Inżynieria Uczenia Maszynowego

Dokumentacja Etap 2

Klasyfikacja użytkowników pod kątem skłonności do zakupu konta premium

Marcin Bagnowski, Kacper Kowalczyk

16 stycznia 2025
Semestr 24Z

Spis treści

1 Repozytorium kodu

Stworzony kod na potrzebę projektu, znajduje się w repozytorium pod linkiem https://gitlab-stud.elka.pw.edu.pl/kkowalcz/ium-24z.git

2 Analiza zadania

2.1 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.

Jak każda szanująca się firma internetowa, Pozytywka zbiera dane dotyczące swojej działalności – są to (analitycy mogą wnioskować o dostęp do tych informacji na potrzeby realizacji zadania):

- lista dostępnych artystów i utworów muzycznych
- baza użytkowników
- historia sesji użytkowników
- techniczne informacje dot. poziomu cache dla poszczególnych utworów

2.2 Definicja zadania biznesowego

Potrzeba biznesowa została wyrażona następującym sformułowaniem:

> "Jakiś czas temu wprowadziliśmy konta premium, które uwalniają użytkowników od słuchania reklam. Nie są one jednak jeszcze zbyt popularne – czy możemy się dowiedzieć, które osoby są bardziej skłonne do zakupu takiego konta?"

Interpretujemy ją jako potrzebę zachęcenia użytkowników do korzystania z kont premium. Zakładamy jednak, że ci są już zachęcani poprzez banery, czy klipy puszczane w aplikacji. Wówczas model miałby za zadanie ustalić, co przekonałoby daną osobę do zakupu konta premium. Jednak w obecnej wersji zadanie jest niemożliwe do zrealizowania, ponieważ konto premium przynosi tylko jedną korzyść - usunięcie reklam. To eliminuje możliwość profilowania korzyści pod konkretnych użytkowników, aby techniki zachęcające były dopasowane do ich preferencji.

Dlatego zdecydowaliśmy się skupić na dosłownym ustalaniu, którzy użytkownicy są zainteresowani kontem premium. Rozwiąże to inny problem biznesowy, który również przyniesie korzyść firmie - **ustalenie, komu opłaca się okazjonalnie wyświetlać reklamy konta premium zamiast zwykłych reklam, a komu lepiej wyświetlać same reklamy komercyjne.**

2.3 Definicja zadania modelowania

Z perspektywy modelowania najbardziej dopasowanym typem zadania będzie **klasyfikacja binarna** - podział użytkowników na grupy skłonnych i niechętnych do zakupu konta premium.

W przypadku niepowodzenia wcześniej wymienionego podejścia sprawdzimy alternatywne opcje:

- Regresja - utworzenie przedziałów prawdopodobieństwa, że osoba z danej grupy jest w danym stopniu zainteresowana zakupem.
- Grupowanie - podział na rozłączne grupy i ustalenie zainteresowania kontem premium w każdej z grup.

2.4 Definicja założeń

- Model będzie uczony offline na historycznych danych dostarczonych z systemu.
- Dane są historyczne, dlatego mimo długiego okresu nie gwarantują idealnej reprezentacji danych aktualnych oraz przyszłych.
- Wnioski i predykcje wyciągamy jedynie na podstawie danych z systemu to znaczy, że nie bierzemy pod uwagę czynników zewnętrznych mogących wpływać na decyzje użytkownika o zakupie konta premium.
- Użytkownik może raz kupić konto premium i jest ono permanentne. Zakładamy tak, ponieważ w zbiorze interakcji użytkowników z systemem jest akcja identyfikująca zakup konta premium, natomiast nie ma żadnej akcji wskazującej na jego anulowanie/rezygnację (opis zbioru interakcji 3.1).

2.5 Kryteria sukcesu

Jak w większości aplikacji, konta premium są mało popularne. Podstawowym kryterium sukcesu, które przyniesie korzyść, jest uzyskanie lepszej trafności przewidywań niż model wybierający klasę większościową. Użytkownicy bez konta premium stanowią **57%**, natomiast posiadacze **43%**. Dalszym kryterium sukcesu będzie uzyskiwanie coraz lepszej skuteczności, co przełoży się na zwiększenie przychodów firmy, poprzez oszczędniejsze korzystanie z banerów zachęcających do kont premium (przy zachowaniu zbliżonego poziomu sprzedaży) na rzecz reklam.

Również jest ważne to, aby decyzje podejmowane przez model były zrozumiałe i wyjaśnialne (**Explainable AI**). Spróbujemy to osiągnąć, tworząc możliwie prosty model przy zachowaniu wysokiej jakości predykcji. Opiszemy również, z jakich atrybutów składa się model, a dla tych bardziej złożonych, jak są one zbudowane.

3 Analiza danych

3.1 Spis posiadanych danych

Z portalu Pozytywka zostały udostępnione nam następujące dane:

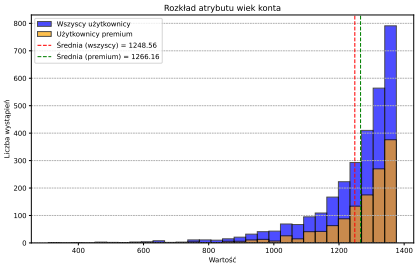
- Zbiór artystów - ponad **27 tys.** przykładów - id, nazwa, lista gatunków muzycznych.
- Zbiór sesji użytkowników - prawie **1,5 mln** przykładów - czas, id użytkownika, id odsłuchiwanego utworu, akcja (puść, pomiń, reklama, kup premium), id sesji.
- Zbiór utworów - prawie **130 tys.** przykładów - id utworu, id artysty, nazwa, popularność, czy wulgarny, data wydania, taneczność, energiczność, klucz, tryb, głośność, mowa, akustyczność, instrumentalność, żywotność, wartościowość, tempo, sygnatura czasowa.
- Zbiór przechowywania utworów - prawie **130 tys.** przykładów - id utworu, klasa przechowywania, koszt przechowywania. Z tego zbioru nie będziemy korzystać.
- Zbiór użytkowników - 3 tys. przykładów - id użytkownika, nazwa, miasto, ulica, lista ulubionych gatunków muzycznych, czy kiedykolwiek kupił konto premium.

3.2 Atrybuty z danych bazowych

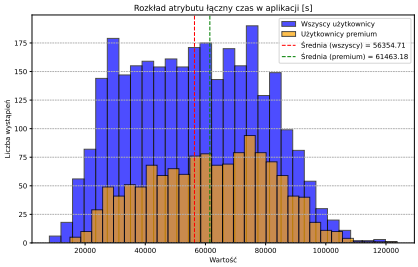
Mamy hipotezę, że następujące informacje mogą być potencjalnie skorelowane z prawdopodobieństwem zakupu konta premium:

3.2.1 Wiek konta

Mierzony jako czas pierwszej interakcji użytkownika z systemem. Rozkład atrybutu pokazuje histogram 1. Zdecydowanie dominują starsi użytkownicy. Również średnia dla osób z kontem premium jest większa. Wiek konta ma rozkład wykładniczy zarówno dla wszystkich użytkowników, jak i tych premium.



Rysunek 1: Rozkład atrybutu - wiek konta w dniach



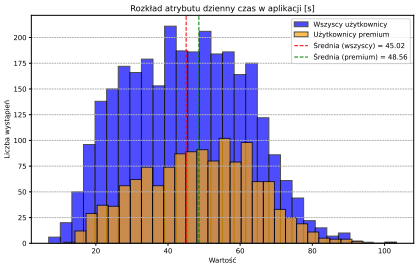
Rysunek 2: Rozkład atrybutu - łączny czas w sekundach

3.2.2 Czas spędzony w aplikacji

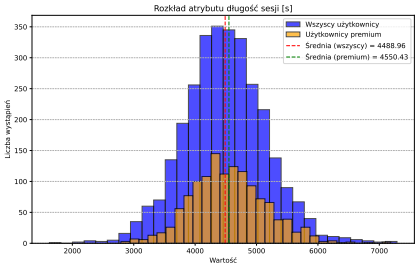
Czas użytkownika w aplikacji liczony na różne sposoby:

- łącznie przez cały okres korzystania - histogram 2
- średnia w ciągu dnia - histogram 3
- średnia w jednej sesji - histogram 4

Widać, że użytkownicy decydujący się na zakup konta premium spędzają więcej czasu w aplikacji, co pokazuje większa średnia dla tych wykresów. Wszystkie te czasy mają charakter rozkładów normalnych.



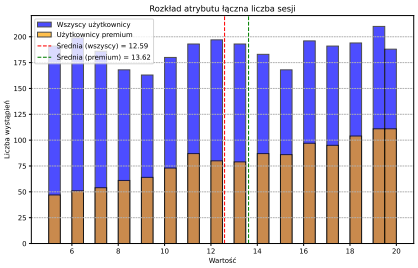
Rysunek 3: Rozkład atrybutu - czas dziennie



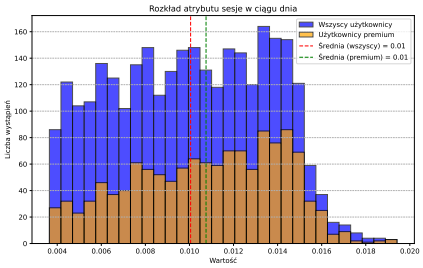
Rysunek 4: Rozkład atrybutu - czas w sesji

3.2.3 Sesje użytkownika

- Łączna liczba sesji użytkownika - histogram 5 - ma rozkład dyskretny cykliczny.
- Średnia liczba sesji w ciągu dnia - histogram 6 - ma rozkład asymetryczny.



