większa lub równa 0,7 (15% obserwacji). W przeciwnym razie przeprowadzany jest ostatni test na zmiennej *speechiness*. Jeśli przekracza ona lub równa się wartości 0,17 to utwór jest uznany za niepopularny i ma to miejsce w 2% obserwacji, a jeżeli ta wartość jest mniejsza to utwór jest zaliczony do klasy pozytywnej. Z modelu wynika, że zaliczeniu utworu do popularnych sprzyja duża ilość obserwujących, mała ilość słów mówionych i mała taneczność.

Rysunek 8. Drzewo klasyfikacyjne zbudowane na zbiorze pierwszym

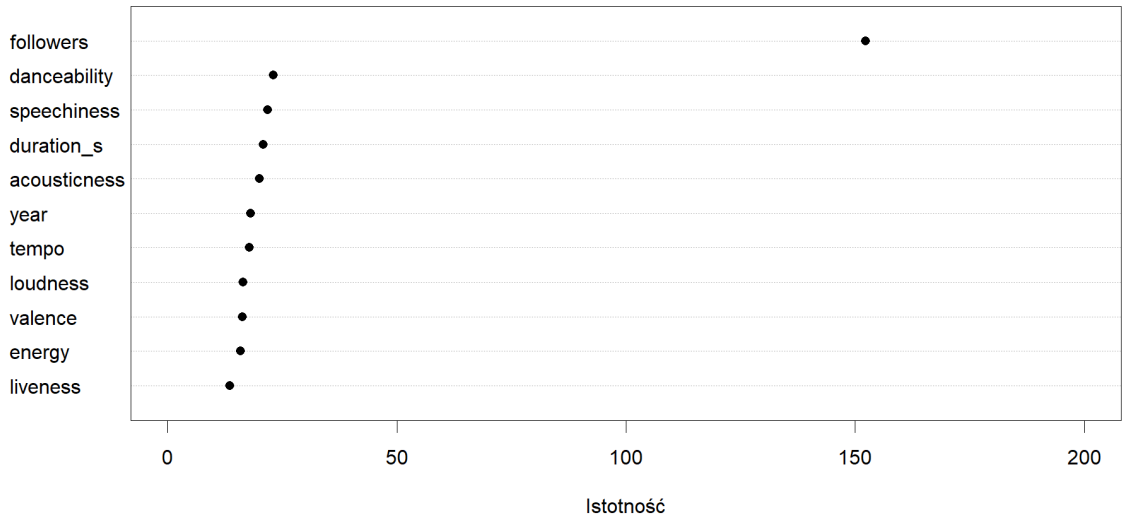


Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Las losowy

Na rysunku 9 przedstawiono wykres istotności zmiennych objaśniających w lesie losowym. Zgodnie z przewidywaniami zmienna *followers* zdecydowanie przeważa istotnością nad resztą zmiennych. Poza liczbą obserwatorów najistotniejsze zmienne są w większości takie same jak w pozostałych modelach z wyjątkiem zmiennej *acousticness*, która nie była istotna w regresji logistycznej, jak również nie pojawiła się w testach w drzewie klasyfikacyjnym.

Rysunek 9. Wykres istotności zmiennych objaśniających w lesie losowym zbudowanym na zbiorze pierwszym



Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Ocena jakości modeli

Wartości zawarte w tabeli 4 wskazują, że las losowy jest najskuteczniejszym ze zbudowanych modeli. Dokładność tej metody wynosi ponad 84%. Najniższą dokładność ma regresja logistyczna (niezależnie czy wykorzystywany jest algorytm *stepwise*) – różnica między jej wartością a wartością dla lasu losowego to ponad 3 punkty procentowe. Wyniki obu regresji logistycznych cechują się wysoką specyficznością i niską czułością w porównaniu z pozostałymi modelami. Oznacza to, że lepiej przypisują obserwacje do klasy negatywnej, ale jednocześnie słabiej klasyfikują do pozytywnej. Las losowy ma największą dokładność oraz miarę F1, co wskazuje na optymalność tej metody. Wartość AUC potwierdza te przypuszczenia.

