CODER HOUSE

DATA SCIENCE

PROYECTO FINAL

¿CÓMO CREAR HITS?

SPOTIFY HITS PREDICTION

CAMADA 16330

GRUPO 8

Melisa Leyton
Micaela Ferrari Navarro
Domingo Rizzi
Luciano Colucci

1. Caso de negocio

El acelerado desarrollo tecnológico desde mediados del siglo XX, en particular el boom de la utilización de dispositivos electrónicos en la producción musical a partir de mediados del mismo siglo, y la simultánea mayor accesibilidad de la tecnología a más gente en cada vez más partes del mundo, llevó a un fenómeno a nivel mundial que sucedió por primera vez en la historia: cualquier persona con acceso a un ordenador, un smartphone o una tablet puede producir su propia música y darse a conocer, hasta alcanzar la fama, como músico, compositor o productor.

El nuevo término acuñado para estos artistas es “Bedroom Producer”, o productor de dormitorio. En el mercado musical actual, algunos de los artistas más exitosos ahora son Bedroom Producers como Grimes, Richie Hawtin, Flume o Billy Eilish. Con la existencia y difundida utilización de las redes sociales y plataformas musicales como iTunes, Soundcloud, Youtube, Spotify, Apple Music, o Deezer (algunas de estas inclusive consideradas redes sociales en sí mismas), los artistas pueden hacer llegar su música a cualquier persona en el mundo que tenga acceso a internet. En 2020, las ganancias totales obtenidas por la industria del streaming musical en línea llegó a los 13.4 billones de dólares en todo el mundo, un valor que cuadruplica el total del año 2015[1].

Esta nueva ola de músicos pueden darse a conocer en el marco del desarrollo de la era digital, y en particular la industria musical digital, presentada aquí en contraposición a los formatos físicos que predominaron hasta principios del siglo XXI. Para tener un pantallazo del gran cambio que sufrió la industria, las ventas físicas a nivel mundial en 2020 cayeron a 31,6 millones de dólares mundialmente, cuando en 2000 eran de 938,9 millones.

La industria musical digital presenta muchas ventajas en comparación con la física. Por ejemplo, hay un seguimiento claro y registrado de la reproducción, descargas y ventas de una pieza musical. La forma de adquisición de la misma está disponible a cualquier comprador en diversos formatos (tarjetas de crédito, crédito online, pagos con billeteras virtuales).

Sin embargo, a pesar de todas estas ventajas, la enorme accesibilidad a cualquier persona para convertirse en un “Bedroom Producer” ha supuesto un desafío para otros actores de la industria de la música, como lo son las grandes productoras de música como Warner o Sony, productores individuales, centros de distribución y venta de música en formato físico, entre otros. Deben competir contra un crecimiento cada vez más acelerado de nuevos artistas emergentes, por lo que deben descubrir nuevos talentos cada vez más rápido y asegurarse de que creen canciones que serán un hit para obtener muchas ganancias de las ventas, principalmente en formato digital. Por otro lado, debieron reforzar su presencia en redes sociales, principalmente aquellas más audiovisuales como Instagram, Tik Tok, y los videos de Twitter y Facebook, y actualizarse en las nuevas plataformas de streaming y descargas de música.

En la realidad moderna, cada vez más acelerada, deben captar la atención de los usuarios de forma rápida y las discográficas tienen poco tiempo para mostrar su producto. Esto llevó a la industria musical a producir más cantidad de singles o EPs (formato parecido a un disco común o “LP”, pero con menor cantidad de canciones) en contraste con los tradicionales discos.

---

[1] https://www.statista.com/statistics/587216/music-streaming-revenue/

Diccionario

- Bedroom Producer: músico aficionado que crea, interpreta y graba su música de forma independiente utilizando un estudio en casa, a menudo considerado un aficionado frente a un productor de discos profesional en la industria discográfica que trabaja en un estudio tradicional con clientes.

- Single: Disco musical pequeño y de corta duración, que contiene generalmente una o dos canciones.

- EPs: sigla inglesa que traducida al español significa reproducción extendida y se utiliza como denominación para un formato de grabación musical. La duración de un EP es muy larga para considerarse como sencillo, y muy corta para considerarse como álbum.

