Aprendizaje Máquina, Febrero 2021.

Popularidad de canciones spotify

Óscar Alvarado[*] Dante Bermúdez[**] Martín García[***]

Resumen

En el presente trabajo se muestran distintas metodologías de Aprendizaje Máquina para la exploración y análisis de ciertas características de canciones de la plataforma de *spotify*. Se obtienen resultados de R^2, al tratar de predecir la popularidad de una canción, que van desde valores negativos hasta 0.67. El mejor método para predicción es el de bosque aleatorio, seguido de LGBM. El peor resultó ser la regresión lineal. Palabras clave: Spotify, Aprendizaje Máquina, Popularidad.

1. Introducción

En los últimos seis años la industria de la música ha presentado crecimiento, contrario al comportamiento que se había visto desde el comienzo del siglo como se pueda observar en la referencia [1]. A partir del año 2011 los números que dicha industria presentaba empezaron a mantenerse constantes constantes y luego fueron de subida, contrario a lo que venía sucediendo; lo que sucedió ese año y vino a revolucionar la manera en que se escuchaba música fue la salida de *Spotify* al mercado.

Spotify es una plataforma que contiene millones de canciones que el usuario puede escuchar sin la necesidad de comprar cada una por separado, hasta es posible usar esta plataforma de manera gratuita con ciertos detalles como tener que escuchar un comercial cada cierto tiempo y ciertas limitantes más. Si se paga una cuota mensual es posible tener acceso a las millones de canciones con las que cuenta esta plataforma sin ninguna limitación ni tener que escuchar comerciales. Actualmente hay más de 150 millones de usuarios que pagan dicha suscripción, que forma parte de los más de 345 millones de usuarios activos mensuales [2]. Además de las tantas canciones que se pueden escuchar por un precio tan bajo a comparación de cómo se consumía la música anteriormente, una ventaja enorme de *spotify* es que no es necesario descargar ninguna canción, ya que se está emitiendo gracias a la velocidad del internet.

Con estos datos en mente, *spotify* dominó el mercado por años de la nueva forma de consumir música, pero poco a poco empezaron a surgir más plataformas como esta y se tenía que innovar para sobrevivir, es por esto que tal como sucede con las plataformas de transmisión de video como *netflix*, *youtube*, etc. *spotify* tomó el camino del análisis de los datos de sus usuarios para brindar un mejor servicio y hasta poder hacer recomendaciones al usuario. *Spotify* además puso a disponibilidad del público su API para que los usuarios también pudieran hacer análisis , exploración y predicción de los datos que recaba dicha plataforma [3]. Esta API ofrece características de cada canción, artista, género y año para que se pueda tener un estándar en el análisis que se puede realizar.

Para el presente trabajo se utilizó PCA (*Principal Component Analysis*) [4], LDA (Linear Discriminant Analysis) [5] y SVD (Singular Value Decomposition) [6] como métodos no supervisados para reducción de dimensionalidad. Como métodos supervisados para el modelado de datos se utilizó regresión lineal [7], regresión polinomial [8], k-vecinos [9], árboles de decisión [10], bosque aleatorio [11] GBT (Gradient Boosting Trees) [12] y LGBM (Light Gradient Boosting Machine) [13].

La ventaja que tiene LGBM sobre otros algoritmos es que es rápido y tiene un manejo adecuado de cantidades grandes de datos, entre más, mejor. Su principal diferencia respecto a otros métodos basados en árboles es su crecimiento vertical, que es lo que lo hace rápido, sin embargo tiene cierta tendencia al sobreajuste.

2. Datos

Los datos utilizados para este trabajo se obtuvieron de un repositorio en *kaggle*[14], que a su vez se obtuvieron desde la API de spotify llamada spotipy, ya que está construida en python. se da una breve explicación de las variables utilizadas a continuación.

- Acústica (Acousticness) Variable numérica de confianza que va de 0 a 1 y señala que tan acústica es una canción, una puntuación de 1.0 indica que una canción tiene altas posibilidades de ser acústica
- Artistas (Artists):
  Columna de datos de tipo "Stringçon el nombre del artista que interpreta la canción.
- Capacidad de baile o bailabilidad (danceability): Variable numérica entre 0 a 1, describe qué tan adecuada

es una pista para bailar en función de una combinación de elementos musicales que incluyen el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general. Entre mas cercano sea a uno el valor de una canción mas bailable es.