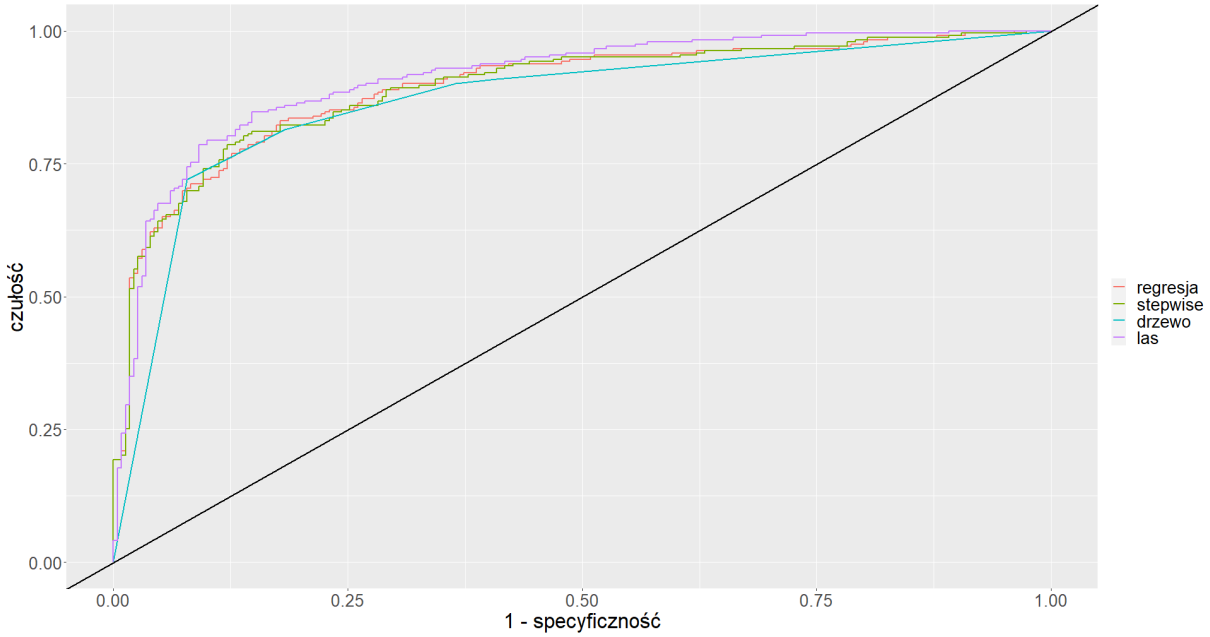
Tabela 4. Ocena jakości zbudowanych modeli na zbiorze pierwszym

	Regresja logistyczna	**Regresja logistyczna stepwise**	**Drzewo klasyfikacyjne**	**Las losowy**
dokładność	0,8076	0,8076	0,8161	0,8414
czułość	0,7202	0,7161	0,8148	0,8231
specyficzność	0,9000	0,9044	0,8174	0,8609
F1	0,7937	0,7927	0,8199	0,8421
AUC	0,8934	0,8929	0,8689	0,9096

Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Krzywe ROC na rysunku 10 nie dają jednoznacznej odpowiedzi, który model najskuteczniej klasyfikuje obserwacje. Dla wysokich wartości specyficzności lepszym wyborem wydaje się jedna z regresji, ale po osiągnięciu pewnego poziomu specyficzności wybór lasu losowego jest pewniejszy. Różnice w wartości AUC są niewielkie, co tym bardziej wskazuje na niejednoznaczność wskazania najlepszego modelu, lecz wartości dokładności i F1 wskazują na wybór lasu losowego.

Rysunek 10. Krzywe ROC dla zbudowanych modeli



Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

2.2.2 Zbiór drugi

Wszystkie modele zostały zbudowane analogicznie do modeli w zbiorze pierwszym.

Regresja logistyczna

Przyjmując poziom istotności α = 0,05, w modelu występuje 9 zmiennych istotnych statystycznie. Jedyne nieznaczące zmienne w modelu to: *liveness* i *duration*_s. Wynik ten jest zaskakujący – po usunięciu liczby obserwatorów długość utworu przestała być istotna statystycznie. Znak przy parametrach strukturalnych nie zmienił się. Kilka zmiennych, które w pierwszym zbiorze nie były istotne statystycznie, wykazują istotność w drugim zbiorze. Wzrost wartości zmiennej *energy* o 1% powoduje spadek prawdopodobieństwa przypisania do klasy pozytywnej średnio o 0,891% ceteris paribus – jest to niezgodne z postawionymi hipotezami, a także z przypuszczeniami na podstawie opisanej literatury. Wzrost zmiennej *loudness* o 1% zwiększa prawdopodobieństwo zaklasyfikowania obserwacji pozytywnie przeciętnie o 0,167% jeżeli pozostałe zmienne pozostaną stałe. Istotny wpływa wykazuje również akustyczność – jej wzrost o 1% powoduje spadek prawdopodobieństwa, że utwór zostanie uznany za popularny średnio o 0,648%. W modelu nie występuje współliniowość – wartości czynnika inflacji wariancji nie przekraczają 4. Wszystkie opisane wyniki znajdują się w tabeli 5.

