



¿Se puede predecir el éxito de una canción en función de sus características? ¿Varía según la plataforma?

Universidad:

Universidad de San Andrés

Carrera:

Lic. Ciencias del Comportamiento

Materia:

Ciencia de Datos

Integrantes:

Aixa Bollini

Martina Brun

Iara Lorain Williams

Fecha de entrega: 07/12

Año: 2024

Introducción

La música es algo presente en nuestras vidas de forma cotidiana, especialmente a través de la plataforma de Spotify. Los artistas deben su éxito y sus ingresos, en gran medida, al éxito de sus canciones y esto se traduce en la cantidad de reproducciones que obtienen. Estudiar qué factores de una canción son más valorados por los oyentes puede ayudar a los artistas a hacer una canción mejor aceptada. Además, entender si existen diferentes tendencias entre plataformas musicales ayuda a entender mejor al público target y ver si existen diferentes segmentos.

Hoy en día casi todos los usuarios de los diferentes países acceden a la plataforma de Spotify para escuchar la música de los diferentes artistas, valorando a estos por sus canciones. El mercado de canciones cada vez mueve a más consumidores, siendo uno que se debe estudiar y entender para futuras intervenciones y modelos de negocios.

Tomando en cuenta la relevancia que tiene la música hoy en día, se quiere estudiar lo siguiente: ¿Se puede predecir el éxito de una canción en función de sus características? ¿Varía según la plataforma? Se busca hacer una predicción a corto plazo, entendiendo que es un mercado que está en constante cambio. Esto es porque los datos de hoy en día pueden no reflejar el mercado en diez años, pero si en el corto plazo. Esta información podría ser de vital importancia para el mercado de discográficas o productoras de música.

Búsqueda bibliografía

Existen algunos estudios que trataron esta idea. Por ejemplo, Vardo et al. (2023) utilizaron distintos métodos como árbol de decisiones, clasificador de vecinos más cercanos y Random Forest, para predecir el éxito de una canción en base a distintas características de las canciones de Spotify. Estas metodologías permitieron identificar la importancia de ciertos

factores que correlacionan con la popularidad de una canción, como por ejemplo, el mes de lanzamiento, la acústica y el tempo.

Otros estudios no obtuvieron resultados muy prometedores, sin embargo, los mismos autores justifican esto por la metodología y tipo de análisis de datos que utilizaron. Este fue el caso de Nijkamp, R. (2018), quien construyó un modelo de predicción por regresión que le permitió concluir que las características de audio de Spotify tienen poco o moderado poder explicativo para un mayor recuento de reproducciones.

Además, Sebastian, Jung y Mayer (2024) utilizan un modelo de aprendizaje automático como Random Forest, XGBoost y OLS para predecir la popularidad de canciones usando una base de datos de 30000 canciones, tomando datos desde 1957 hasta 2020. En este estudio se analizan características como el género, duración e instrumentalidad, siendo el primero uno de los mayores predictores.

Dimolitsas, Kantarelis y Fouka (2024) recopilaron datos de Spotify para predecir el éxito de canciones mediante algoritmos como Random Forest, SVM, Logistic Regression, y k-Nearest Neighbors. También aplica PCA para reducir la dimensionalidad de los datos y destaca variables importantes como la energía y el tempo.

Mordor Intelligence. (s. f.) busca hacer un análisis predictivo de la música del 2024 al 2029. Busca entender como va a crecer el tamaño del mercado, medido en dólares, estimando que va a crecer un 8.54% en los próximos 5 años. Además, le agrega relevancia al hecho de que hoy en día es muy simple poder acceder a la música, ya que se encuentra en la nube.

Para finalizar, Mordor Intelligence (s. f.) examina el crecimiento de aplicaciones de música, impulsado por el aumento en usuarios de smartphones, de internet, y las compras dentro de la aplicación. Destaca la influencia de la pandemia en el auge de las plataformas

de música digital y la integración de inteligencia artificial en apps como Spotify y Pandora para mejorar la experiencia del usuario.