- LP: es un disco de vinilo de tamaño grande, de 30,5 cm de diámetro, en el cual se puede grabar, en formato analógico, un máximo de unos 20 a 25 minutos de sonido por cada cara. Los LP suelen constar de unas ocho, diez o doce canciones, dependiendo de su duración, y están grabados a una velocidad de 33 y 1/3 revoluciones por minuto (RPM).

2. Problema de negocio

Con la creciente cantidad de posibilidades para los artistas de componer y crear su propia música, inclusive desde la comodidad de su casa, subirla a internet y convertirla en un hit, las discográficas que representan a los músicos están teniendo cada vez más problemas en discernir qué canción se convertirá en un hit o cuál ganará más popularidad que otra.
Cabe mencionar que, si bien el problema de negocio mencionado es de las discográficas, el modelo desarrollado en este proyecto también puede ser vendido a los mencionados Bedroom Producers.

3. Contexto: hipótesis iniciales

Este trabajo toma como verdadera la hipótesis de que existen canciones que tienen ciertos atributos, los cuales predisponen al track a convertirse en un hit con mayor probabilidad que otros tracks. Estos atributos son intrínsecos de la canción, o extrínsecos a la misma (es decir, pueden ser variables contextuales como el álbum al que pertenece, la época del año de lanzamiento, situación social de la población que lo escuchó por primera vez, entre otros).
Libros como “The Song Machine” de John Seabrook, y varios papers como “Hit Song Science Is Not Yet a Science” y “Automatic Prediction of Hit Songs” apoyan la teoría de que las canciones que se convierten en hits o ganan mucha popularidad están de hecho diseñadas para ser convertirse en tales, ya que gozan de estas cualidades particulares que otras canciones no tienen las convierten en un potencial track exitoso para ciertos artistas. Tomando estas teorías como verdaderas, sólo los atributos intrínsecos a la canción (tempo, energía, timbre, intensidad, entre otros) definirán si esta será un hit o no.

4. Objetivo del modelo

La utilización de un modelo supervisado de clasificación de Machine Learning nos permitirá predecir con cierta seguridad si una canción será o no un éxito. El objetivo del modelo es determinar con la mayor precisión posible si determinada pieza musical será o no un hit, en función de las características y atributos de esa canción.

5. Desarrollo del Modelo

Bases de datos

Se tomó como base de datos el dataset "Spotify and Genious Track Dataset". Los datos se obtuvieron de la API de Spotify, y son relativos a los Tracks, Albums y Artistas. Está dividido en tres documentos .csv según este criterio. El primer archivo de *tracks* contiene en él las *audio features*, o características intrínsecas de la canción como la letra, energía, género y más, además de la información principal de una canción como su nombre, artista y duración. El segundo y tercer documento contienen datos contextual, como el álbum, el artista, la fecha de lanzamiento de un álbum, qué álbum es de ese artista y más. Para desarrollar el modelo de Machine Learning se utilizó el dataset de Audio Features.

Descripción del modelo desde el negocio

Se utilizó sólo la información intrínseca de las canciones, contenidas en el dataset de Audio Features, con el objetivo de desarrollar el modelo en función de los atributos de las canciones en las que se basan las investigaciones mencionadas anteriormente, para poder determinar si será exitosa o no.
Consecuentemente, esto otorga a un Bedroom Producer, o a una discográfica, un mayor control del "nivel de éxito" de una canción. Mientras más información se tenga sobre cuáles son los determinantes para que un track se convierta en un hit, los mismos se pueden definir más claramente y la probabilidad de lograr un éxito con un track aumentará. En un futuro posible, planteamos la posibilidad de que pueda crearse una "Máquina de Hits", si se obtienen los ingredientes para la "receta del éxito" para una canción.