Rysunek 5: Rozkład atrybutu - łączna liczba sesji



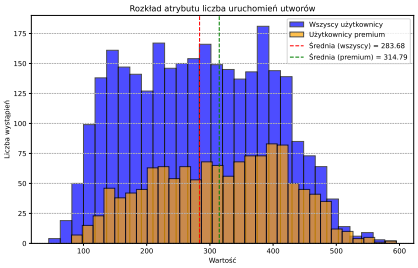
Rysunek 6: Rozkład atrybutu - liczba sesji na dzień

3.2.4 Interakcje użytkownika

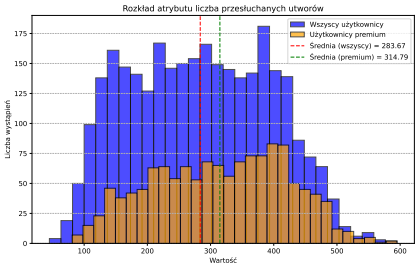
Bierzemy pod uwagę liczbę interakcji użytkownika z systemem, czyli:

- Liczba uruchomionych utworów - histogram 7 - ma rozkład przypominający normalny.
- Liczba przesłuchanych utworów - histogram 8 - ma rozkład przypominający normalny.
- Liczba pominiętych utworów - histogram 9 - histogram ze wszystkimi użytkownikami ma rozkład log-normalny, natomiast dla użytkowników premium ciężko określić konkretny rozkład.
- Liczba polubień utworów - histogram 10 - histogram ze wszystkimi użytkownikami ma rozkład gamma, natomiast dla użytkowników premium ciężko określić konkretny rozkład.

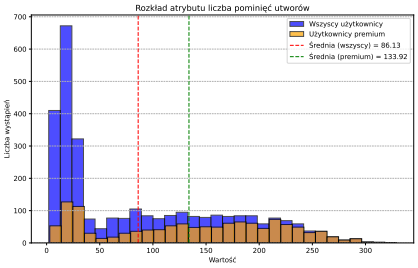
Zdecydowanie widać, że użytkownicy premium dokonują większej liczby interakcji w systemie dla każdej z możliwych akcji. Najbardziej pokazuje to atrybut określający liczbę pominiętych utworów, gdzie wielu użytkowników niebędących premium (różnica między niebieskim i pomarańczowym wykresem) wcale lub rzadko pomija słuchane utwory.



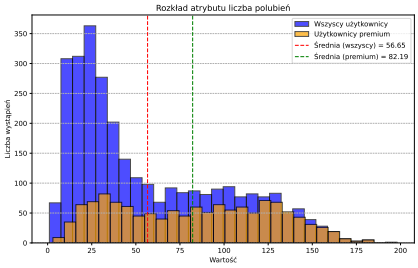
Rysunek 7: Rozkład atrybutu - liczba uruchomionych utworów



Rysunek 8: Rozkład atrybutu - liczba przesłuchanych utworów



Rysunek 9: Rozkład atrybutu - liczba pominiętych utworów



Rysunek 10: Rozkład atrybutu - liczba polubień utworów

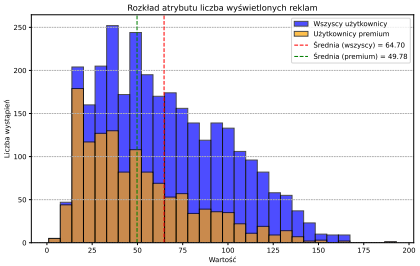
3.2.5 Reklamy

Dla akcji związanych z reklamami chcemy sprawdzić następujące statystyki:

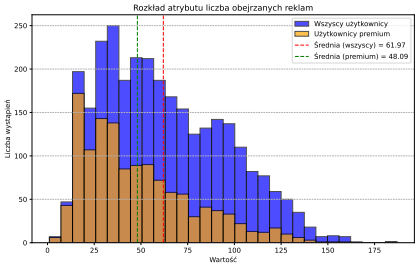
- Liczba wyświetlonych reklam użytkownikowi - histogram 11.
- Liczba obejrzanych reklam (użytkownik nie wyszedł z aplikacji po zobaczeniu reklamy) - histogram 12.

Oba histogramy mają rozkład przypominający rozkład gamma. Użytkownicy premium rzadziej oglądają, a więc również rzadziej pomijają reklamy. Jednak ta statystyka jest trochę zniekształcona, ponieważ została policzona dla całego okresu korzystania z aplikacji. Użytkownicy od momentu wykupienia konta premium przestają widzieć reklamy, co wpływa na całą statystykę.

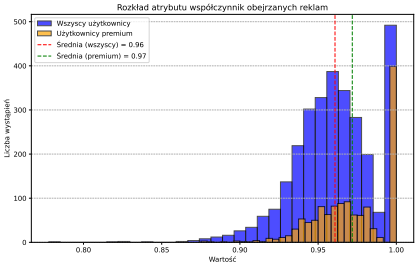
Dla przejrzystości utworzyliśmy współczynnik obejrzanych do wyświetlonych reklam - histogram 13. Tutaj dla obu przypadków (ogólnego i użytkowników premium) widzimy zbliżony rozkład tego współczynnika. Jednak nieproporcjonalnie więcej dla użytkowników premium było osób, które oglądały całe reklamy. Można wywnioskować, że te osoby były niechętne do rozwiązania z zamykaniem aplikacji, aby rozpocząć nową sesję i pozbyć się reklamy, tylko zdecydowały się na zakup konta premium. Osoby, które częściej wychodziły z aplikacji przy wyświetlaniu reklamy, rzadziej dokonywały zakupu konta premium.



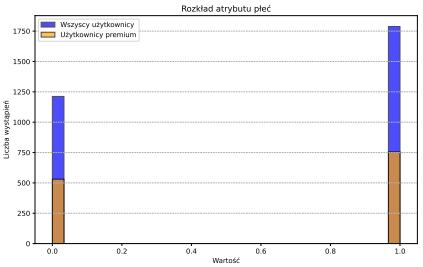
Rysunek 11: Rozkład atrybutu - liczba wyświetlonych reklam



Rysunek 12: Rozkład atrybutu - liczba obejrzanych reklam



Rysunek 13: Rozkład atrybutu - współczynnik obejrzanych reklam



Rysunek 14: Rozkład atrybutu - płeć

3.2.6 Płeć użytkownika

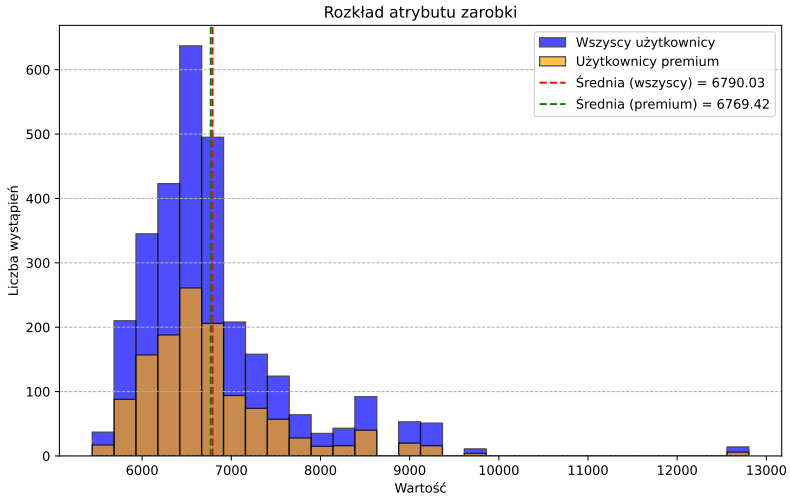
Podstawowe dane osobowe użytkownika, takie jak płeć (kobieta - imię kończy się na literę "a", mężczyzna - przeciwny przypadek), mogą mieć pośredni wpływ na kupowanie konta premium. Histogram 14 prezentuje rozkład ze względu na płeć i widzimy przewagę mężczyzn (wartość 1) względem liczby kobiet (wartość 0). Wykres ma rozkład Bernoulliego. Co ciekawe, różnica ta jest znacznie mniejsza dla użytkowników premium, wygląda na to, że kobiety częściej decydują się na zakup konta premium.

3.2.7 Zarobki użytkownika

Miejsce zamieszkania pozwala na średnie oszacowanie zarobków danej osoby. Przypuszczamy, że zamożniejsze osoby mogą mieć mniej oporów przed dokonaniem zakupu. Do oszacowania zarobków przyjęliśmy średnią pensję z danej miejscowości na podstawie danych z GUS. Dane pobraliśmy za pomocą API GUS link. Rozkład tego atrybutu przypomina rozkład Gamma i prezentuje go histogram 15. Z podobieństwa obu rozkładów widać, że użytkownicy premium nie zarabiają więcej niż ci bez konta premium.

