IUM - Etap 1

Jakub Kieruczenko 318669
Karol Ziarek 318749

1. Temat

“Mamy pomysł na świetne narzędzie dla naszych użytkowników – zacznijmy generować listę top utworów, ale z wyprzedzeniem tygodniowym! Będziemy szybsi od konkurencji i zwiększymy zaangażowanie.”

2. Kontekst

W ramach projektu wcielamy się w rolę analityka pracującego dla portalu „Pozytywka” – serwisu muzycznego, który swoim użytkownikom pozwala na odtwarzanie ulubionych utworów online. Praca na tym stanowisku nie jest łatwa – zadanie dostajemy w formie enigmatycznego opisu i to do nas należy doprecyzowanie szczegółów tak, aby dało się je zrealizować. To oczywiście wymaga zrozumienia problemu, przeanalizowania danych, czasami negocjacji z szefostwem. Same modele musimy skonstruować tak, aby gotowe były do wdrożenia produkcyjnego – pamiętając, że w przyszłości będą pojawiać się kolejne ich wersje, z którymi będziemy eksperymentować.

Pozytywka zbiera dane dotyczące swojej działalności – są to:
• lista dostępnych artystów i utworów muzycznych
• baza użytkowników
• historia sesji użytkowników
• techniczne informacje dot. poziomu cache dla poszczególnych utworów

3. Biznesowa definicja problemu

a) Cel biznesowy

Wygenerowanie listy top utworów z wyprzedzeniem tygodniowym, tak by wyprzedzić konkurencję i zwiększyć zaangażowanie. Zadanie do zrealizowania poprzez identyfikację cech utworów, które powiązane są z ich słuchalnością, a na ich podstawie przewidzenie przyszłej wartości słuchalności dla utworów. Ma to skutkować zwiększeniem zaangażowania użytkowników portalu.

Zaangażowanie - tygodniowa ilość odsłuchań w serwisie. Celem systemu jest osiągnięcie satysfakcjonującego przyrostu zaangażowania.

Słuchalność - współczynnik popularności utworu wyznaczany przy użyciu tygodniowego przyrostu polubień i odtworzeń, oraz współczynnika mówiącego o tym, jaka część utworu została średnio przesłuchana.

b) Biznesowe kryteria sukcesu

Poprawiamy istniejące podejście (niekoniecznie stosujące UM)?
Nie wiemy nic konkretnego o istniejącym podejściu. W tabeli tracks znajduje się atrybut *popularity,* którego nazwa wprost sugeruje popularność, nie znamy jednak jego konkretnego przeznaczenia. Być może był poprzednim wskaźnikiem kwalifikacji utworów do toplisty, lecz równie dobrze może być on cechą utworu - “rozrywkowością”. Nie nadaje się jednak w naszym przypadku do użycia jako wartość determinująca dla top listy utworów, bo jej wartość jest dla każdego utworu niezmienna w czasie.

Spełniamy jasno sprecyzowane wymaganie klienta?
Zwiększenie zaangażowania rozumiemy jako zwiększenie przyrostu łącznej liczby odtworzeń w serwisie w porównaniu z konkurencją. Aby je uzyskać, klient chce utworzyć top listę z wyprzedzeniem tygodniowym poprzez zidentyfikowanie cech utworów, które korelują z ich wysoką słuchalnością w danym tygodniu i przewidzenie przyszłej słuchalności.

Staramy się przewyższyć konkurencję/standardy branżowe?
Tak, jest to realizowane poprzez wygenerowanie top listy z wyprzedzeniem tygodniowym. Pozostałych metryk, takich jak np. zaangażowanie, nie jesteśmy w stanie porównywać z konkurencją, bo nie mamy dostępu do ich danych z tego zakresu.

Kryteria biznesowe sukcesu:

- zwiększenie zaangażowania o co najmniej 5% w porównaniu do średniego zaangażowania przez ostatni rok

Metoda pomiaru kryteriów sukcesu:

- Testy A/B
  Wstępny pomiar kryterium sukcesu można przeprowadzić poprzez podzielenie użytkowników do dwóch grup: eksperymentalnej i kontrolnej. Użytkownicy nie mają świadomości brania udziału w testach, wiedzy o obecnej grupie i możliwości jej zmiany.
  Grupa kontrolna ma dostęp do dotychczasowej top listy, a eksperymentalna - do przewidzianej z użyciem naszego modelu. Możemy wtedy porównać ilość odsłuchań generowanych przez użytkowników w każdej grupie, by określić czy nowe rozwiązanie jest skuteczne.

Faktyczny pomiar skuteczności i spełnienia kryterium można uzyskać jedynie poprzez wdrożenie modelu i zmierzenie realnych wyników.

4. Definicja zadania modelowania

a) Zadanie

Obliczenie słuchalności każdego utworu oraz przewidzenie tej wartości dla następnego tygodnia. Top Lista wyznaczana jest jako lista 20 utworów o największej wartości słuchalności w danym tygodniu.
W wersji podstawowej skorzystamy z modelu regresji liniowej, który jest względnie prosty i szybki do zaimplementowania. W późniejszych fazach będziemy wykorzystywali bardziej skomplikowane metody przewidywania szeregów czasowych by wybrać ten o najlepszym działaniu według analitycznych kryteriów sukcesu. Obiecujący wydaje się przede wszystkim model ARIMA, natomiast dokładnie zweryfikują to testy.

b) Dane

Lista utworów muzycznych (tracks):

- id - unikalny identyfikator utworu
- name - nazwa utworu
- popularity - popularność utworu
- duration_ms - czas trwania utworu
- explicit - informacja o zawieraniu nieodpowiednich (np. dla dzieci) treści
- id_artist - identyfikator artysty utworu
- release_date - data wydania utworu
- danceability - taneczność utworu
- energy - energia, dynamiczność utworu
- key - tonacja utworu
- loudness - głośność utworu
- speechiness - ilość wokalu w utworze
- acousticness - wskaźnik ilości elementów akustycznych utworu
- instrumentalness - wskaźnik koncertowości utworu
- valence - wskaźnik emocjonalności utworu
- tempo - tempo utworu

Historia sesji (sessions):

- session_id - unikalny identyfikator utworu
- timestamp - czas rozpoczęcia sesji
- user_id - identyfikator użytkownika w danej sesji
- track_id - identyfikator utworu, którego dotyczy sesja
- event_type - typ zdarzenia w danej sesji

Zdecydowaliśmy się nie korzystać z trzech pozostałych tabel.