Variables

La variable target será el grado de popularidad de un track. La variable es numérica decimal, y se encuentra como "popularity" en el dataset.
Para obtener una variable categórica de la misma, se creó la variable "popularity_cat" que define dos categorías de popularidad, asociada a la variable númerica de "popularity" de manera directa y positiva (a mayor popularity, será más probable que el track sea un hit):

hit
no hit

Si bien nuestra variable target acorde al problema de negocio sabíamos de qué se iba a tratar, no es lo primero que hicimos con las variables, primero no podemos olvidarnos del Data Wrangling y Data Cleaning.
Reducimos la dimensionalidad del dataset que inicialmente tenía más variables, mejorando así la capacidad de procesamiento de datos, y llegando a 31 columnas en total que describen los atributos intrínsecos de una canción.
Dentro de las variables escogidas, debíamos revisar valores vacíos, u outliers, es decir, valores llamativamente fuera de la norma. Una vez detectados estos valores, podemos ya sea dropear la línea o rellenarla, esto siempre a criterio del científico de datos a cargo. Afortunadamente, la información ya traída de la API de Spotify tiene pocos problemas de este tipo, ya que por defecto no trae valores vacíos.

Mediante análisis univariado, bivariado, y multivariado de las columnas, se busca una posible correlación de éstas con la variable target. Para este tipo de análisis, se toma las columnas, o variables, por separado o en conjunto, y se analiza por ejemplo, si es una variable numérica, se ve las principales medidas estadísticas,o por ejemplo, si se trata de una variable categórica, poder graficar cuál es la distribución de las canciones en ciertas categorías. También se estudia la relación entre las variables, es decir la correlación, por ejemplo. Esto nos permitirá detectar colinealidad y el mismo modelo que desarrollemos nos permitirá eliminar columnas facilitando el procesamiento de los datos.

Modelo supervisado de clasificación

Definimos hacer un modelo de Machine Learning supervisado ya que los datos de entrada están etiquetados; es decir, cada track tiene asociado un nivel de popularidad determinado. Además, el modelo es de clasificación ya que queremos predecir en qué categoría de popularidad estará el track.
Para poder predecir esto, y ya que nos basamos en el dataset de tracks subidos a Spotify con las variables que mide la misma plataforma, será necesario que los tracks con los cuales se alimente al modelo sean canciones que estén en la plataforma de Spotify; esto hará que tengan ciertos valores para cada una de las variables del modelo que permitirá al mismo predecir la categoría de popularidad a la cual pertenece.

Algoritmos de clasificación

Se desarrolló el modelo supervisado con cuatro algoritmos diferentes de clasificación: KNN, Árbol de decisión, Regresión Logística y Random Forest. Es este último modelo el que mejor ha performado para los datos con los que contamos.
Los algoritmos Random Forest son sensibles a la data desbalanceada, ya que el algoritmo tiende a sesgar hacia la categoría con mayor cantidad de datos. La regresión logística también es sensible a la data desbalanceada ya que afecta la pendiente de regresión. Entonces, para poder desarrollarlos correctamente, previo a esto se realizó un balanceo de las categorías sobre la muestra obtenida.

Mejora de modelos

Se aplicaron modelos de mejora sobre el árbol de decisión, al cual se aplicó Grid Search y Randomized Search, y a Random Forest, al cual se aplicó Randomized Forest.

Boosting y Modelos de ensamble

El objetivo de Boosting, que es un enfoque de Machine Learning y una serie de algoritmos aplicables a un modelo, es mejorar el rendimiento del algoritmo de aprendizaje al tratarlo como una "caja negra" que se puede llamar repetidamente. Si bien el algoritmo de aprendizaje base puede ser rudimentario y moderadamente inexacto, no es del todo trivial ni poco informativo y debe obtener resultados mejores a los que se podrían obtener en forma aleatoria. La idea fundamental detrás de Boosting es elegir conjuntos de entrenamiento para el algoritmo de aprendizaje base de tal manera que lo obligue a inferir algo nuevo sobre los datos cada vez que se lo llame.
En este trabajo, se aplicaron los modelos Adaboost, XG Boost, Light GBM y Gradient Boost.