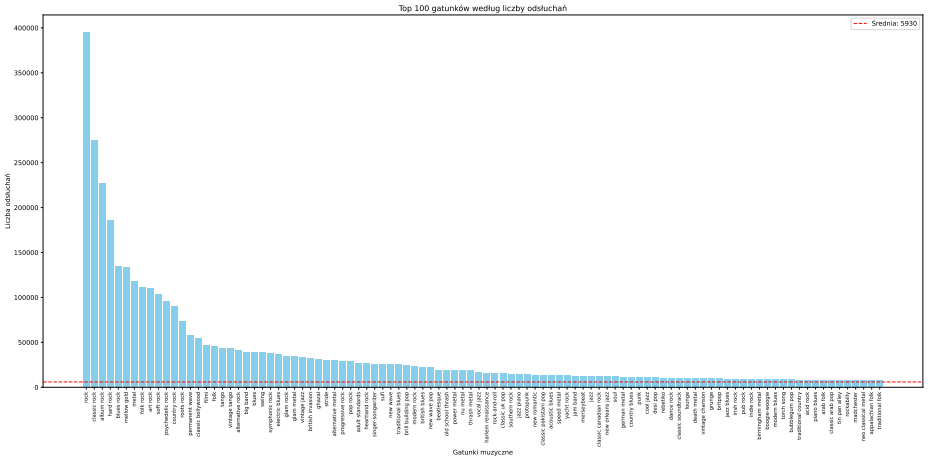
3.2.8 Gatunek słuchanej muzyki

Osobowość danego użytkownika może wpływać na preferencje muzyczne i skłonność do zakupu konta premium. Łącznie w systemie jest 3953 różnych gatunków muzycznych. Z racji tak dużej liczby sprawdziliśmy, jak często użytkownicy słuchają danego gatunku. Rysunek 16 prezentuje liczbę przesłuchań dla każdego z 100 najczęściej słuchanych gatunków. Wykres ma rozkład wykładniczy, co informuje nas, że wiele dalszych gatunków z poza wykresu jest bardzo rzadko słucha-



Rysunek 15: Rozkład atrybutu - zarobki

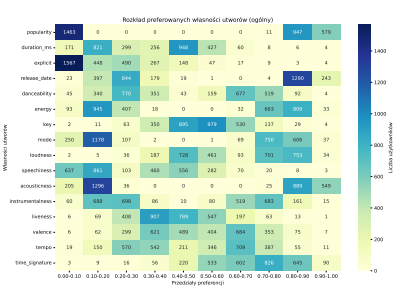
nych. Z tego powodu, że gatunków jest tak sporo, a dysproporcje między nimi są znaczące, nie będziemy brali tego atrybutu pod uwagę. Rozważaliśmy również złączenie mało popularnych gatunków w jeden pod nazwą "pozostałe gatunki"jednak to jeszcze bardziej zaburzyłoby rozkład gatunków. Dodatkowym argumentem za rezygnacją z tego atrybutu jest obecność atrybutu popularity bezpośrednio przy słuchanym utworze, co już daje nam informacje, czy użytkownik słucha popularnej muzyki bez potrzeby sprawdzania popularności samego gatunku.



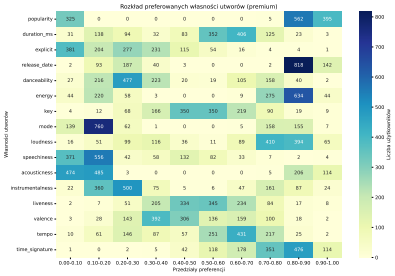
Rysunek 16: Rozkład atrybutu - Liczba odsłuchań gatunku

3.2.9 Własności słuchanych utworów

Zbadamy, czy własności utworów takie jak ich popularność, data wydania, wulgarność czy energiczność mają wpływ. Rozkład tych własności prezentują dwie heatmapy: jedna dla wszystkich użytkowników 17, druga dla użytkowników premium 18. Z porównania wykresów wynika, że użytkownicy premium częściej słuchają utworów popularnych, dłuższych, nowszych, mało tanecznych, energicznych i z niskim poziomem akustyczności i instrumentalności. W dalszej części dokładniej zbadamy te zależności wykorzystując macierz korelacji.



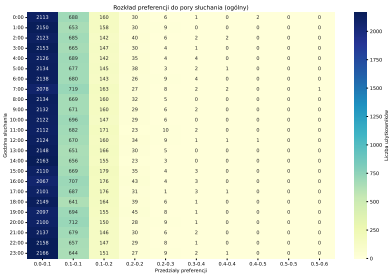
Rysunek 17: Rozkład atrybutu - własności utworów (ogólna)



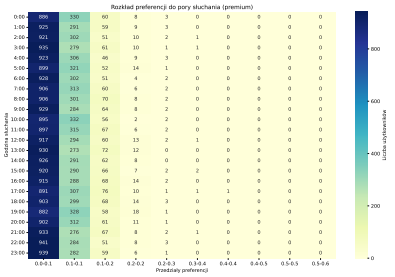
Rysunek 18: Rozkład atrybutu - własności utworów (użytkownicy premium)

3.2.10 Pora słuchania

Odtwarzanie muzyki na imprezach (czyli intensywnie w godzinach nocnych) byłoby wygodniejsze bez reklam. Rozkład preferencji słuchania prezentują dwie heatmapy: jedna dla wszystkich użytkowników 19, druga dla użytkowników premium 20. Z obu heatmap widać, że o każdej godzinie są użytkownicy, którzy korzystają z aplikacji. Liczba użytkowników jest rozłożona równomiernie i ciężko wskazać trend, kiedy więcej użytkowników korzysta z aplikacji, biorąc pod uwagę zarówno wszystkich, jak i tylko użytkowników premium. Wartość preferencji występuje w zakresie [0, 1] natomiast nie zamieszczaliśmy kolumn dla wartości powyżej 0,6 z racji braku takich wartości.



Rysunek 19: Rozkład atrybutu - pora słuchania (ogólna)



Rysunek 20: Rozkład atrybutu - pora słuchania (użytkownicy premium)

3.2.11 Potencjalne (nieutworzone) atrybuty

Przed otrzymaniem ostatecznej wersji danych z systemu myśleliśmy nad utworzeniem następujących atrybutów, jednak z racji braku danych musimy z nich zrezygnować:

- Wiek użytkownika.
- Model urządzenia - wskazuje na zamożność danej osoby.
- Pomijanie reklam (jeżeli jest).
- Dla osób, które użyły linku referencyjnego (jeżeli jest w aplikacji): Czy polecający użytkownik ma konto premium? - Mógłby je również polecić.

- Zainteresowanie okresem próbnym (jeżeli jest) - wskazuje na zainteresowanie kontem premium.
- Dokonanie innych zakupów w aplikacji (jeżeli są) - wskazują, że użytkownik jest skłonny wydać pieniądze w aplikacji.

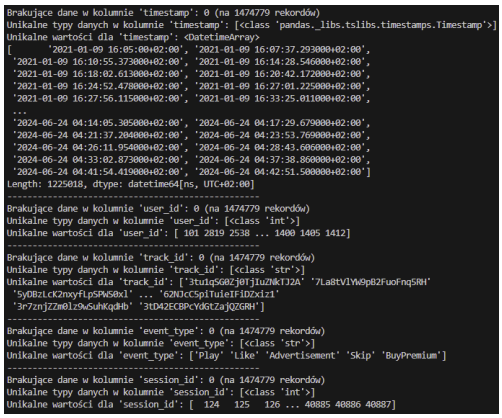
Łącznie dysponujemy odpowiednio dużą liczbą danych, obejmującą większość informacji, jakie są konieczne do wytrenowania modelu o zadowalającej jakości.

3.3 Dostosowanie danych

Dane nieliczbowe dwustanowe np. płeć przeniesiemy do przestrzeni liczbowej kodowaniem binarnym. Datę zamienimy na wartość liczbową mierzoną w sekundach lub dniach. Następnie do wszystkich danych zastosujemy normalizację do zakresu [0, 1], bądź standaryzację. Poprawi to dostosowanie danych do modelu.

3.4 Braki i błędne wartości w danych

Przeanalizowaliśmy otrzymane dane pod kątem potencjalnych braków w danych oraz błędnych wartości. W tym celu sprawdziliśmy bezpośrednie występowanie wartości null oraz jakie są poszczególne unikalne wartości, typy dla danych dla każdego zbioru. Zdjęcie 21 prezentuje wyniki dla zbioru sesji, zdjęcia 22 i 23 dla użytkowników, zdjęcie 24 dla zbioru artystów, a 25 i 26 dla zbioru utworów. Wszystkie atrybuty są kompletne i mają prawidłowe wartości, z wyjątkiem atrybutu "mode"ze zbioru utworów. Ten jest mocno wybrakowany i oprócz wartości 0, 1, posiada wiele wartości typu null. Z tego powodu zdecydowaliśmy się na usunięcie tego atrybutu.



