Inżynieria Uczenia Maszynowego

Bartosz Pawlak
Kacper Tomczykowski

Temat

Większość serwisów udostępniających muzykę czy filmiki, poleca coś swoim użytkownikom. Przyszedł czas, abyśmy zaczęli robić tak samo.

Określenie celu biznesowego

Celem biznesowym jest polecanie naszym użytkownikom nowych piosenek, które mogą być popularne w nadchodzącym miesiącu. Dzięki temu będą oni bardziej zadowoleni z serwisu i potencjalnie będą w nim spędzać więcej czasu co przyczyni się do zwiększenia przychodów platformy. Chcemy stworzyć model, który będzie tworzył playlistę z utworów, w gatunku ulubionym przez użytkownika, których on jeszcze nie zna. Taka playlista będzie miała od 10 do 20 utworów. Jeśli słuchacz zna większość piosenek ze swojego ulubionego gatunku uzupełnimy listę najlepszymi utworami z innych gatunków. Nowi słuchacze którzy nie mają jeszcze przypisanego ulubionego gatunku dostaną playlistę składającą się ze wszystkich najpopularniejszych piosenek .

Zadania modelowania

W ramach zadania analitycznego zajmujemy się przyporządkowaniem utworom dostępnym na portalu Pozytywka miary, która określi ich szansę na popularność w przyszłości, na podstawie danych zebranych przez ten portal. Popularność piosenki będzie mierzona przez ilość jej odtworzeni i zdobytych lików w przeciągu miesiąca od predykcji(dokładniej poniżej). Celem jest określenie, które utwory będą popularne oraz które z tych utworów pasują do użytkownika, dla którego przygotowywana jest playlista.

Ponieważ funkcja celu jest ciągła nasz model będzie regresją. W ramach prostego modelu planujemy użyć regresji liniowej, a w ramach bardziej złożonego – sieci neuronowej.

Biznesowe kryterium sukcesu

Po wdrożeniu naszego modelu system będzie przydzielał użytkowników do grupy kontrolnej i grupy doświadczalnej. Po przydzieleniu do grupy, użytkownik pozostanie w niej na zawsze. Użytkownicy będą do nich przydzielani losowo. Grupa kontrolna w każdym miesiącu będzie korzystać z portalu w niezmieniony sposób. Grupa doświadczalna będzie korzystać z portalu z zaproponowanymi przez model playlistami. Co miesiąc będzie przeprowadzana weryfikacja czy osoby z grupy doświadczalnej spędzają więcej czasu w serwisie na słuchaniu poleconych przez nas utworów niż osoby z grupy kontrolnej. Dzięki tym statystykom będzie można określić jak bardzo opłacalny jest nasz model, jednakże wnioski będzie można wyciągnąć dopiero w dłuższej perspektywie czasu. Sukcesem będzie możliwe większa ilość czasu spędzonego w serwisie i ilość odsłuchań poleconych utworów w grupie badawczej względem grupy kontrolnej.

Analityczne kryterium sukcesu

Po wytrenowaniu modelu na części danych zebranych do początku 2024 roku wygenerujemy predykcję które utwory będą dalej popularne. Następnie porównamy te predykcje z danymi z kolejnych miesięcy. Celem będzie działanie lepsze niż losowe, czyli wskazywanie piosenek otrzymujących więcej lików i odtworzeń niż mediana wyników wszystkich piosenek.

Analiza danych

Pierwsze dane otrzymane od klienta były złej jakości.
Wiele danych zawierało błędy – 27% sessions, 10% artists, 17%tracks.
Ilość otrzymanych danych była znacząco zbyt mała.
Nie mieliśmy żadnych danych użytkowników.

Po kilku rozmowach z klientem udało się uzyskać znacznie lepsze dane które są wystarczające do stworzenia modelu.

Dane składają się z:
5000 użytkowników
22412 utworów
1667 artystów
174121419 sesji z czego play-93081043, like-30449874, skip-26885983

1. Znaleźliśmy 8 błędów w danych utworów (dodatnia głośność i tempo równe zero) ale poza tym dane wydają się być prawidłowe.
2. Część piosenek nie ma żadnych lików ani odtworzeń, często mają też popularity równe zero. Pokazuje to, że dużo piosenek nie jest słuchanych.
3. Informacje o gatunku piosenek zawarte są tylko w danych o jej wykonawcy.
4. Daty wydania piosenek mają różne formaty dlatego postanowiliśmy je zgeneralizować tylko do roku.
5. Dane z sesji inne niż play, like i skip nie są istotne dla naszego projektu więc nie bierzemy ich pod uwagę.