6. Métricas y desempeño

Descripción de las métricas a tomar en cuenta:

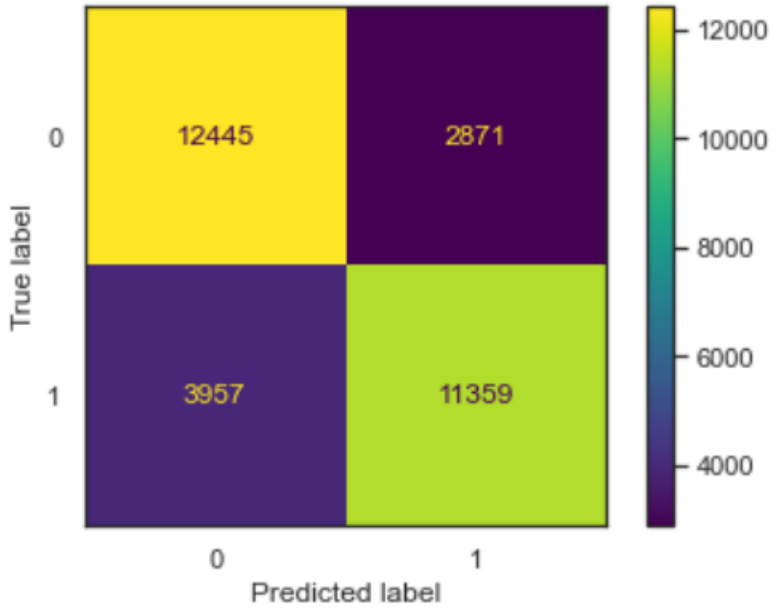
- Accuracy (Machine Learning Models): El número de clasificaciones que un modelo predice correctamente, dividido por el número total de predicciones realizadas. Es una forma de evaluar el rendimiento de un modelo.
- Precision: Mide la calidad del algoritmo en base a los verdaderos positivos y los falsos positivos
- Recall score: Es la habilidad del algoritmo de encontrar verdaderos positivos
- f1: Media entre Precision y Recall score
- AUC: Esta métrica describe la performance de un modelo de manera directa y positiva: a mayor AUC, mejor es la performance del modelo entre categorías. Puede interpretarse como la probabilidad de que el modelo de cómo resultados verdaderos positivos cuando corresponde.
- Accuracy (Boosting Models): Esta medida es básicamente el número total de predicciones correctas dividido por el número total de predicciones.

7. Resultados y análisis

Los resultados obtenidos con el Algoritmo de clasificación de Random Forest demostraron ser el que mejor ajusta a la predicción de hits para el dataset utilizado.
Las métricas obtenidas fueron:

Accuracy: 0,77
Precision: 0,80
Recall score (verdaderos positivos): 0,74
F1 Score: 0,77

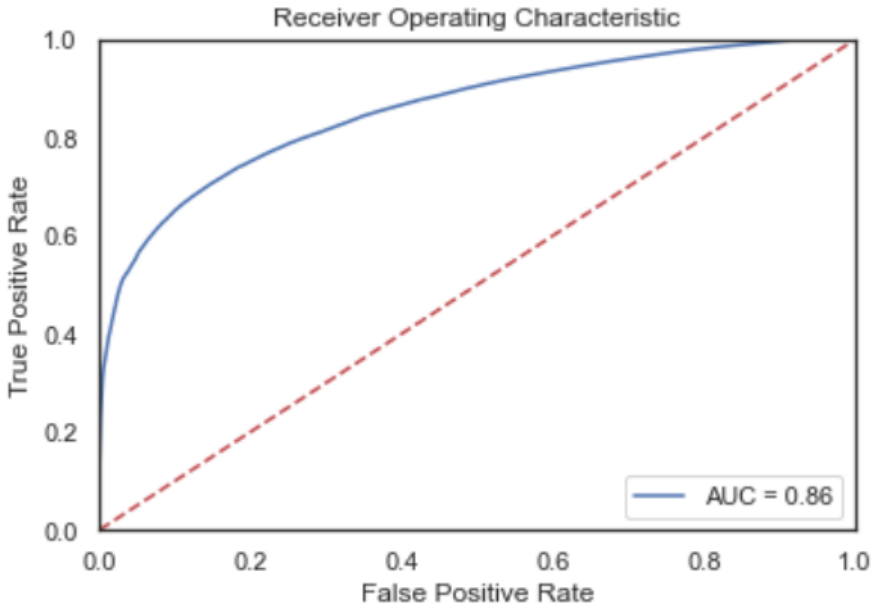
Vemos a partir de la matriz de confusión que el algoritmo cometió errores de clasificación en menos del 20% de las clasificaciones. Cometió aproximadamente un 36% de errores de tipo II al momento de clasificar, lo que podría conducir a que se prediga que las canciones no son hits en base a sus atributos de audio, pero que tienen todo el potencial para poder serlo. Este error podría llevar a una estrategia de marketing de no promocionar la canción y por lo tanto que no se vuelva popular. Sin embargo, el algoritmo tiene la capacidad de encontrar hits en más de un 80% de canciones.