Tabela 5. Wyniki estymacji regresji logistycznej na zbiorze drugim

Zmienna	Parametr strukturalny	Błąd standardowy	p-value	Istotność
danceability	-2,2150	0,4913	6.52e-06	***
energy	-2.8630	0,5966	1.59e-06	***
loudness	0,1548	0,0333	3.27e-06	***
mode	0,2457	0,1184	0,0380	*
speechiness	-2,6587	0,6167	1.62e-05	***
acousticness	-1,0435	0,3120	0,0008	***
liveness	0,6889	0,4869	0,1571	
valence	-1,1976	0,3084	0,0001	***
tempo	-0,0061	0,0023	0,0073	**
duration_s	-0,0020	0,0014	0,1490	
year	-0,1286	0,0144	< 2e-16	***

‘***’ p= 0,001 ‘**’ p = 0,01 ‘*’ p = 0,05

Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Regresja logistyczna z wykorzystaniem algorytmu *stepwise*

W zbiorze drugim regresja logistyczna wykorzystująca algorytm *stepwise* do selekcji zmiennych użytych w modelu daje dokładnie takie same rezultaty co model zawierający wszystkie predyktory. Oznacza to, że nie jest możliwe zbudowanie modelu z lepszymi wynikami biorąc pod uwagę zmienne zawarte w zbiorze drugim. Wyniki tej metody znajdują się w tabeli 6.

Tabela 6. Wyniki estymacji regresji logistycznej z wykorzystaniem algorytmu stepwise na zbiorze drugim

Zmienna	Parametr strukturalny	Błąd standardowy	p-value	Istotność
danceability	-2,2150	0,4913	6.52e-06	***
energy	-2.8630	0,5966	1.59e-06	***
loudness	0,1548	0,0333	3.27e-06	***
mode	0,2457	0,1184	0,0380	*
speechiness	-2,6587	0,6167	1.62e-05	***
acousticness	-1,0435	0,3120	0,0008	***
liveness	0,6889	0,4869	0,1571	

valence	-1,1976	0,3084	0,0001	***
tempo	-0,0061	0,0023	0,0073	**
duration_s	-0,0020	0,0014	0,1490	
year	-0,1286	0,0144	< 2e-16	***

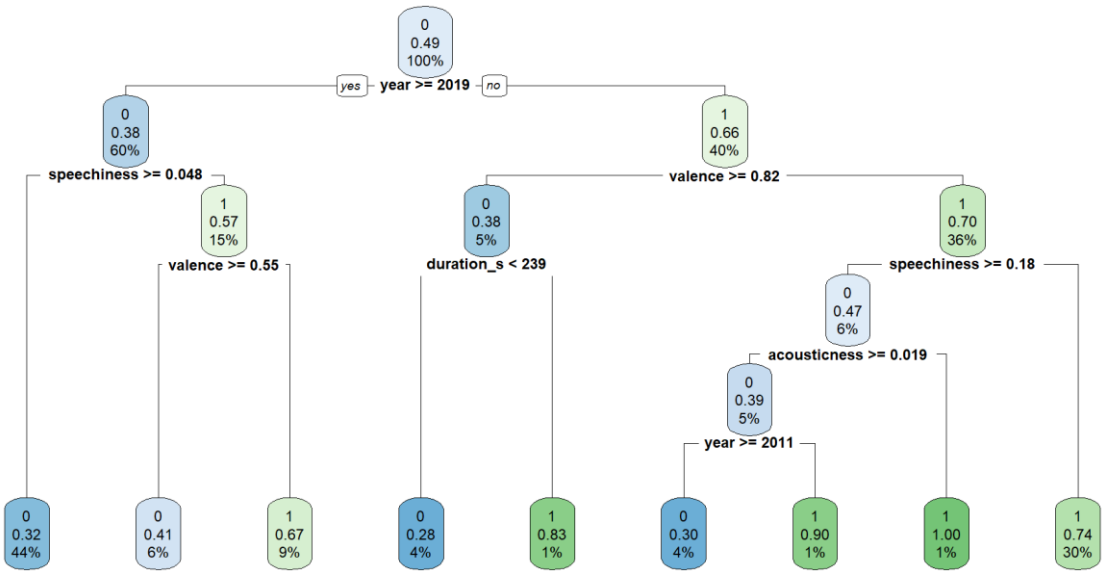
‘***’ p= 0,001 ‘**’ p = 0,01 ‘*’ p = 0,05

Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Drzewo klasyfikacyjne

Drzewo z rysunku 11 generuje 9 reguł decyzyjnych i ma głębokość 5. Pierwszy test sprawdza czy dany utwór został wydany w 2019 roku lub później. Jeżeli tak się dzieje to przeprowadzamy kolejny test – tym razem na zmiennej *speechiness*. Jeżeli jest ona większa lub równa 0,048 to przypisuje się obserwację do klasy negatywnej. Taka sytuacja ma miejsce w 44% obserwacji. W przeciwnym wypadku przeprowadzany jest test na zmiennej *valence*. Sprawdza on czy miara pozytywności przekracza lub równa się 0,55. Jeżeli tak się dzieje to utwór zaliczany jest jako niepopularny i dzieje się tak w 6% obserwacji, a w przeciwnym wypadku klasyfikujemy ją pozytywnie dla 9% rekordów. W przypadku gdy utwór został wypuszczony przed 2019 rokiem przeprowadzany jest test na zmiennej *valence*. Jeżeli jest większa lub równa 0,82 to sprawdzamy czy długość utworu jest mniejsza niż 239 sekund. W takiej sytuacji utwór jest uznawany za niepopularny, a gdy jest odwrotnie zakłada się, że jest on popularny. Wracając do testu, w którym sprawdzano czy miara pozytywności jest niemniejsza od 0,82 – jeżeli nie jest wykonujemy test na zmiennej określającej ilość słów mówionych w utworze. Jeśli jest mniejsza od 0,18 to dla 30% obserwacji przypisujemy utwór do klasy pozytywnej. W przeciwnym razie przeprowadzany jest test na zmiennej *acousticness*. Jeżeli jest mniejsza niż 0,019 to utwór jest uznany za popularny, a w przeciwnym wypadku wykonujemy ostatni test sprawdzając czy utwór powstał przed rokiem 2011. W takiej sytuacji zakłada się, że jest on popularny. Z modelu wynika, że zaliczeniu utworu do popularnych sprzyja wczesny rok wydania, mała ilość słów mówionych, nieduża pozytywność, mała akustyczność i dłuższy czas trwania.