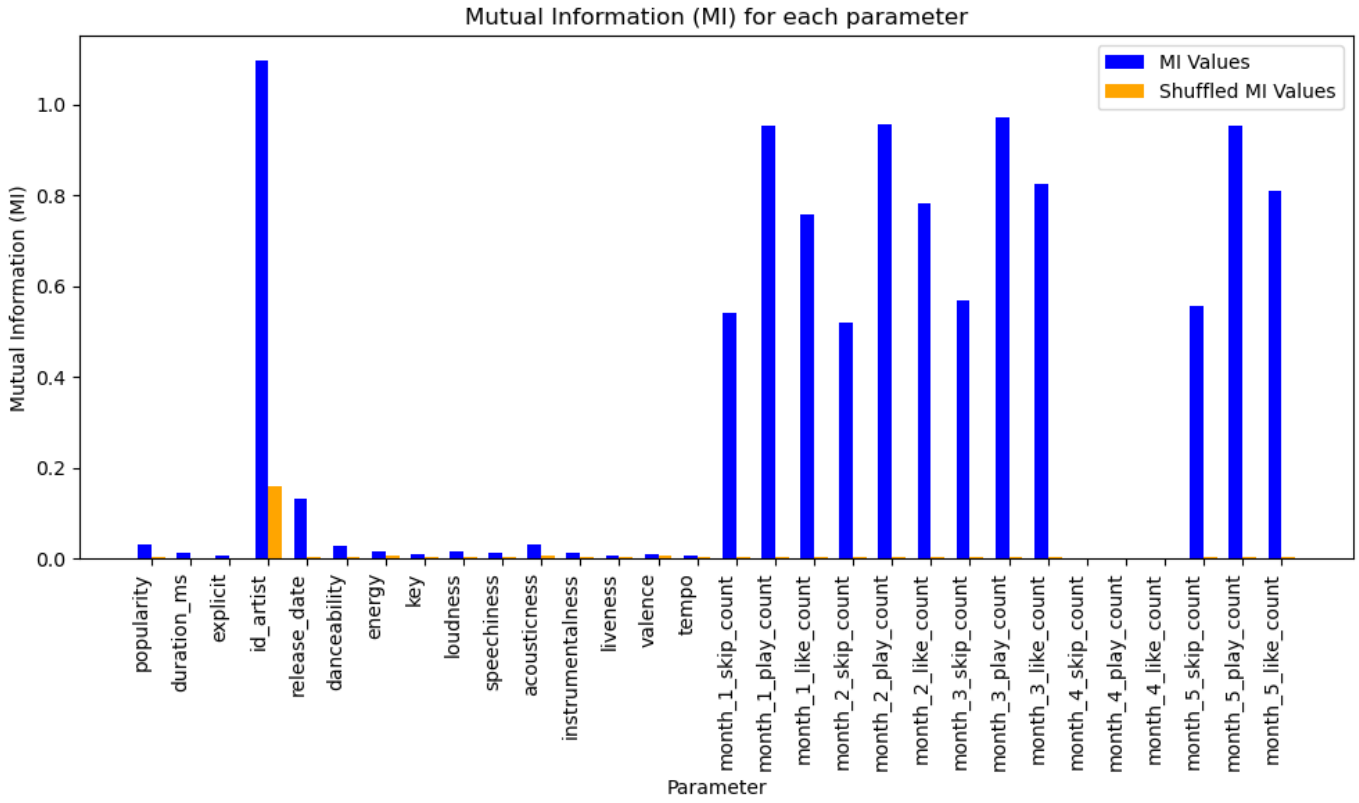
Wstępne założenia dotyczące treningu

Planujemy oceniać trafność predykcji poprzez analizę ilości lików i odtworzeni jakie otrzymuje utwór w następnym miesiącu po wytrenowaniu modelu.
y = ilość odtworzeń + ilość lików * 3
Są to dwa parametry określające na ile piosenka podoba się użytkownikom. Ilość lików jest mnożona przez trzy ponieważ jest ich średnio ok. trzy razy mniej niż odtworzeń, a zależy nam aby oba miały równy wpływ na ocenę popularności.

Wstępny zbiór danych treningowych zawiera wszystkie dane techniczne o utworach oprócz nazwy utworu (która nie ma wpływu na jego popularność z perspektywy modelu). Oprócz tego dodaliśmy informacje o ilości lików, odtworzeni i skipów utworu z ostatnich 5 miesięcy.