- Duración en milisegundos(Duration ms): Variable numérica con la duración de la canción en milisegundos.
- Energía (Energy): Variable que cuantifica una medida perceptiva de intensidad y actividad. Las canciones con un valor cercano a 1 tienden a tener un tempo rápido y ruidoso mientras que las canciones con un numero cercano a 0 son mas tranquilos.
- Explicito (Explicit): Variable que expresa si una canción cuenta con palabras consideradas antisonantes u ofensivas en donde si se tiene un valor de 0 la canción no es explicita y con un valor de 1 la canción si tiene contenido explicito.
- ID: ID de identificación único asignado por spotify que involucra tanto variables numéricas como alfabéticas del tipo string.
- Instrumentalidad (Instrumentalness): Variable que cuantifica el nivel de voces dentro de una canción es decir si la variable se acerca a 0 la canción tiene muchas voces y entre mas cerca este de 1 mas instrumental es, la canción se compone solo de instrumentos
- Llave (Key): La clave general estimada de la pista. Los enteros se asignan a los tonos utilizando la notación estándar de clase de tono . Por ejemplo, 0 = C, 1 = C / D , 2 = D, y así sucesivamente.
- Vivacidad (liveness): Variable numérica que refleja la presencia de una audiencia en la grabación. Los valores mas cercanos a 1 representan una mayor probabilidad de que la pista se haya interpretado en vivo.
- Volumen (loudness): Es una variable numérica entre 0 y 1, La sonoridad es una variable numérica general es el promedio de decibelios (dB), se toman los valores de decibelios durante toda la canción y se promedian ese promedio es el que se ve reflejado en el valor.
- Volumen (loudness): Variable numérica entre 0 y 1 que representa el contenido melódico dentro de una canción entre mas cercano a uno sea mas melódica es la canción.
- Nombre (Name): Variable de tipo string que representa el nombre de la canción.
- Popularidad (Popularity): Variable que cuantifica que tan popular es una canción del 0 al 100 donde 0 es muy impopular y 100 es muy popular sin tomar en cuenta el año
- Fecha de lanzamiento(release date): Fecha en que fue lanzada la canción algunas cuentan con el formato e información de año/mes/día aunque algunas canciones solo cuentan con la información del día
- Discurso (speechiness): Variable numérica entre 0 y 1,si el habla de una canción es superior a 0,66, probablemente esté compuesta de palabras habladas, una puntuación entre 0,33 y 0,66 es una canción que puede contener tanto música como palabras, y una puntuación inferior a 0,33 significa que la canción no tiene ningún habla.
- Tempo: Variable numérica entre 0 y 1 que representa la velocidad de la canción entre mas cercano a 0 mas tranquila es la canción y entre mas cercano a uno mas rápida es la canción
- Positividad (Valence): Variable numérica entre 0 y 1 que describe la positividad musical que transmite una pista. Las pistas con valencia alta suenan más positivas (por ejemplo, feliz, alegre, eufórico), mientras que las pistas con valencia baja suenan más negativas (por ejemplo, triste, deprimido, enojado)
- Año (Year): Variable numero que indica el Año de lanzamiento de la canción.

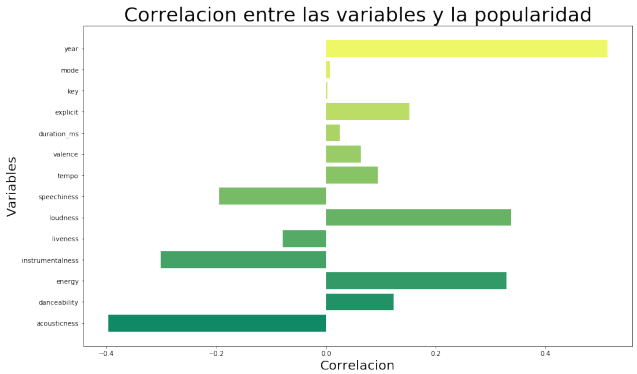

FIGURA 1: Gráfica que representa la correlación entre las distintas variables y la popularidad