Matriz de confusión Random Forest

La matriz de confusión estudiada en el modelo de Random Forest se interpreta:

- Predijo que la canción no es un hit (0) y verdaderamente no eran hits: 12445 clasificaciones
- Predijo que la canción es un hit (1) y verdaderamente lo es: 11359 clasificaciones
- Predijo que la canción no es un hit(0), pero verdaderamente eran hits(1): 3957 clasificaciones (Error tipo II)
- Predijo que la canción es un hit(1), pero verdaderamente no lo eran(0): 2871 clasificaciones (Error tipo I)



Curva ROC AUC

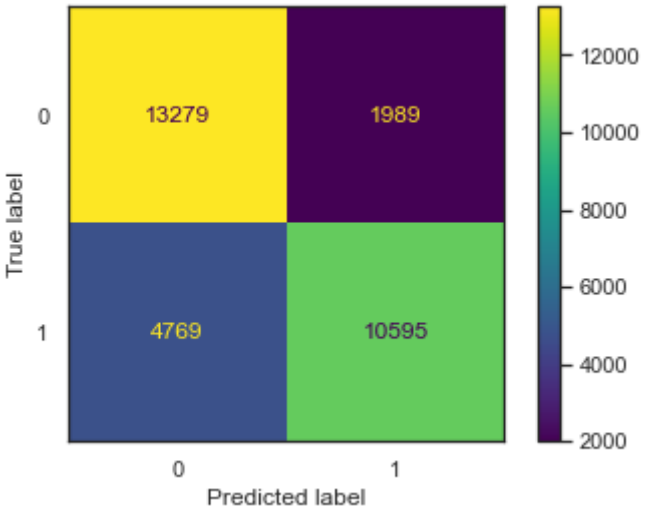
Con un AUC de 0,86, comparativamente con los otros modelos, podemos decir que la performance del Árbol de decisión es buena por su AUC elevado.

Como resultados de la aplicación de las mejoras al Árbol de decisión, y a Random Forest, podemos decir:

- En el primer caso se observa que el AUC para el modelo GridSearch es menor que para el del árbol de decisión. La performance del modelo es mucho menor, por esto y, como consecuencia de un AUC elevado en el modelo de árbol de decisión, esto lleva a concluir que el modelo de árbol de decisión se encuentra sobre ajustado.
- Luego, con Randomized Search, Se obtuvo una performance mucho menor del modelo con respecto al modelo de árbol de decisión original, probablemente porque la combinación aleatoria de los modelos no encontró el mejor ajuste, o porque el modelo original se encuentra sobre ajustado
- En el segundo caso, con Randomized Forest, en ambos casos se obtuvo una performance mucho menor del modelo con respecto al modelo de árbol de decisión original, probablemente por las mismas razones que en Randomized Search para árboles de decisión.