Rysunek 11. Drzewo klasyfikacyjne zbudowane na zbiorze drugim

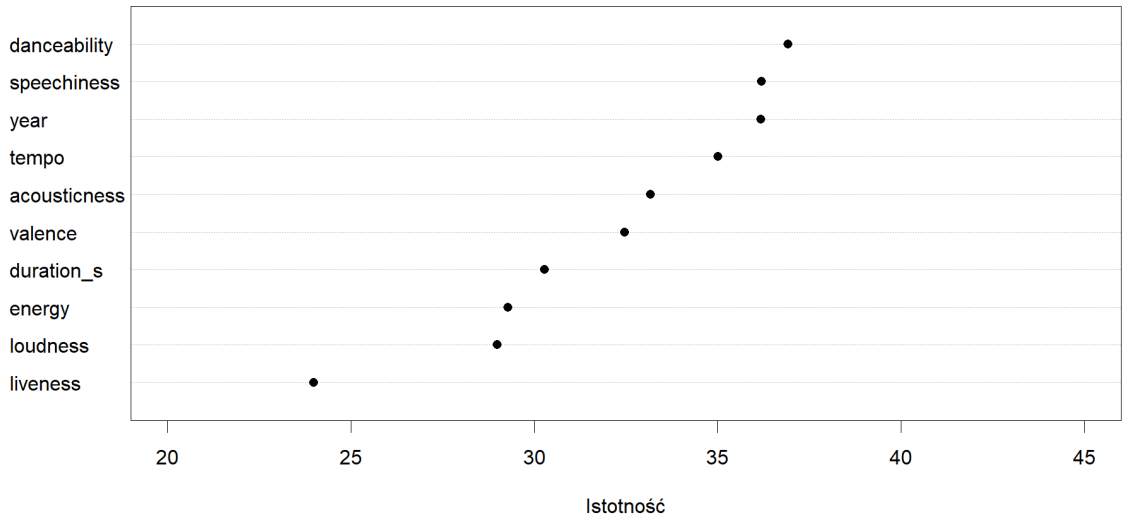


Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Las losowy

Rysunek 12 zawiera wykres wskazujący na brak dominującej zmiennej w modelu. Najistotniejsze zmienne to taneczność, ilość słów mówionych i rok wydania utworu. Predyktory o najwyższych wartościach w większości pokrywają się z poprzednimi modelami używającymi zbioru drugiego. W porównaniu z lasem losowym ze zbioru pierwszego można zauważyć znaczny spadek relatywnej istotności zmiennej *duration_s*, co zauważono również w regresji logistycznej.

Rysunek 12. Wykres istotności zmiennych objaśniających w lesie losowym zbudowanym na zbiorze drugim



Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Ocena jakości modeli

Wartości zawarte w tabeli 7 wskazują na zdecydowane obniżenie jakości modeli po wyrzuceniu zmiennej *followers*. Najskuteczniejszym z nich wydaje się być ponownie las losowy. Dokładność tej metody wynosi prawie 70%. Najniższą dokładność ma drzewo klasyfikacyjne– różnica między jej wartością a wartością dla lasu losowego to ponad 5 punktów procentowych. Ocena regresji logistycznej z wykorzystaniem algorytmu *stepwise* ma takie same wyniki jak regresja ze wszystkimi zmiennymi ze zbioru tak jak wspomniano wyżej. Wyniki obu regresji logistycznych cechują się ponownie wysoką specyficznością i niską czułością w porównaniu z pozostałymi modelami. Oznacza to, że lepiej przypisują obserwacje do klasy negatywnej, ale jednocześnie słabiej do klasy pozytywnej. Las losowy ma największą dokładność, miarę F1 i wartość AUC. Można więc stwierdzić, że najlepiej prognozuje popularność utworów.

Tabela 7. Ocena jakości zbudowanych modeli na zbiorze drugim

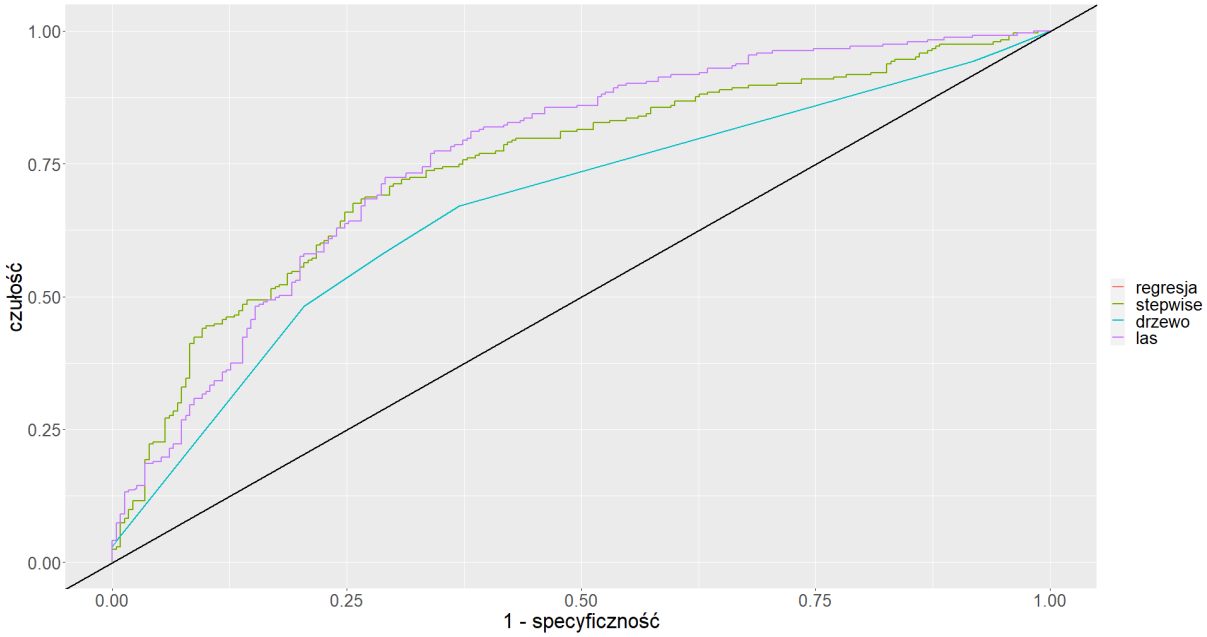
	Regresja logistyczna	**Regresja logistyczna stepwise**	**Drzewo klasyfikacyjne**	**Las losowy**
dokładność	0.6913	0.6913	0,6448	0,6998
czułość	0.6296	0.6296	0,5844	0,6831