Base de datos

<https://www.kaggle.com/datasets/abdulszz/spotify-most-streamed-songs?resource=download>

Los datos disponibles son:

Información general: “track_name”: Nombre de la canción. “artist(s)_name”: Nombre de el/los artistas. “artist_count”: Número de artistas que participan en la canción.

“released_year”, “released_month”, “released_day”: Fecha de lanzamiento.

Métricas de las plataformas: “in_spotify_playlists”: Cantidad de playlists de Spotify donde esta la canción. “in_spotify_charts”: Clasificación de la canción en las listas de Spotify.

“streams”: Total de reproducciones en Spotify. “in_apple_playlists”, “in_apple_charts”:

Presencia en las playlists y listas de Apple Music. “in_deezer_playlists”, “in_deezer_charts”:

Presencia en las playlists y listas de Deezer. “in_shazam_charts”: Clasificación de la canción en las listas de Shazam.

Métricas de la canción: “bpm”: Beats por minuto (tempo de la canción). “key”: Tonalidad de

la canción. “mode”: Modo mayor o menor. “danceability_%”: Idoneidad de la canción para

bailar. “valence_%”: Positividad del contenido musical de la canción. “energy_%”: Nivel percibido de energía de la canción. “acousticness_%”: Presencia acústica.

“instrumentalness_%”: Proporción de contenido instrumental. “liveness_%”: Presencia de

elementos de actuación en vivo. “speechiness_%”: Cantidad de palabras en la canción.

Metodología

Métodos para la etapa exploratoria

Debido a que la base de datos no cuenta con valores atípicos, nulos o faltantes, no se realizó una etapa de limpieza. Como primer paso, se hizo una matriz de correlación (figura 1) entre todas las variables numéricas. En este análisis, no encontramos una correlación importante entre alguna de las características de la canción y la cantidad de reproducciones. Solo se ve una correlación alta entre algunas características (como energy y accoustiness -0.58) lo cual tiene sentido, y entre la cantidad de reproducciones y la presencia en playlist. Sin embargo, solo estamos analizando relaciones entre dos variables y quizás existe una combinación de características que permita predecir el éxito de una canción.

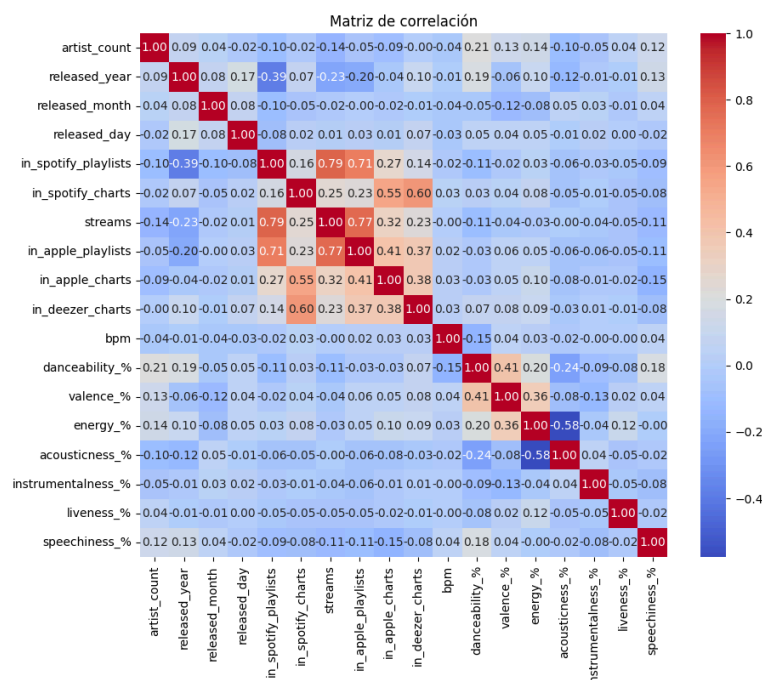


Figura 1: Matriz de correlación de todas las variables numéricas de la base de datos.

Continuando con el análisis, observamos el promedio de reproducciones en Spotify, el cual dio como resultado 514.137.400.