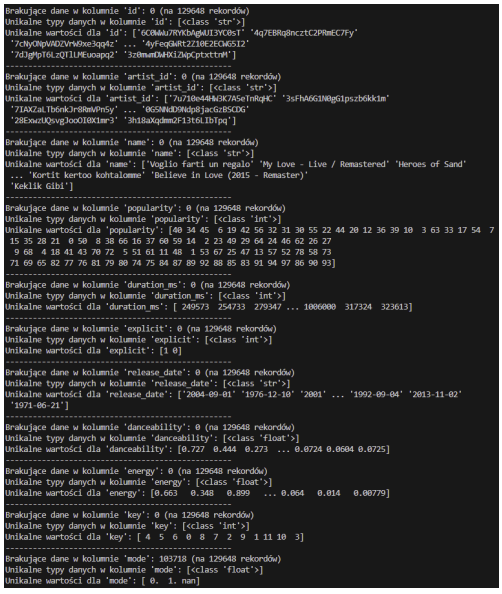
Rysunek 21: Braki i typy wartości dla zbioru sesji

Rysunek 22: Braki i typy wartości dla zbioru użytkowników cz. 1

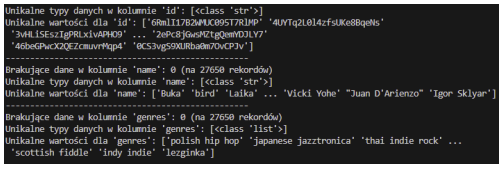
3.5 Korelacja danych

Do zbadania korelacji atrybutów z wartością przewidywaną (czy użytkownik kupi konto premium) oraz korelacji między atrybutami wykorzystaliśmy macierz

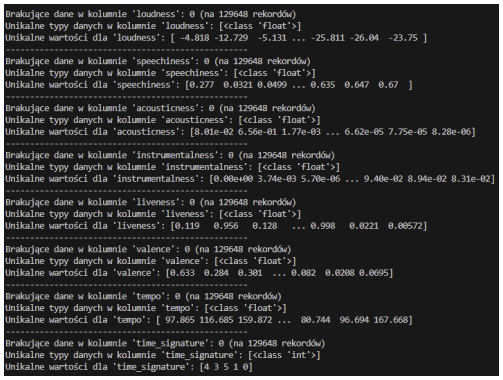
Rysunek 23: Braki i typy wartości dla zbioru użytkowników cz. 2



Rysunek 25: Braki i typy wartości dla zbioru utworów cz. 1



Rysunek 24: Braki i typy wartości dla zbioru artystów

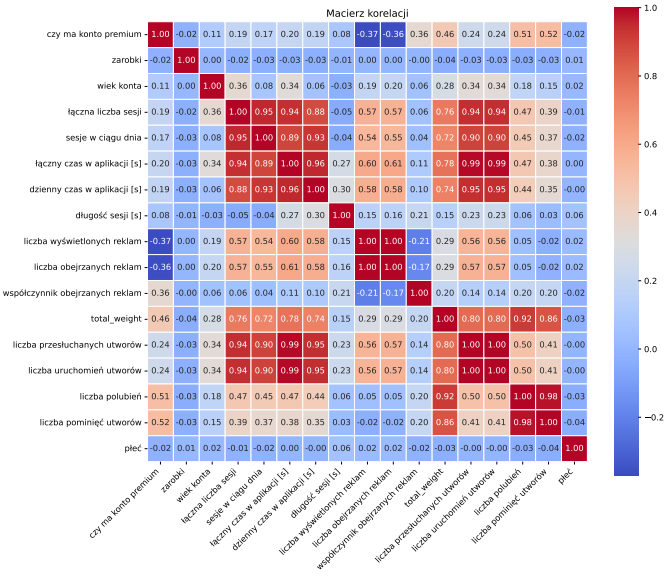


Rysunek 26: Braki i typy wartości dla zbioru utworów cz. 2

kowariancji. Jej wyniki prezentuje rysunek 27. Wiele z wcześniej wymienionych atrybutów jest słabo skorelowana z wartością, którą chcemy przewidywać - czy użytkownik kupi konto premium. Najmocniej jest skorelowana z atrybutami dotyczącymi interakcji użytkownika z systemem (3.2.4) oraz liczbą obejrzanych reklam. Atrybuty opisujące czas spędzony w systemie przynoszą już słabszą informację. Zaskakujący może być również fakt braku korelacji zarobków użytkownika. Brak pojedynczych, silnie skorelowanych atrybutów z wartością przewidywaną pokazuje, że problem przewidywania nie należy do łatwych. Może to spowodować, że nasz model w przypadku braku satysfakcjonujących wyników będzie wymagał dostarczenia nowych atrybutów.

Co istotne, chcąc uprościć nasz model, moglibyśmy zrezygnować z kilku atrybutów silnie skorelowanych między sobą, jako że nie wnoszą dodatkowej informacji do modelu, a jedynie go komplikują. Takie parametry to między innymi:

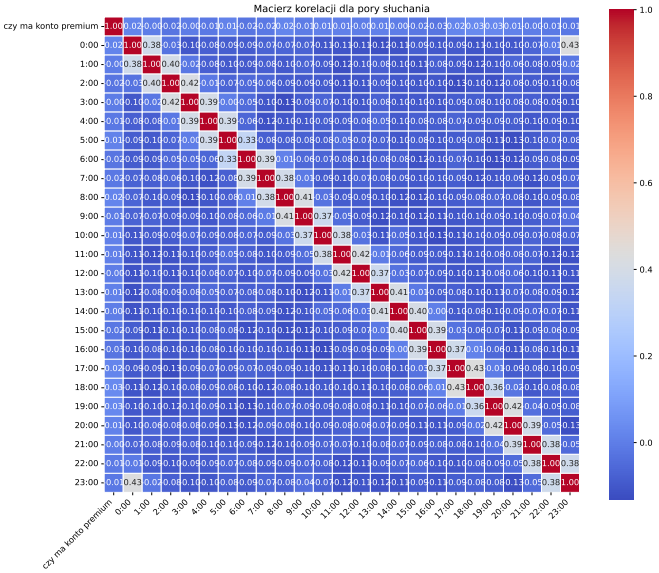
- liczba przesłuchanych utworów - liczba uruchomień utworów - korelacja 1.00
- liczba obejrzanych reklam - liczba wyświetlonych reklam - korelacja 1.0
- łączny czas w aplikacji - liczba przesłuchanych utworów - korelacja 0.99
- liczba polubień utworów - liczba pominięć utworów - korelacja 0.98
- łączny czas w aplikacji - dzienny czas w aplikacji - korelacja 0.96



Rysunek 27: Ogólna macierz korelacji

- łączna liczba sesji - sesje w ciągu dnia - korelacja 0.95
- i parę więcej...

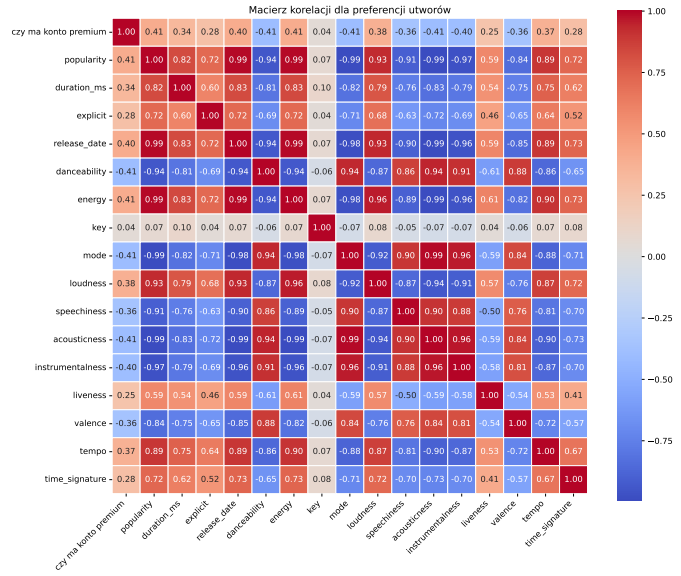
Dla przejrzystości dla złożonych atrybutów, czyli preferencji pory słuchania i preferencji właściwości utworów, stworzyliśmy osobne macierze korelacji. Rysunek 28 prezentuje macierz dla pory słuchania. Z macierzy jasno wynika, że żadna z godzin słuchania nie jest powiązana z chęcią do zakupu konta premium. Z tego powodu powinniśmy nie uwzględniać preferencji pory słuchania w naszym modelu.



Rysunek 28: Macierz korelacji dla pory słuchania

Rysunek 29 prezentuje macierz własności utworów. Niemal wszystkie właściwości utworów są skorelowane na poziomie około 0,4. To oznacza, że wiele z nich niesie informacje o predykcji i warto je uwzględnić w modelu. Wyjątkiem jest atrybut "key", który jest zupełnie nieskorelowany z atrybutem premium oraz z

żadnym innym. Dodatkowo warto zwrócić uwagę na korelacje między właściwościami jak np. energy i release_date na poziomie 0,99. Takie atrybuty powielają się i można z nich zrezygnować utraty informacji.



Rysunek 29: Macierz korelacji dla własności utworów

To pokazuje, że nasz obecny zestaw atrybutów może zostać znacząco zmniejszony bez straty niesionej informacji.

3.6 Podział danych

Dostarczone dane nie są jasno podzielone na zbiór treningowy i testowy. Zdecydowaliśmy się dokonać podziału według liczby użytkowników, tak aby w zbiorze treningowym znalazło się 80% użytkowników, a w testowym 20%. Do podziału wykorzystamy technikę Consistent Hashing modulo 5 według id użytkownika (co piąty użytkownik trafi do zbioru testowego), która zapewni losowość przy podziale na zbiory. Z podziałem zbioru użytkowników wiąże się odpowiedni podział zbioru ich sesji (interakcji z systemem) - sesje danego użytkownika trafią do tego samego zbioru co ten użytkownik.