specyficzność	0,7565	0,7565	0,7087	0,7174
F1	0,6770	0,6770	0,6283	0,7004
AUC	0,7460	0,7460	0,6699	0,7640

Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Krzywe ROC, znajdujące się na rysunku 13, dla regresji i lasu losowego kilka razy przecinają się, więc nie jesteśmy w stanie stwierdzić, który model klasyfikuje najlepiej korzystając tylko z poniższego wykresu. Wyraźnie widać jednak przewagę obu wyżej wymienionych metod nad drzewem klasyfikacyjnym. Wartości AUC wskazują jednak nieznaczną przewagę lasu losowego.

Rysunek 13. Krzywe ROC dla zbudowanych modeli na zbiorze drugim



Źródło: Opracowanie własne na podstawie danych ze Spotify Web API

Zakończenie

W pracy przeprowadzona została próba identyfikacji czynników wpływających na popularność utworów muzycznych na rynkach krajów należących do Unii Europejskiej z premierą w latach 2000-2022 oraz przewidywania sukcesu odnoszonego przez dany utwór przy użyciu metod uczenia maszynowego. Wykonana analiza pozwoliła na wyciągnięcie pewnych wniosków na temat tego, co wpływa na badaną popularność oraz które z używanych modeli pozwalają na najskuteczniejszą jej predykcję. Umożliwiła również weryfikacje hipotez postawionych w rozdziale 1.1.

Największą istotność wśród zmiennych zastosowanych w modelu wykazała zmienna określająca liczbę obserwatorów artysty wykonującego daną piosenkę. Okazała się ona znacznie przewyższać wpływ innych zmiennych, co zostało zobrazowane przez wykres istotności zmiennych wykorzystanych do budowy lasu losowego. Jest to spodziewany wynik ze względu na korelację zmiennej objaśnianej ze zmienną *followers*. Została ona jednak zawarta w modelu z kilku wymienionych wcześniej w pracy powodów. Były to próba pokazania wpływu innych użytkowników Spotify na decyzję o słuchaniu danego utworu (im większa liczba obserwujących tym łatwiej przyciągnąć kolejnych słuchaczy oraz prawdopodobna chęć przyporządkowania się gustom szerszej grupy) oraz uwzględnienie efektu „supergwiazdy" opisanego w jednym z wymienionych w niniejszej pracy badań. Znaczenie liczby osób obserwujących danego twórcę jest również zauważalne w wynikach oceny jakości testowanych modeli w pracy. Modele na zbiorze drugim (bez zmiennej *followers*) osiągnęły znacznie słabsze wyniki. Dokładność najskuteczniejszego modelu spadła o ponad 14 punktów procentowych. Drugą co do wartości istotności w obu zbiorach była taneczność utworu. Jej wzrost powodował spadek prawdopodobieństwa przypisania obserwacji do klasy pozytywnej, co jest niezgodne z przedstawioną w rozdziale pierwszym literaturą. Jednym z powodów takiej różnicy może być badany zakres – przykładowo słuchacze w krajach Unii Europejskiej mogą wykazywać inne skłonności niż osoby w Stanach Zjednoczonych. Poza tanecznością wysoką istotnością wykazała się zmienna określająca ilość mowy w utworze. *Speechiness* miała również negatywną relację ze zmienną *song_popularity* – jej wzrost powodował spadek prawdopodobieństwa, że dany utwór jest popularny. W tym przypadku jednak jest to zgodne z przypuszczeniami postawionymi zgodnie z poruszającą ten temat literaturą. Z czynników nienależących do grupy cech dźwiękowych znaczący wpływ miał rok wydania danego utworu. Jego relacja ze zmienną zależną była niespodziewana – im piosenka była wydana wcześniej, tym większe było prawdopodobieństwo osiągnięcia przez nią sukcesu. Poprzednie badania oraz

sposób w jaki tworzony jest indeks popularności przemawiały za odwrotnym oddziaływaniem m.in. dlatego, że popularność mierzona przez Spotify zależy od tego jak niedawno miały miejsce odtworzenia utworu. Oznacza to, że utwory, które są odtwarzane często obecnie będą miały większą popularność od tych odtwarzanych wiele razy w przeszłości. Interesujące rezultaty wykazały również zmienne dotyczące długości trwania, energiczności oraz głośności utworu. *Duration_s* przestała być istotna gdy badany był zbiór drugi, natomiast *energy* i *loudness* stały się wtedy dopiero istotne.