Además se realizaron histogramas para ver las distribuciones de las variables numéricas (figura 2). Puede observarse que “bpm”, “danceability_%” y “energy_%” parecerían seguir una distribución normal.

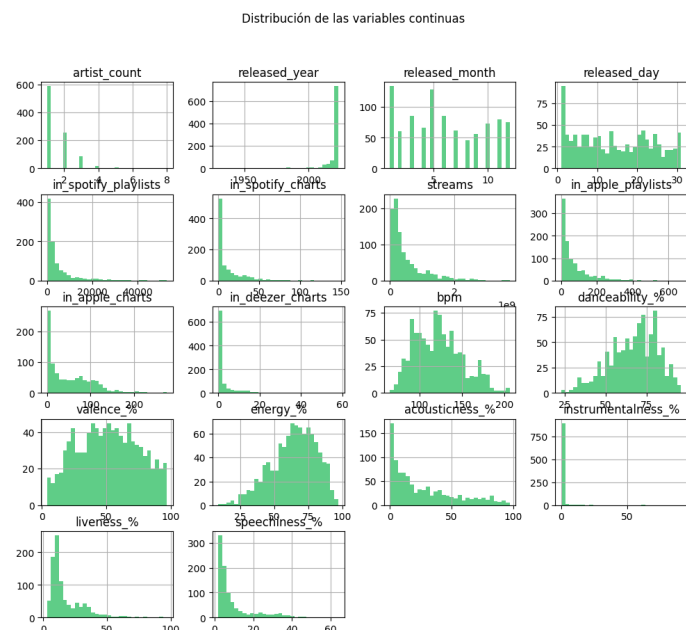


Figura 2: Histogramas de las variables continuas de la base de datos.

Por último, se analizaron la cantidad de reproducciones según el año de lanzamiento de las canciones(figura 3). Puede verse un pico en la cantidad de reproducciones en canciones lanzadas entre los 70 y 80, probablemente debido a alguna moda en el año cuando se publicó el dataset (como la última temporada de la serie de Stranger Things), aunque es solo una hipótesis .

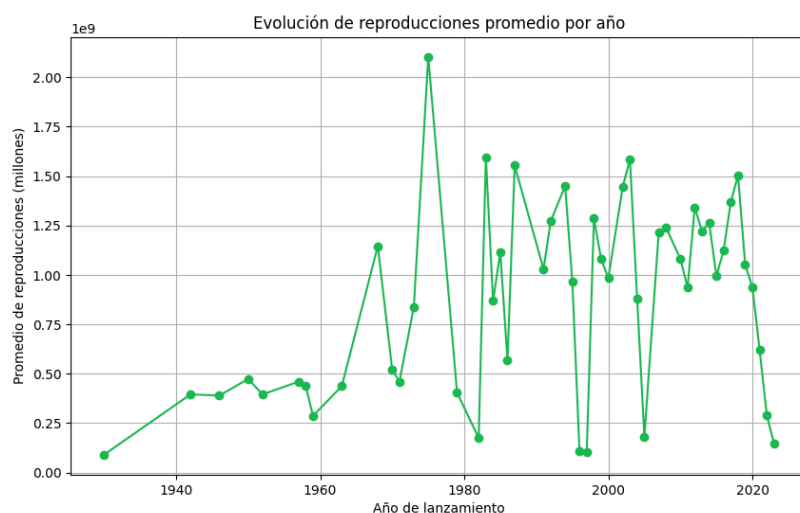


Figura 3: Las reproducciones de las canciones tomando en cuenta el año en el que se lanzaron.

Análisis predictivo

Se debe crear una variable categórica “éxito” que valga 1 o 0 dependiendo si la cantidad de streams > x número con el fin de poder realizar una regresión logística. Una vez que tengamos la variable “éxito”, se puede realizar una regresión logística con las métricas de la canción como predictores para calcular la $Pr(Y|X)$ y usando el clasificador de Bayes para la predicción. Además, podemos realizar un análisis discriminante lineal con más de una categoría para calcular también la $Pr(Y|X)$. Luego de estos dos análisis, finalmente podríamos decidir cual es mejor comparando el accuracy de cada modelo a través del análisis ROC. Para finalizar, podemos ver si este modelo funciona para otras plataformas haciendo nuevas variables categóricas de “éxito” que tengan en cuenta la cantidad de reproducciones promedio de cada plataforma.