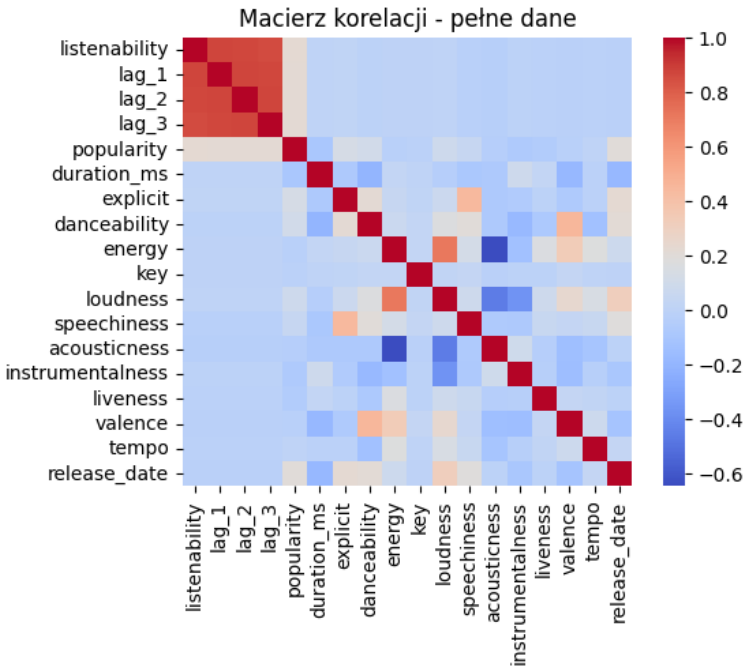
Tworzymy wektory cech dla każdego utworu przez agregację danych.

Wektor wejściowy:
Zawierał będzie dane w formacie: [track_id][week][features], gdzie features to zmienne zawierające informacje o utworze, wstępnie wybrane atrybuty to:

- duration_ms - czas trwania utworu
- explicit - informacja o zawieraniu nieodpowiednich (np. dla dzieci) treści
- release_date - rok wydania utworu
- danceability - taneczność utworu
- energy - energia, dynamiczność utworu
- key - tonacja utworu
- loudness - głośność utworu
- speechiness - ilość wokalu w utworze
- acousticness - wskaźnik ilości elementów akustycznych utworu
- instrumentalness - wskaźnik koncertowości utworu
- valence - wskaźnik emocjonalności utworu
- tempo - tempo utworu
- listenability_lag1 - wartość słuchalności 1 tydzień wstecz
- listenability_lag2 - wartość słuchalności 2 tygodnie wstecz
- listenability_lag3 - wartość słuchalności 3 tygodnie wstecz

Wyszczególnione w wektorze atrybuty są dodane z założeniem, że niosą informację o zmiennej celu, natomiast po przeanalizowaniu obecnego zbioru danych możemy stwierdzić, że wszystkie podkreślone atrybuty nie niosą istotnej informacji o zmiennej celu - wartość bezwzględna korelacji ze zmienną celu żadnego z tych atrybutów nie przekracza 4%. Jedyne atrybuty, które niosą istotne informacje o zmiennej celu to jej historyczne wartości w poprzednich tygodniach (korelacja kolejnych lagów równa ~85%).
Z tego powodu w testach podstawowego modelu używaliśmy wektora wejściowego jedynie z wartościami historycznymi zmiennej, taki model osiągał słabe wyniki.

Testowana była również korelacja atrybutów piosenek z uwzględnieniem różnego podziału danych: bez piosenek o zerowej liczbie powiązanych sesji, dane sumaryczne, czy z wyłączeniem konkretnych tygodni. Również te testy wykazują niską korelację atrybutów statycznych piosenek ze zmienną celu. Dane w obecnej formie wydają się być niewystarczające do zbudowania modelu o jakości satysfakcjonującej klienta.



Zmienna celu:
“Słuchalność” - zmienna wyliczana dla każdego utworu w danym tygodniu, jako: $S = Lo * Pr + Lp$
Gdzie:
S - słuchalność, Lo - liczba odtworzeń, Lp - liczba polubień
Pr - współczynnik średniej długości słuchania utworu (0.0 - 1.0)

c) Kryteria sukcesu

MSE - Mean Square Error: wartość pomocnicza określająca skuteczność modelu w dokonywaniu predykcji słuchalności. Nie mówi jednak ona o tym czy utworzona top lista składa się z pożądanych utworów. Im mniejsze MSE, tym lepsze są predykcje modelu.
Za satysfakcjonujący wynik uznamy MSE <= 2. (Wyniku testowego modelu regresji.)

Do określania skuteczności wyznaczania top listy planujemy wykorzystać miary rankingujące takie jak nDCG@K. Z przewidzianych wartości wybieramy 20 najlepszych sortując je, a następnie wyznaczymy nDCG@20 w celu zmierzenia trafności przewidzianej top listy. Za sukces uznajemy osiągnięcie wartości >= 0,8.