Como resultado de la aplicación de los algoritmos de Boosting, concluimos que XG Boost nos provee buenos resultados de predicción con bajos esfuerzos de procesamiento; esta última característica es la única que diferencia a XG Boost como mejor algoritmo respecto de los otros, ya que en general todos los modelos han obtenido performance similares. Las métricas obtenidas fueron:

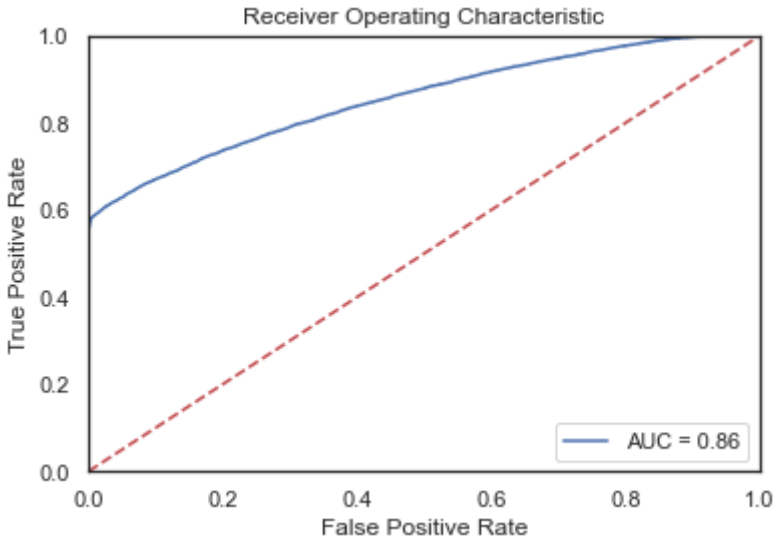
Accuracy: 0,78
Precision: 0,84
Recall score (verdaderos positivos): 0,69
F1 Score: 0,76



Matriz de confusión XGBOOST

De la matriz de confusión estudiada se interpreta que el algoritmo de XGBoost:

- Predijo que la cancion no es un hit (0) y verdaderamente no eran hits: 13279 clasificaciones
- Predijo que la cancion es un hit (1) y verdaderamente lo es: 10595 clasificaciones
- Predijo que la canción no es un hit(0), pero verdaderamente eran hits(1): 4769 clasificaciones (Error tipo II)
- Predijo que la canción es un hit(1), pero verdaderamente no lo eran(0): 1989 clasificaciones (Error tipo I)



Curva ROC AUC para XG BOOST

El algoritmo posee métricas aceptables a la hora de realizar la clasificación y poder encontrar verdaderos hits. En la ROC AUC se observa que a bajas tasas de falsos positivos, el algoritmo tiene un rate de 60% de verdaderos positivos (canciones que son hits). Por lo tanto, cuando la tasa de verdaderos positivos de este algoritmo es alta, la tasa de falsos positivos es muy baja.

8. Conclusiones

El mejor algoritmo de clasificación de canciones que serán un HIT es Random Forest, ya que sus métricas de performance son superiores a las de los demás modelos. Adicionalmente, sabemos que existen y pueden utilizarse algoritmos de Boosting para mejorar la precisión de los modelos de clasificación tradicionales de Machine Learning; en nuestro caso, la aplicación de estas mejoras elevó la precisión de la predicción hasta en un 5%, según la comparación de métricas referidas como el Accuracy.
Finalmente, como objetivo principal que teníamos en este trabajo, podemos concluir que es posible predecir con alta precisión si una canción se convertirá en un HIT o no según sus atributos intrínsecos, mediante un modelo de Machine Learning supervisado de clasificación.

9. Mejoras Futuras

Se indican las siguientes propuestas como posibles mejoras futuras al desarrollo de cada una de las etapas de este trabajo:

- El modelo está basado en bases de datos obtenidas de la API de Spotify. Como se indicó al principio, es importante conocer que el mercado de la música se expande mucho más allá de esta herramienta; Spotify es tan solo una conocida plataforma de streaming e instrumento de marketing para los productores o artistas. En un futuro ambicioso, se propone aumentar la cantidad de datos a los cuales se aplica este modelo con información de otras plataformas de streaming como Apple Music, por ejemplo.
- La API de Spotify provee los atributos de la música que ellos han decidido obtener de cada una de las canciones. Si se quisiera determinar el valor de alguno diferente, tendría que llevarse a cabo un proceso aparte para lograrlo. Cabe mencionar además que no existe razón para considerar a los atributos determinados por la API de Spotify como los atributos intrínsecos que mejor pueden describir si un track será un hit o no.