Oprócz identyfikacji czynników tematem niniejszej pracy była możliwość predykcji popularności osiąganej przez badany utwór. Do tego celu zastosowano cztery modele uczenia maszynowego: regresję logistyczną, regresję logistyczną z wykorzystaniem algorytmu *stepwise,* drzewo klasyfikacyjne oraz las losowy. Ocenę jakości tych metod oraz ich rezultaty należy podzielić na dwa zbiory. W pierwszym z nich dokładność wahała się pomiędzy 80-85%. Podobne wartości przyjmowała miara F1, a AUC wynosiło około 0,9. Najskuteczniejszym modelem biorąc pod uwagę te miary okazał się las losowy z zauważalną przewagą nad pozostałymi modelami z wyjątkiem krzywych ROC. Regresja logistyczna (niezależnie od wykorzystania algorytmu *stepwise*) lepiej klasyfikowała obserwacje do klasy negatywnej jednak znacząco słabiej robiła to dla klasy pozytywnej. Na podstawie tych informacji zdecydowano, że najlepszym klasyfikatorem jest w tej sytuacji las losowy. W zbiorze drugim różnice między jakością regresji a jakością lasu losowego były mniejsze. Jedynie drzewo klasyfikacyjne zostało w tyle z istotnie mniejszymi wartościami dla każdej z miar. Pomimo zmniejszenia rozbieżności las losowy pozostał najskuteczniejszym klasyfikatorem. Ważne jest jednak, aby pamiętać, że różnice były niewielkie, co było szczególnie widoczne na krzywych ROC oraz w wartościach dokładności (mniej niż 1 punkt procentowy różnicy). Modele zbudowane na zbiorze drugim wykazują znacznie niższą skuteczność od tych na zbiorze pierwszym. Dokładność na poziomie blisko 70% nie jest wynikiem w pełni satysfakcjonującym, ale pokazuje jednak występowanie wpływu cech dźwiękowych na sukces piosenek.

Przeprowadzono również weryfikację hipotez postawionych w rozdziale 1.1 niniejszej pracy. Wyniki estymacji modeli wskazują na prawdziwość hipotezy pierwszej o dużej istotności zmiennej określającej liczbę obserwatorów danego twórcy oraz jej pozytywnym wpływie na sukces odniesiony przez dany utwór muzyczny. Istotność zmiennych *danceability* i *tempo* została potwierdzona, ale zależność między nimi a zmienną objaśnianą jest odwrotna niż zakładano – ich wzrost powoduje spadek prawdopodobieństwa zaklasyfikowania obserwacji do klasy pozytywnej. Zmienna *energy* była istotna statystycznie tylko w zbiorze

drugim. Jej wpływ na popularność jest również przeciwny do tego, którego oczekiwano na podstawie dostępnych badań. Hipoteza trzecia jest zgodna z wynikami niniejszej analizy – zmienna *speechiness* jest istotna statystycznie i wykazuje negatywny wpływ na zmienną objaśnianą.

Rezultaty osiągnięte w analizie badanego zjawiska pokazują, że badane determinanty mają istotny wpływ na zmienną zależną – popularność utworów muzycznych. Na podstawie niniejszej pracy można jednak wywnioskować, że w celu prognozowania rozpoznawalności piosenek nie są one wystarczające. Literatura poruszająca ten temat sugeruje wpływ wielu innych czynników m.in. tekst utworu, gatunek, strategie marketingowe zastosowane przy jego promocji czy wzmianki na portalach społecznościowych. Uwzględnienie tych zmiennych prawdopodobnie zwiększyłoby skuteczność predykcji.