Otro análisis que podemos hacer con los datos que tenemos y consideramos más apropiado es random forest con árboles categóricos. Este método nos permite clasificar las canciones entre exitosas y no exitosas, evitando el overfitting que tiene un solo árbol y descorrelacionando los árboles entre sí para encontrar el resultado más óptimo. Se utilizará el coeficiente de Gini para medir la pureza de los nodos. Además, para elegir la profundidad de los árboles, utilizaremos cross validation, probando las siguientes profundidades: [3, 5, 10, 15, None].

Limitaciones y conclusiones

En este trabajo se cuenta con algunas limitaciones, tanto de la base de datos como de la metodología elegida. En primer lugar, la base cuenta con datos viejos. Teniendo en cuenta que el mercado y tendencias de la música muta constantemente, es posible que haya habido cambios en el mercado desde el momento de la recolección de los datos hasta el


presente, por lo que estos podrían no reflejar las condiciones de éxitos actuales y futuras cercanas. En segundo lugar, hay que tener en cuenta de qué manera definimos “éxito”. En nuestro caso, el criterio elegido para clasificar una canción como exitosa es la cantidad de reproducciones, sin embargo, esto puede no capturar completamente el concepto, ya que factores externos como el reconocimiento del artista, el impacto cultural o las campañas de marketing podrían estar relacionados con lo que se busca predecir; más estos aspectos no se incluyen en la base. Por último, hay que tener en cuenta que algunas métricas están correlacionadas, lo que puede influir en los resultados predictivos al sesgar la importancia que se le atribuye a ciertas características.

Si bien el estudio cuenta con limitaciones, se puede concluir que este tipo de estudio es relevante, ya que el mercado de la música está creciendo con el paso del tiempo, y todavía no hay muchos estudios específicos acerca de esto. Para agregar, un análisis de esto va a servir a los artistas para que puedan decidir qué tipo de música sacar si es que les interesa tener más reproducciones. Por último, este tipo de estudio va a hacer que los artistas puedan decidir estratégicamente en qué plataforma les conviene publicar sus canciones, para poder llegar al target correspondiente y tener mayor cantidad de reproducciones.

A modo de conclusión final, se puede ver que la mayoría de los estudios previos realizan regresiones para las predicciones, sin embargo, en el presente proyecto se proponen otros métodos que podrían ser más efectivos y precisos.

Como recomendaciones futuras, se propone ampliar el análisis a más plataformas, además de Spotify, y considerar variables externas, para así obtener un panorama más completo del éxito de una canción.

Link código

 Untitled4.ipynb

<https://colab.research.google.com/drive/1XID6qgRQPJPYiffHtPjilC9Qdxbin0qK?usp=sharing>

Referencias bibliográficas

Music App market Size | *Mordor Intelligence*. (s. f.).

<https://www.mordorintelligence.com/industry-reports/music-app-market/market-size>

Nijkamp, R. (2018). Prediction of product success: explaining song popularity by audio

features from Spotify data (*Bachelor's thesis*, University of Twente).

Panorama del mercado musical Volumen del mercado | *Mordor Intelligence*. (s. f.).

<https://www.mordorintelligence.com/es/industry-reports/music-market-landscape>

Sebastian, N., Jung, & Mayer, F. (2024, 1 marzo). Beyond Beats: A Recipe to Song Popularity?

A machine learning approach. *arXiv.org*. <https://arxiv.org/abs/2403.12079>

SpotHitPY: A study for ML-Based song hit prediction using Spotify. (2024). *Ar5iv*.

<https://ar5iv.org/html/2301.07978>

Vardo, L., Jerkić, J., & Žunić, E. (2023, March). Predicting song success: Understanding track

features and predicting popularity using spotify data. *In 2023 22nd International*

Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE. DOI:

10.1109/INFOTEH57020.2023.10094172