Pozostałe zbiory danych: artystów, utworów, przechowywania utworów są wspólne i nie wymagają podziału. Dodatkowo zastosujemy walidację krzyżową ze współczynnikiem k=5 lub k=10 (decyzję podejmiemy znając czas trenowania modelu) podczas trenowania modelu. To pozwoli na lepszą ocenę dobranych wartości hiperparametrów do modelu.

4 Sposoby modelowania danych

Dostarczone dane dotyczące interakcji użytkownika z systemem są oznaczone datą - znacznikiem czasowym. To wprowadza możliwość analizy tych danych jako szeregów czasowych. W związku z tym chcielibyśmy sprawdzić, czy dodanie tej informacji do modelu wpłynie na jakość jego predykcji. Dlatego nasz model wytrenujemy w dwóch wersjach, podając inne przekształcenie danych - z i bez uwzględnienia szeregów czasowych.

4.1 Model bez szeregów czasowych

Przy przekształcaniu danych bazowych na atrybuty zamierzamy przetestować wersję bez uwzględniania zmian atrybutów w czasie. Dane, które są zmienne w czasie np. czas od ostatniego logowania, policzymy jako średnią dla tego użytkownika. Atrybuty określające dane stałe w czasie takie jak płeć lub ulubione gatunki muzyczne, pozostaną bez zmian. Wprowadza to założenie, że zachowanie użytkownika związane ze skłonnością do zakupu konta premium jest w dłuższej perspektywie stałe.

4.2 Model z szeregami czasowymi

Dane niezmienne w czasie będą wprowadzone do modelu bez zmian. Natomiast z danych zmiennych spróbujemy utworzyć trendy użytkownika w następujący sposób:

1. Dane zostaną podzielone na mniejsze odcinki czasowe - jeden miesiąc.
2. Dla tych mniejszych okresów zostanie wyciągnięta średnia.
3. Następnie policzymy różnice między tymi okresami jak zmieniała się wartość.
4. Wartość trendu dla wybranego okresu wyznaczymy mnożąc ze sobą różnicę między końcem, a początkiem okresu z liczbą okresów. To znacząco wzmocni wartość trendów, które utrzymują się dłużej.

Trend będziemy wyznaczać na okres do momentu zakupu konta premium przez użytkownika, zatem dla użytkowników, którzy wcale nie dokonali zakupu, będzie to cały okres użytkowania. Oprócz samego trendu będziemy odnotowywali:

- Największy wzrost - różnica między największą wartością, a najmniejszą wcześniejszą wartością.
- Największy spadek - analogicznie.
- Wartość początkowa (bazowa) - aby algorytm nie traktował jednakowo osoby, która stale słucha średnio 3 piosenek, z taką która słucha 8. Sama wartość trendu tego nie pokaże.

5 Miary jakości

Do mierzenia jakości obu modeli (z i bez uwzględnienia szeregów czasowych) zastosujemy te same metryki:

- Confusion Matrix - prezentuje dla każdej z klas, ile razy była rozpoznana poprawnie oraz z jakimi klasami była mylona. Do jej opisu są wykorzystywane następujące wartości:
  - TP - true positives, liczba przykładów poprawnie zaklasyfikowane jako pozytywne.
  - TN - true negatives, liczba przykładów błędnie zaklasyfikowane jako pozytywne.

  - FP - false positives, liczba przykładów poprawnie zaklasyfikowane jako negatywne.
  - FN - false negatives, liczba przykładów błędnie zaklasyfikowane jako negatywne.

- Accuracy - liczba przykładów zaklasyfikowanych do tej klasy.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- Precision - dokładność przewidywań danej klasy.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall - czułość modelu w rozpoznawaniu danej klasy.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- F1-score - średnia harmoniczna między Precision i Recall danej klasy.

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- ROC AUC - pole pod krzywą charakterystyki ROC, obrazujące zdolność modelu do odróżniania klas.

$$\text{ROC AUC} = \int_0^1 \text{TPR}(t)\, d(\text{FPR}(t))$$

  gdzie:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

- AP - średnia precyzja obliczana na podstawie krzywej Precision-Recall, dobrze odzwierciedla jakość modelu w przypadku nierównoważnych klas.

$$\text{AP} = \sum_n (R_n - R_{n-1}) \cdot P_n$$

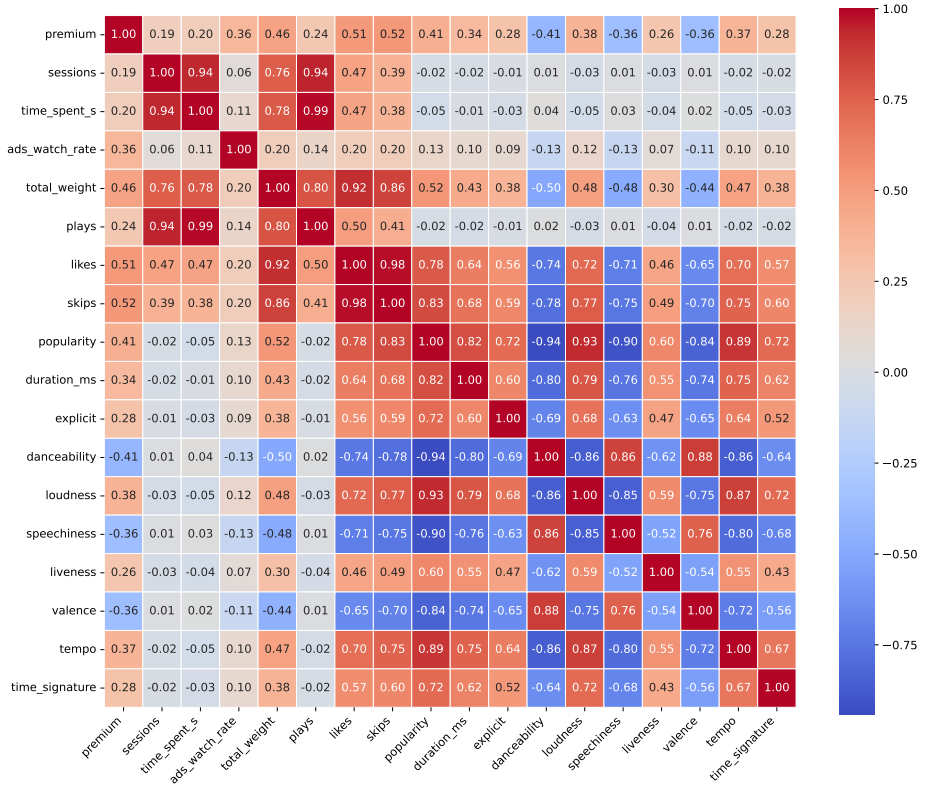
  gdzie: P_n to precyzja w punkcie n na krzywej, a R_n to czułość w punkcie n.

6 Model SVM bez szeregów czasowych

W celu sprawdzenia naszych założeń i atrybutów, stworzyliśmy model SVM z parametrami $c = 100$, gamma ≈ 0.064, z jądrem Gaussowskim RBF. Wykorzystaliśmy łącznie 17 atrybutów modelowanych w sposób nieuwzględniający szeregów czasowych (4.1). Atrybuty zostały dokładnie opisane w punkcie 3.2. Wszystkie atrybuty zostały poddane normalizacji. Rysunek 30 przedstawia macierz korelacji wykorzystanych atrybutów w modelu.

Widoczna selekcja atrybutów opierała się głównie na wyborze atrybutów najmocniej skorelowanych z przewidywaną wartością "premium".

Pozwoliło nam to na osiągnięcie precyzji predykcji na poziomie 85% mierzonej metryką Accuracy. Model zdecydowanie spełnia kryterium sukcesu, dając większą precyzję predykcji od modelu wybierającego klasę większościową stanowiącą 57%.



Rysunek 30: Macierz korelacji atrybutów modelu bez szeregów czasowych

6.1 Metryki modelu

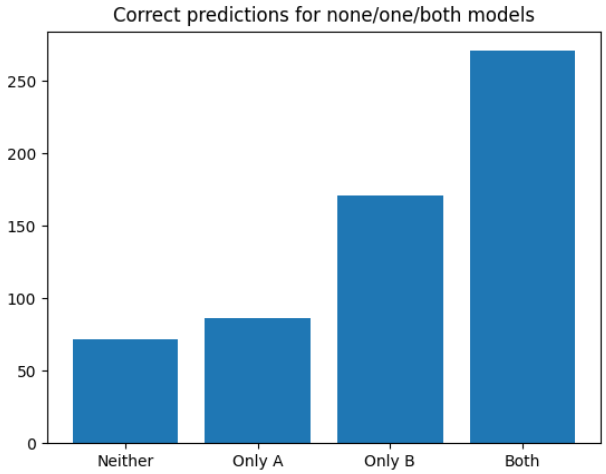
Oto wyniki modelu na podstawie miar jakości opisanych w punkcie 5:

- accuracy: 0.8467
- precision: 0.8756
- recall: 0.7243
- f1_score: 0.7928
- roc_auc: 0.8271
- average_precision: 0.7459

Model uzyskał wysoką dokładność (0.85) i precyzję (0.88), co świadczy o dobrej skuteczności klasyfikacji. Recall (0.72) wskazuje na obszar, który może wymagać poprawy, jednak F1-score (0.79) pokazuje dobry kompromis między precyzją a recall. ROC AUC (0.83) potwierdza zdolność modelu do rozróżniania klas, a średnia precyzja (0.75) świadczy o stabilnych wynikach przy różnych progach decyzyjnych.