Bibliografia

1. Araujo C., Cristo M., Giusti R., *Predicting Music Popularity on Streaming Platforms*, Federal University of Amazonas, Manaus 2019
2. Duda A., Jonek-Kowalska I., *Analiza danych wpływających na popularność produktów na międzynarodowym rynku muzycznym*, "Management and Quality – Zarządzanie i jakość" 2022, vol. 4, no. 3
3. Górska R., Staszkiewicz P., *Zastosowanie algorytmu lasów losowych do prognozowania modyfikacji opinii biegłego rewidenta*, „Zarządzanie i finanse" 2017, nr 3
4. Interiano M., Kazemi K., Wang L., Yang J., Yu Z., Komarova N., *Musical trends and predictability of success in contemporary songs in and out of the top charts*, The Royal Society Publishing, Irvine 2018
5. International Federation of the Phonographic Industry, *Industry data*, https://www.ifpi.org/our-industry/industry-data/ (dostęp 23.04.2023)
6. Kirpsza A., *Zastosowanie regresji logistycznej w studiach nad Unią Europejską*, w: *Metody jakościowe i ilościowe w badaniu organizacji i działania Unii Europejskiej*, K. Ławniczak (red.), Wydział Dziennikarstwa i Nauk Politycznych Uniwersytetu Warszawskiego, Warszawa 2013
7. Lavengood M., *Pitch and Pitch Class,* https://viva.pressbooks.pub/openmusictheory/chapter/pitch-and-pitch-class/ (dostęp 23.04.2023)
8. Łapczyński M., *Drzewa klasyfikacyjne w badaniach satysfakcji i lojalności klientów,* „StatSoft Polska" 2003
9. Nijkamp R., *Prediction of product success: explaining song popularity by audio features from Spotify data*, 11th IBA Bachelor Thesis Conference, Enschede 2018
10. Osborne J. (red.), *Best practices in quantitative methods*, SAGE Publications, Inc., Thousand Oaks 2008
11. Raza A. H., Nanath K., *Predicting a Hit Song with Machine Learning: Is there an apriori secret formula?*, „2020 International Conference on Data Science, Artificial Intelligence and Business Analytics (DATABIA)", Medan 2020
12. Salganik M. J., Dodds P. S., Watts D. J., *Experimental study of inequality and unpredictability in an artificial cultural market*, "Science" 2006, vol. 311
13. Spotify, *About Spotify*, https://newsroom.spotify.com/company-info/ (dostęp 23.04.2023)
14. Spotify for Developers, *Web API Documentation*, https://developer.spotify.com/documentation/web-api (dostęp 23.04.2023)
15. Tsiara E., Tjortjis C., *Using Twitter to Predict Chart Position for Songs,* w: "Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology", vol. 583, Maglogiannis I., Iliadis L., Pimenidis E. (eds.), Springer, Neos Marmaras 2020
16. Vujovic Z., *Classification Model Evaluation Metrics*, "International Journal of Advanced Computer Science and Application" 2021, vol. 12
17. M. Wang, J. Wright, R. A. Buswell, A. Brownlee, *A comparison of approaches to stepwise regression for global sensitivity analysis used with evolutionary optimization*, "Proceedings of BS 2013: 13th Conference of the International Building Performance Simulation Association" 2013

18. G. Wilczewski, *InTrees: Modularne podejście do Drzew Decyzyjnych*, Uniwersytet Mikołaja Kopernika, Toruń 2008

Spis tabel

Spis rysunków

Streszczenie

Celem pracy jest identyfikacja czynników wpływających na popularność utworów muzycznych w krajach Unii Europejskiej w latach 2000-2022 oraz prognozowanie odnoszonego przez nie sukcesu. Do tego celu skorzystano z danych dostępnych w Spotify Web API, a do modelowania skorzystano z metod uczenia maszynowego: regresji logistycznej, regresji logistycznej z wykorzystaniem algorytmu *stepwise*, drzewa klasyfikacyjnego oraz lasu losowego. Wszystkie analizy zostały wykonane w języku programowania R. W pierwszym rozdziale dokonano przeglądu literatury poruszającej podobną tematykę, przedstawiono proces pozyskania danych oraz opis zmiennych, a także opisano metodologię. W drugim rozdziale zaprezentowano proces eksploracji danych oraz dokonane zmiany w bazie danych. Przedstawiono również wyniki estymacji modeli, które budowano na dwóch zbiorach zmiennych oraz ocenę ich jakości.

Słowa kluczowe: popularność utworów, Spotify, uczenie maszynowe, regresja logistyczna, drzewo decyzyjne, las losowy