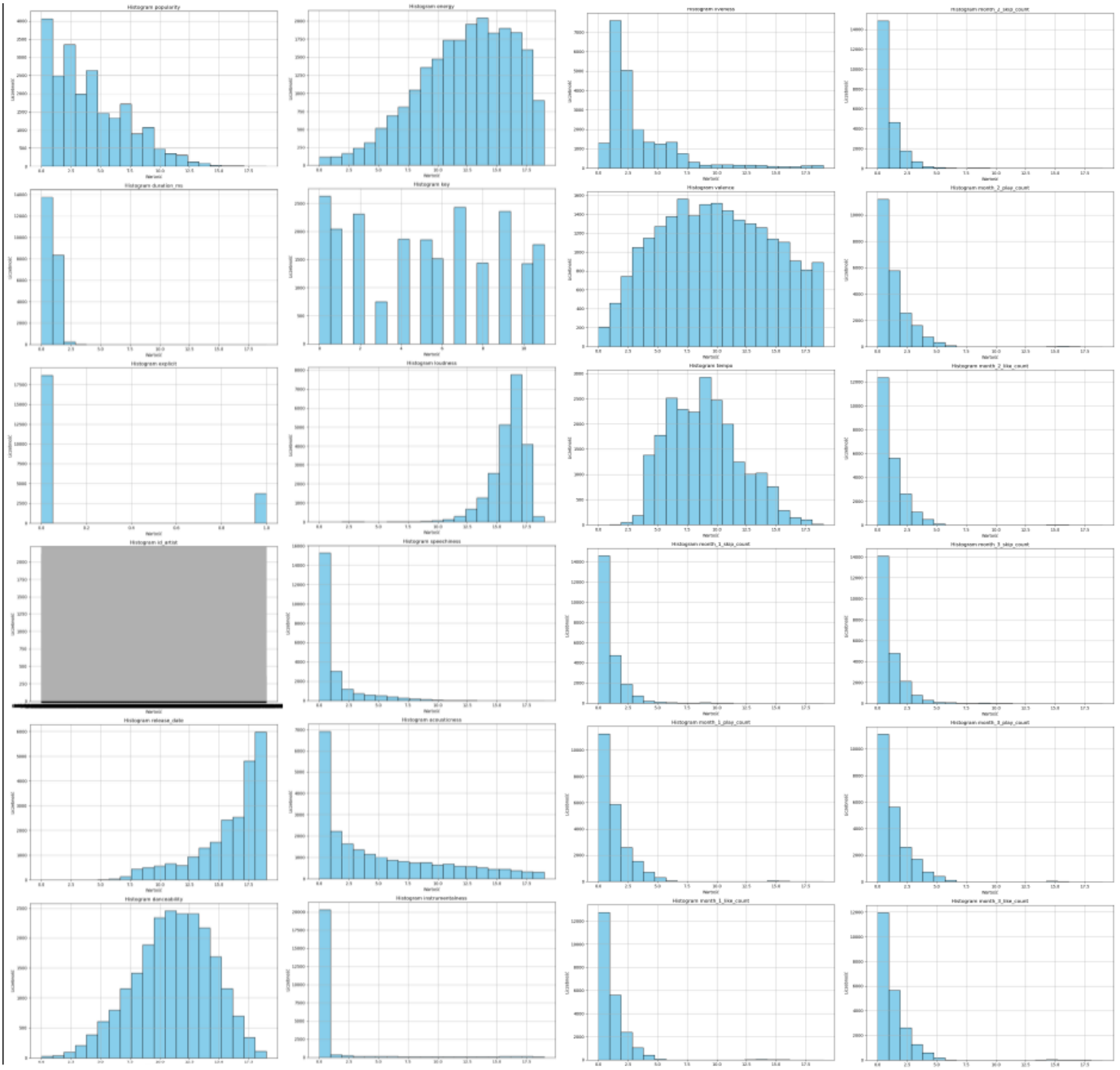
Żeby sprawdzić czy posiadane dane są wystarczające do wytrenowania modelu policzyliśmy współczynnik MI dla każdej z nich i funkcji popularności y. Dane które są ciągłe (większość) zdyskretyzowaliśmy na 20 przedziałów.

W celu porównania wyników policzyliśmy też MI dla losowo zmieszanych danych i wyniki przedstawiliśmy na wykresie.



Wartości MI wahają się od 0 do 1.096, a dla losowo zmieszanych danych nie przekraczają 0,16, a zwykle są bliskie zeru. Duża część danych wydaje się być informatywna.
Na odstawie tych wyników stwierdzamy, że posiadane dane są wystarczające do stworzenia modelu.

Wykresy rozkładów zmiennych wyglądają mniej więcej tak jak się spodziewaliśmy. Niekiedy pojawiają się wartości odstające ale są one zgodne z oczekiwaniami.


Histogram popularity
Histogram energy
Histogram liveness
Histogram month_2_skip_count
Histogram duration_ms
Histogram key
Histogram valence
Histogram month_2_play_count
Histogram explicit
Histogram loudness
Histogram tempo
Histogram month_2_like_count
Histogram id_artist
Histogram speechiness
Histogram month_1_skip_count
Histogram month_3_skip_count
Histogram release_date
Histogram acousticness
Histogram month_1_play_count
Histogram month_3_play_count
Histogram danceability
Histogram instrumentalness
Histogram month_1_like_count
Histogram month_3_like_count