6.2 Testy A/B

Do porównania modelu SVM (bez szeregów czasowych) przeprowadziliśmy testy A/B, w których porównaliśmy jego działanie z najprostszym modelem wybierającym klasę większościową (0 - użytkownik nie kupi premium). Model bez

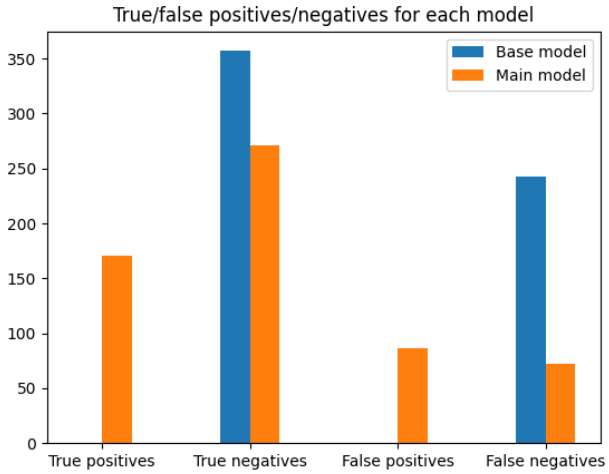


Rysunek 31: Poprawność precyzji obu modeli

szeregów czasowych został oznaczony jako "Main model", natomiast model większościowy jako "Base model". Na rysunku 31 została przedstawiona liczba poprawnych predykcji obu modeli.

Co nie powinno zaskakiwać, wiele przypadków było poprawnie klasyfikowanych przez oba modele - model A wybierał klasę większościową, stąd taka duża liczba poprawnych klasyfikacji. Jednak widać znaczącą różnicę w liczbie przykładów poprawnie klasyfikowanych przez model B w stosunku do modelu A. Również przypadki błędnie klasyfikowane przez oba modele muszą należeć do klasy 1, inaczej model większościowy klasyfikowałby je poprawnie. Z perspektywy modelu B, przykłady źle klasyfikowane przez oba modele można rozumieć jako false negatives. Natomiast te klasyfikowane tylko przez model A jako false positives.

Następny rysunek 32 prezentuje wartości z macierzy pomyłek dla obu modeli.

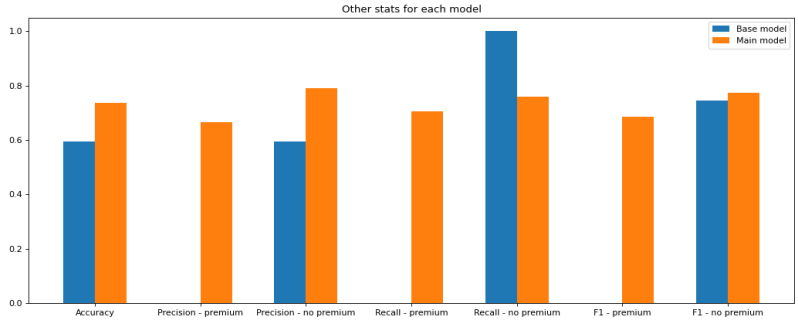


Rysunek 32: Wartości macierzy pomyłek obu modeli

Model bazowy nie ma, ani prawdziwych, ani fałszywych pozytywów, co wynika z faktu, że zawsze wybiera klasę 0. Model główny źle klasyfikuje podobną liczbę przypadków pozytywnych i negatywnych. Marginalnie większa jest liczba źle klasyfikowanych przypadków negatywnych (false positives), chociaż nie tak

duża jak ogólna przewaga liczby przypadków negatywnych 57% do przypadków pozytywnych 43%.

Kolejny rysunek 33 jest zestawieniem miar jakości obu modeli.



Rysunek 33: Miary jakości obu modeli

Metryka accuracy jest wyższa dla modelu bez szeregów czasowych. Wartości metryk dla klasy pozytywnej są jedynie dla modelu głównego, ponieważ model większościowy zawsze wybiera klasę 0. Recall dla modelu większościowego, dla klasy negatywnej jest równy 1, ponieważ nie ma żadnych przypadków fałszywych pozytywnych. Samo oddzielne porównanie wartości recall i precision dla przypadków negatywnych obu modeli może nie być tak istotne jak ich łączne zestawienie w metryce F1. Tutaj widać przewagę modelu głównego nad modelem większościowym nawet na klasie większościowej - "no premium".

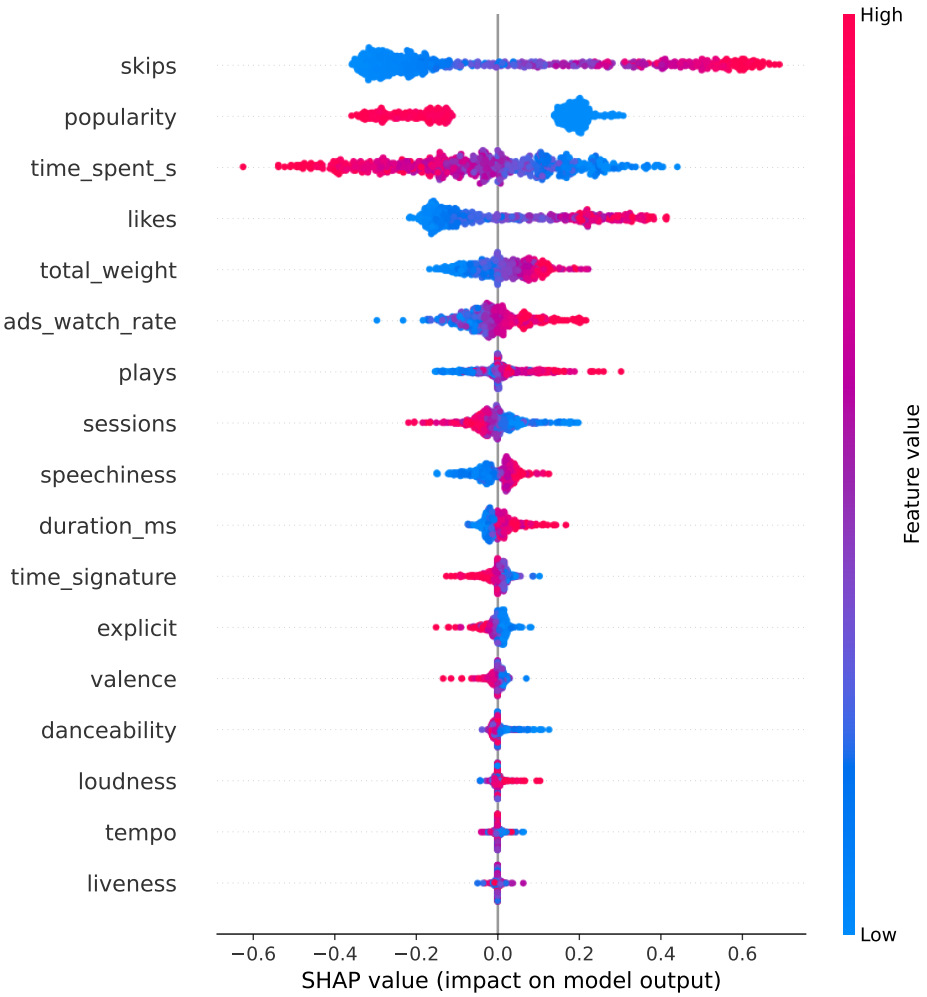
6.3 Wyjaśnialność modelu (XAI)

W celu lepszego zrozumienia działania modelu zastosowaliśmy metodę SHAP jako globalną metodę wyjaśnialności modelu. Niestety przez zastosowanie modelu SVM (z jądrem RBF) nie mogliśmy przeprowadzić bezpośredniej analizy wag modelu. Natomiast metoda SHAP dobrze działa również w takim przypadku. Polega na obliczeniu współczynników Shapley'a dla atrybutów w pojedynczych przypadkach (co jest metodą lokalną) i ich agregacji w celu pokazania globalnej wyjaśnialności modelu. Rysunek 34 pokazuje wpływ atrybutów na wyniki modelu.

Atrybuty na wykresie są posegregowane według ich wpływu na wynik, najbardziej wpływowe atrybuty są umieszczone u góry wykresu. Mocno wyróżnia się wykres parametru popularity, który dobrze rozróżnia przypisanie do zmiennej premium. Niższe wartości parametru oznaczają wzmocnienie przypisania do klasy "premium", wyższe do klasy, że użytkownik nie kupi premium. Jednak największe znaczenie ma atrybut skips, który też ma największą korelacje z atrybutem "premium", co można odczytać z macierzy korelacji 30.