En la figura uno observamos que las variables mas correlacionadas con la popularidad son el año, la sonoridad y la energía lo que nos dice de primera mano que las canciones mas populares son aquellas con niveles de decibelios altos y que tienen mucha energía es decir canciones ruidosas y rápidas que mayormente tienden a ser canciones para bailar La figura uno ilustra el numero de canciones que utilizan una llave (nota) como nota general. En la figura 3 se gráfica el promedio de popularidad que tuvieron las canciones hechas en un año como podemos observar las canciones mas populares tienden ha ser canciones de años recientes lo cual es bastante razonable ya que la mayoria de los usuarios de spotify son gente joven que disfruta de las canciones recientes por lo que las canciones antiguas no son muy populares. En la figura 4 podemos ver como evolucionan los componentes de las canciones como el ritmo o la instrumentación con respecto a las nuevas canciones que salen año con año.

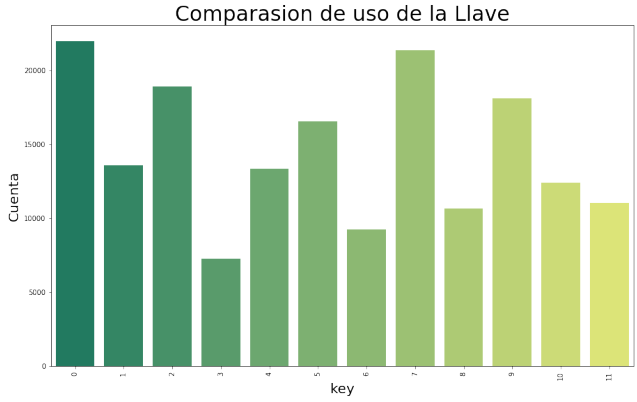

FIGURA 2: Número de canciones que usan una determinada llave

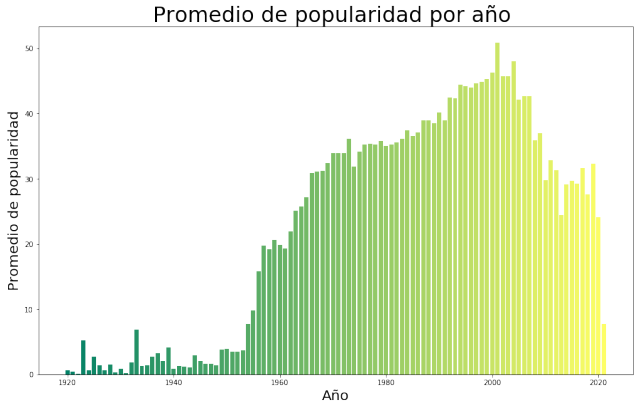

FIGURA 3: Gráfica que muestra la popularidad promedio de las canciones emitidas ese año

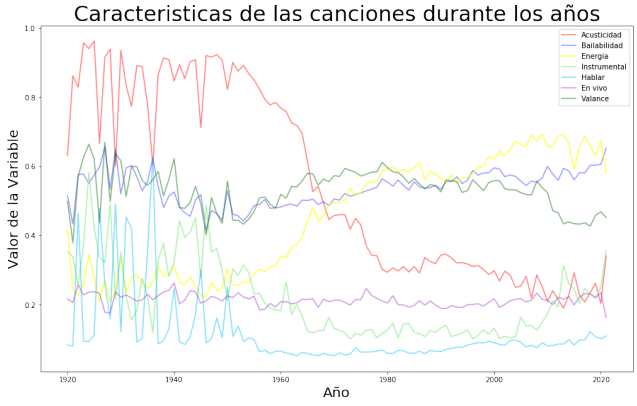

FIGURA 4: Figura que muestra el comportamiento de los elementos con respecto a las canciones que salen cada año

Podemos observar dos cosas bastante interesantes y es que las canciones modernas son cada vez menos acústicas y la diferencia es enorme si las comparamos con las canciones mas antiguas ya que estas tienen altos niveles de acusticidad, a medida que van pasando los años las canciones cada vez tienen una mayor energía es decir son cada vez mas ruidosas y rápidas.