Co ciekawe, mimo znacznie mniejszej korelacji ze zmienną celu, wysokie rozróżnienie daje atrybut time_spent_s. Natomiast inne silniej skorelowane atrybuty z przewidywaną klasą, jak np. likes lub total_weight występują po atrybucie time_spent_s mimo ponad dwukrotnie większej wartości korelacji. To pokazuje, że korelacja niekoniecznie przekłada się na ostateczny wpływ w przewidywaniu klasy.

Wynika to z tego, że korelacja mierzy liniową zależność atrybutów, natomiast współczynniki liczone w metodzie SHAP biorą pod uwagę wszystkie interakcje między atrybutami, uchwycając również nieliniowe zależności oraz interakcje.



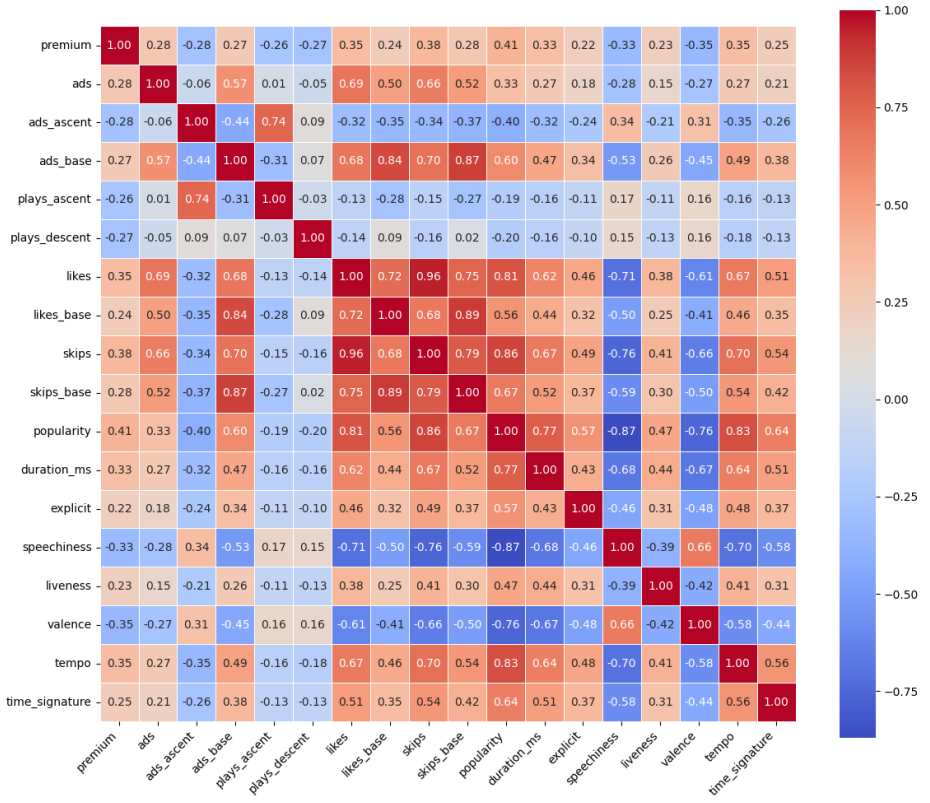
Rysunek 34: SHAP - wpływ atrybutów na działanie modelu

Zatem najbardziej znaczące atrybuty dla przewidywań to skips, popularity oraz time_spent_s, a w dalszej kolejności likes oraz total_weight.

7 Model SVM z szeregami czasowymi

Atrybuty, których wartości są zmienne w czasie, zostały odpowiednio zamodelowane z podziałem na okresy, zgodnie z opisem w punkcie (4.2). Pojedynczy okres trwa trzy miesiące i dla niego była wyciągana średnia danego zmiennego atrybutu. Atrybuty zmienne (powiązane z czasem) powstały na podstawie zbioru sesji użytkowników. Pod uwagę brane były jedynie zachowania użytkowników przed zakupem konta premium. To pozwala na lepsze odwzorowanie realnego stosowania modelu. Model będzie stosowany do detekcji, czy dany użytkownik kupi konto premium, a więc do modelu trafią dane użytkownika, który nie ma konta premium i to zostało odwzorowane w przygotowaniu zbioru danych.

Rysunek 35 przedstawia macierz korelacji dla zastosowanych atrybutów.



Rysunek 35: Macierz korelacji atrybutów modelu z szeregami czasowymi

Po jej analizie, do działania modelu wykorzystaliśmy 17 atrybutów:

- ads - suma reklam wyświetlonych użytkownikowi ze wszystkich okresów (cały okres użytkowania do zakupu premium).
- ads_ascent - największy wzrost liczby wyświetlonych reklam. Liczony jako różnica między okresem z najwyższą wartością, a okresem z najmniejszą wartością występującym wcześniej.
- ads_base - wartość liczby obejrzanych reklam w pierwszym okresie.
- plays_ascent - największy wzrost liczby odtworzeń utworów.
- plays_descent - największy spadek liczby odtworzeń utworów. Spadek jest liczony analogicznie do wzrostu - różnica między najmniejszą, a największą wcześniej występującą wartością.

- likes - suma polubień utworów ze wszystkich okresów.
- likes_base - wartość liczby polubień utworów w pierwszym okresie.
- skips - suma pominięć utworów ze wszystkich okresów.
- skips_base - wartość liczby pominięć utworów w pierwszym okresie.
- popularity, duration_ms, explicit, speechiness, liveness, valence, tempo, time_signature - atrybuty określające własności utworów, opisane w punkcie 3.2.9 - nie podlegające zmianom w czasie.

Widoczna selekcja atrybutów opierała się na wyborze atrybutów najmocniej skorelowanych z przewidywaną wartością "premium"oraz możliwie niskim skorelowaniu między sobą. Dodatkowo wszystkie atrybuty zostały poddane standaryzacji.

Model został przetestowany dla różnych parametrów i najwyższą dokładność uzyskał dla $c = 1$, gamma $= 10^{-7}$ i jądra wielomianowego poly. Pozwoliło to na osiągnięcie precyzji predykcji na poziomie 76,(6)% mierzonej metryką Accuracy. Model zdecydowanie spełnia kryterium sukcesu, dając większą precyzję predykcji od modelu wybierającego klasę większościową stanowiącą 57%.

7.1 Metryki modelu

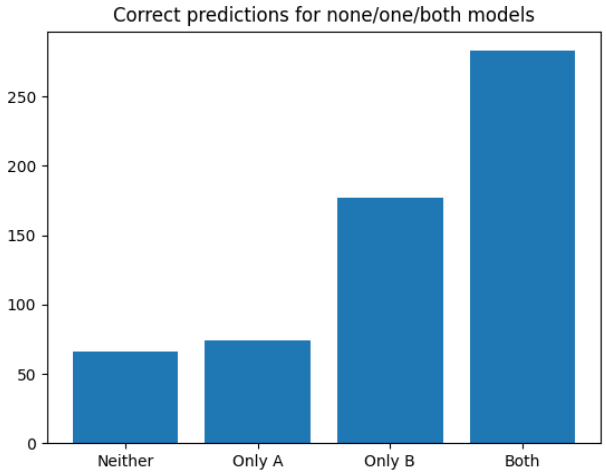
Oto wyniki modelu na podstawie miar jakości opisanych w punkcie 5:

- accuracy: 0.7667
- precision: 0.7052
- recall: 0.7284
- f1_score: 0.7166
- roc_auc: 0.7606
- average_precision: 0.6236

Uzyskane wyniki przez model z szeregami czasowymi są nieco niższe od wersji bez szeregów. Jedynie miara recall pozostała bez zmian, a nawet jest marginalnie wyższa. Największa różnica w wynikach jest na mierze precision, bo aż o (0.17). Niższe wyniki najprawdopodobniej wynikają z zastosowania mniejszego zbioru treningowego - wykorzystaliśmy dane sesji użytkowników jedynie do momentu zakupu konta premium, czyli bez analizy dalszego zachowania użytkownika po dokonaniu zakupu.

7.2 Testy A/B

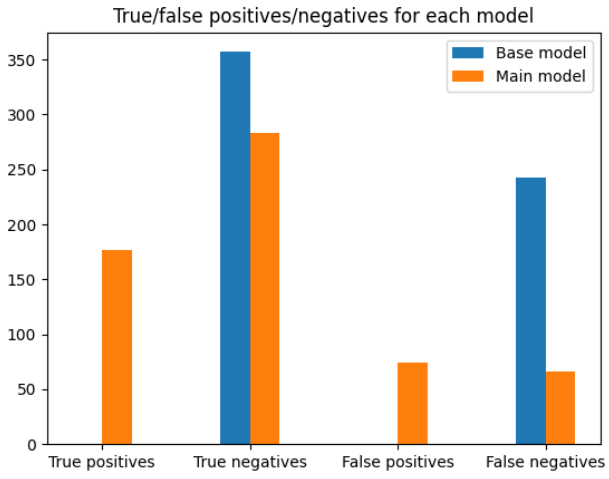
Do porównania modelu SVM z szeregami czasowymi, również przeprowadziliśmy testy A/B, porównując go z modelem wybierającym klasę większościową (analogicznie jak dla modelu bez szeregów czasowych 6.2). Model z szeregami czasowymi został oznaczony jako "Main model", natomiast model większościowy jako "Base model". Na rysunku 36 została przedstawiona liczba poprawnych predykcji obu modeli.