3. Metodología

Los datos se dividieron en dos conjuntos, uno de prueba y otro de "no prueba".

En la primera parte realizó una reducción de dimensionalidad para obtener visualizaciones de los datos así como menos variables a la hora de probar modelos. Los datos se redujeron a tres variables y se probaron tres métodos: Análisis de componentes principales, descomposición de valores singulares y análisis lineal discriminante.

Se probaron los siguientes modelos

- Regresión lineal
- Regresión polinomial
- K - vecinos
- Árbol de decisión
- Bosque aleatorio
- Gradient Boosting Tree
- Light Gradient Boosting Machine

Para cada modelo, se realizó una validación cruzada de cuatro pliegues utilizando el conjunto de no prueba para obtener así un conjunto de entrenamiento y uno de validación para cada pliegue.

Posteriormente se intentó el procedimiento descrito anteriormente pero estandarizando los datos antes de ser reducidos. También se hizo otro intento en el que no se hacía reducción de dimensionalidad.

Posteriormente se realizó búsqueda de hiperparámetros para los modelos que se consideraron con potencial. En algunos modelos se hizo la búsqueda exhaustiva en una malla de hiperparámetros (grid search)[15] y en otros se realizó una búsqueda aleatoria (randomized search)[16].

4. Resultados y análisis

La reducción a tres dimensiones con PCA explicaba 46 % de la varianza. Las correlaciones entre las variables y las primeras dos componentes se puede apreciar en la figura 5.

Para evaluar los modelos, se utilizó el coeficiente de determinación r^2 como métrica en el conjunto de prueba.

En la tabla 1 se pueden ver los desempeños de los modelos con el enfoque de realizar la reducción de dimensiones.

En busca de mejorar resultados, se realizó la estandarización antes de la reducción. Los resultados de dicha modificación se pueden ver en la tabla 2.

Se podría pensar que la estandarización no es útil, pero la situación realmente es la siguiente

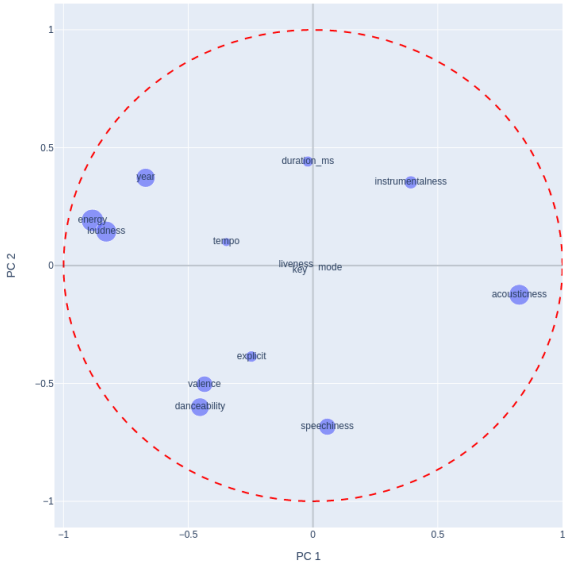

FIGURA 5: Correlaciones entre variables y las primeras dos componentes

	PCA	SVD	LDA
Regresión lineal	0.273	0.271	0.368
Regresión polinomial	0.304	0.395	0.407
K - vecinos	0.283	0.282	0.455
Árbol de decisión	0.49	0.302	0.464
Bosque aleatorio	0.525	0.48	0.473
GBT	0.471	0.27	0.448
LGBM	0.05	-6.169	0.038

CUADRO 1: Resultados de los modelos con reducción de dimensionalidad

	PCA	SVD	LDA
Regresión lineal	0.23	0.23	0.368
Regresión polinomial	0.33	0.326	0.407
K - vecinos	0.296	0.296	0.455
Árbol de decisión	0.329	0.312	0.464
Bosque aleatorio	0.314	0.317	0.473
GBT	0.38	0.261	0.448
LGBM	0.027	-6.767	0.038

CUADRO 2: Resultados de los modelos con estandarización y reducción de dimensionalidad

- Si no se hace la estandarización, resulta que la variabilidad se la lleva principalmente la variable de duración, porque está en milisegundos y eso hace que se tengan valores grandes en comparación con las otras. Con una sola componente ya se tenía el 99 % de la variabilidad
- Si se realiza la estandarización, se resuelve el problema anterior, pero el problema es que con 3 componentes solamente se consigue explicar el 46 % de la varianza, lo cual puede explicar porque "no está ayudando" la estandarización.

Dicho lo anterior, se realizó la última prueba de solamente estandarizar las variables. Los resultados se pueden ver en la tabla 3.

	R^2	Tiempo [s]
Regresión lineal	0.368	0.957
Regresión polinomial	0.591	13.3
K - vecinos	0.539	165
Árbol de decisión	0.601	2.98
Bosque aleatorio	0.671	111
GBT	0.603	18.1
LGBM	0.636	1.56

CUADRO 3: Resultados de los modelos con estandarización

Respecto a la búsqueda de hiperparámetros se llegó a lo siguiente

- Random forest
  - RandomizedSearch
  - n_estimators = 100
  - min_samples_leaf = 1
  - max_depth = 15
  - bootstrap = True
  - R^2 = 0.676
- LGBM
  - RandomizedSearch
  - boosting_type = dart
  - learning_rate = 1.124
  - max_depth = 5
  - min_ data = 35
  - n_ estimators = 79
  - num_leaves = 100
  - sub_feature = 0.999
  - R^2 = 0.636

Ya que el mejor modelo fue el bosque aleatorio, se obtuvieron los coeficientes de importancia, los cuales se ven en la tabla 4.

característica	importancia
year	0.534975
instrumentalness	0.078800
loudness	0.050508
duration_ms	0.047704
energy	0.038958
acousticness	0.037943
danceability	0.036465
liveness	0.036304
valence	0.035960
speechiness	0.034949
tempo	0.034516
key	0.015117
explicit	0.014658
mode	0.003144

CUADRO 4: Importancia de características

name	popularity	year
drivers license	100	2021
Mood (feat. iann dior)	96	2020
positions	96	2020
DÁKITI	95	2020
BICHOTA	95	2020
34+35	94	2020
Whoopty	94	2020
WITHOUT YOU	94	2020
Therefore I Am	94	2020
LA NOCHE DE ANOCHE	94	2020

CUADRO 5: Top 10 canciones

5. Conclusiones

Respecto a la popularidad, si revisamos las canciones mas populares (tabla 5), vemos que las canciones más populares son de los últimos años.

Considerando que el año es determinante y con ayuda del gráfico de la figura 5, se puede decir que las canciones tienden a ser más populares si son mas animadas, intrépidas, es decir, que en promedio se tengan amplitudes más grandes y que sean intensas, rápidas, etc.

Se observa que los modelos no presentan resultados mejores a 0.7 pero que sí subieron considerablemente al hacer la búsqueda de los mejores hiperparámetros con validación cruzada y que una gran diferencia que hay entre los dos mejores es el tiempo. Se recomienda el uso de metodologías más robustas para tareas de regresión.

Referencias

[1] https://www.statista.com/statistics/272305/global-revenue-of-the-music-industry/

[2] https://newsroom.spotify.com/company-info/

[3] Spotify, (2021). Web api reference. Recuperado de https://developer.spotify.com/documentation/web-api/reference/

[4] scikit-learn developers, (2020). LinearDiscriminantAnalysis. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[5] scikit-learn developers, (2020). LinearDiscriminantAnalysis. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html

[6] scikit-learn developers, (2020). TruncatedSVD. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

[7] scikit-learn developers, (2020). LinearRegression. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[8] scikit-learn developers, (2020). PolynomialFeatures. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

[9] scikit-learn developers, (2020). KNeighborsRegressor. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

[10] scikit-learn developers, (2020). DecisionTreeRegressor. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[11] scikit-learn developers, (2020). RandomForestRegressor. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[12] scikit-learn developers, (2020). GradientBoostingRegressor. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

[13] Microsoft Corporation, (2021). LGBMRegressor. Recuperado de https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html

[14] Kaggle, (2021). Spotify Dataset 1921-2020, 160k+ Tracks. Recuperado de https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

[15] scikit-learn developers, (2020). GridSearchCV. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[16] scikit-learn developers, (2020). RandomizedSearchCV. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html