Rysunek 36: Poprawność precyzji obu modeli

Wykres jest bardzo zbliżony do tego z testu A/B z modelem bez szeregów czasowych. Również wiele przypadków było poprawnie klasyfikowanych przez oba modele oraz wystąpiła znacząca różnica w liczbie przykładów poprawnie klasyfikowanych przez model B w stosunku do modelu A. Analogicznie do poprzedniego przykładu, z perspektywy modelu B, przykłady źle klasyfikowane przez oba modele można rozumieć jako false negatives. Natomiast te klasyfikowane tylko przez model A jako false positives.

Następny rysunek 37 prezentuje wartości z macierzy pomyłek dla obu modeli.

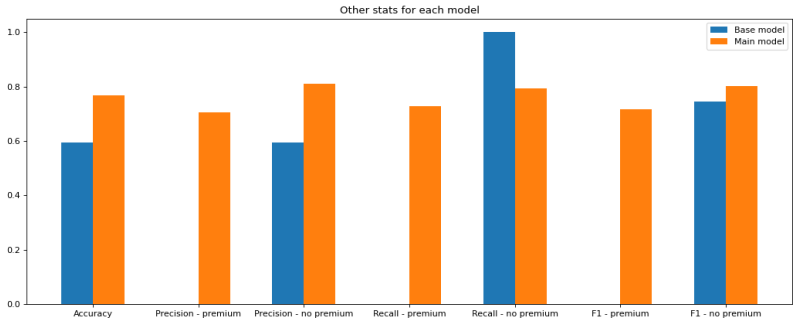


Rysunek 37: Wartości macierzy pomyłek obu modeli

Ten wykres jest również mocno podobny do tego z poprzedniego testu A/B. Model z szeregami również źle klasyfikuje podobną liczbę przypadków pozytywnych i negatywnych, i także marginalnie więcej przypadków negatywnych (false positives).

Kolejny rysunek 38 jest zestawieniem miar jakości obu modeli.

Metryka accuracy jest wyższa dla modelu z szeregami. Statystyki modelu większościowego są bez zmian względem poprzedniego testu. Również widać przewagę modelu głównego nad modelem większościowym nawet na klasie większo-



Rysunek 38: Miary jakości obu modeli

ściowej - "no premium".

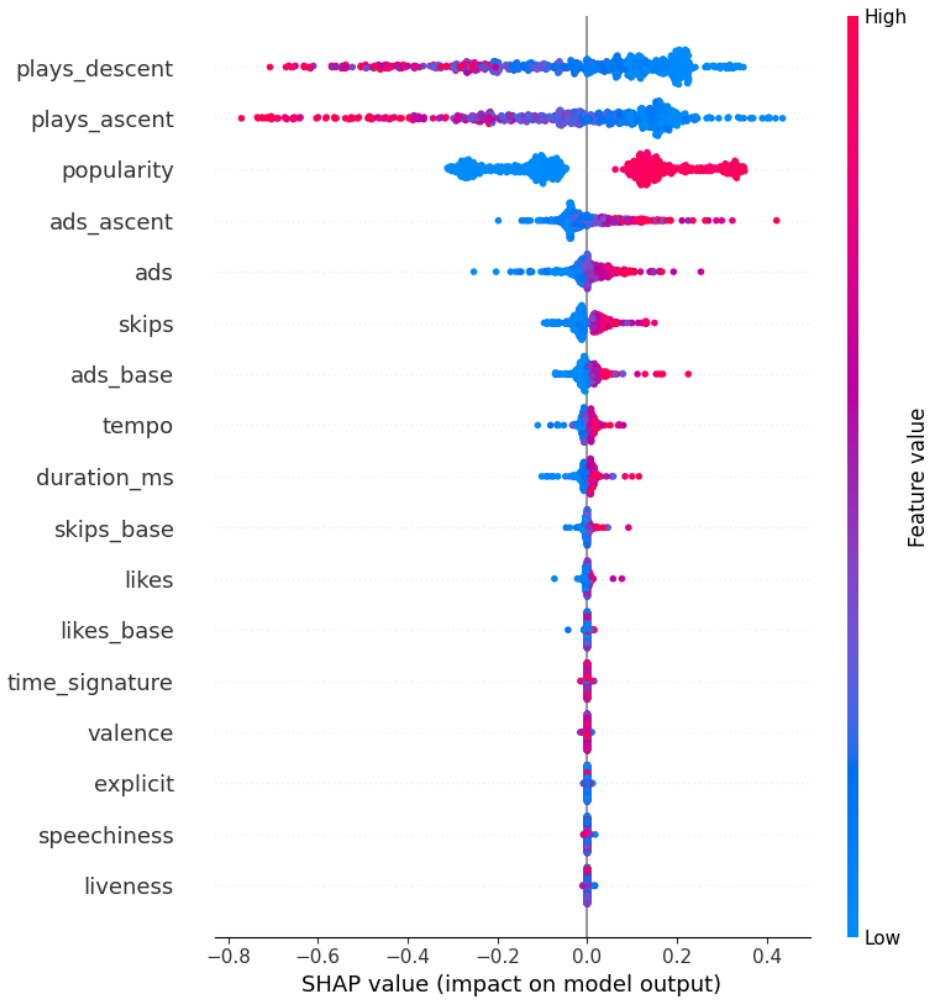
7.3 Wyjaśnialność modelu (XAI)

W celu lepszego zrozumienia działania modelu z szeregami czasowymi również zastosowaliśmy metodę SHAP jako globalną metodę wyjaśnialności modelu. Tutaj także nie mogliśmy przeprowadzić bezpośredniej analizy wag modelu. Rysunek 39 pokazuje wpływ atrybutów na wyniki modelu.

Na wykresie mocno wyróżnia się parametr popularity, który dobrze rozróżnia przypisanie do zmiennej premium. Wyższe wartości parametru oznaczają wzmocnienie przypisania do klasy "premium", niższe do klasy, że użytkownik nie kupi premium. To jest też zgodne z macierzą korelacji, z której wynika, że to ten atrybut jest najbardziej skorelowany z przewidywaną wartością "premium".

Co ciekawe, mimo mniejszej korelacji ze zmienną celu, największe rozróżnienie dają atrybuty plays_descent oraz plays_ascent. Mają one główne znaczenie w klasyfikacji przykładów mimo mniejszej korelacji. Z drugiej strony, następne w kolejności atrybuty pod kątem siły skorelowania z przewidywaną klasą po atrybucie "popularity"to: skips, likes, liveness. Mimo to nie mają już tak silnego wpływu na przewidywaną wartość "premium". To ponownie pokazuje, że korelacja niekoniecznie przekłada się na ostateczny wpływ w przewidywaniu klasy, tak jak przy analizie modelu bez szeregów czasowych.

Zatem najbardziej znaczące atrybuty dla przewidywań to plays_descent, plays_ascent oraz popularity, a w dalszej kolejności ads_ascent, ads oraz skips.



Rysunek 39: SHAP - wpływ atrybutów na działanie modelu

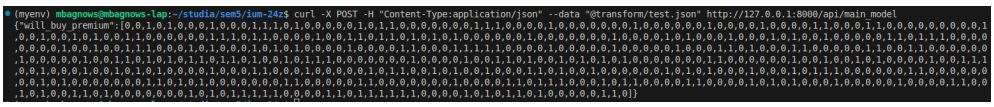
8 Uruchomienie modelu

Po wykonaniu niezbędnej konfiguracji zgodnie z instrukcją zamieszczoną w pliku README.md można uruchomić mikroserwis zawierający działający model wykonując polecenie "python3 microservice.py". Rysunek 40 prezentuje uruchomienie mikroserwisu.

```
(myenv) mbagnows@mbagnows-lap:~/studia/sem5/ium-24z$ python3 microservice.py
 * Serving Flask app 'microservice'
 * Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:8000
Press CTRL+C to quit
```

Rysunek 40: Uruchomienie mikroserwisu

Następnie, przy uruchomionym mikroserwisie można wysłać zapytanie o dokonanie predykcji dla dostarczonych danych. Rysunek 41 prezentuje przykładowe zapytanie do mikroserwisu wraz odpowiedź zwróconą przez mikroserwis - predykcjami modelu.



Rysunek 41: Wysłanie zapytania o predykcje

9 Podsumowanie

Oba podejścia do modelowania danych - bez szeregów czasowych oraz z szeregami, spełniły kryterium sukcesu w przewidywaniu, czy dany użytkownik zdecyduje się na zakup konta premium w serwisie "Pozytywka". Oba pozwolą na rozwiązanie problemu: komu opłaca się okazjonalnie wyświetlać reklamy konta premium zamiast zwykłych reklam, a komu lepiej wyświetlać same reklamy komercyjne. Jednak model bez szeregów uzyskał nieco lepsze wyniki, dlatego powinien być stosowany w pierwszej kolejności. Mimo to warto dla pewności wypróbować też działanie modelu z szeregami czasowymi. Bowiem w realnym środowisku modele mogą zachować się lekko inaczej i warto mieć pewność wyboru lepszego modelu. Alternatywnie można zastosować oba modele, uśredniając wynik prawdopodobieństwa przynależności do danej klasy z obu modeli i na tej podstawie przypisywać ostateczną etykietę.