Maquinita, maquinita…¿tendrá éxito mi temita?

Machine learning aplicado a la predicción de la popularidad de canciones en función de sus letras y características musicales

Olga Sanz de Sousa

Silvia Yáñez López

Proyecto final de bootcamp en Data Science y Machine Learning

Contenido

1. Datos generales del proyecto

Este proyecto ha sido realizado por Olga de Sanz de Sousa y Silvia Yáñez López como proyecto final del **Bootcamp en Machine Learning y Data Science** de IDBOOTCAMPS (promoción septiembre de 2023).

El título del proyecto es: *Machine learning aplicado a la predicción de la popularidad de canciones en función de sus letras y características musicales*.

2. Beneficiarios del proyecto

Este proyecto aborda el desafío de comprender y prever la recepción del público hacia las canciones. En la actualidad, la industria musical se enfrenta a la saturación de contenido, y los artistas y productores buscan métodos efectivos para destacar en un mercado abrumador. Este proyecto propone una solución al emplear algoritmos de aprendizaje automático que analizan tanto las letras como las características musicales, permitiendo a los profesionales de la música tomar decisiones informadas sobre la producción y promoción de canciones. Al resolver este problema, el proyecto trata de ofrecer una herramienta valiosa para artistas emergentes, casas discográficas y profesionales de la industria, mejorando sus estrategias de lanzamiento y aumentando las posibilidades de éxito en un entorno competitivo.

Este proyecto está dirigido a músicos, productores, y expertos de la industria musical que buscan maximizar el impacto de sus creaciones. La capacidad de prever la popularidad de una canción antes de su lanzamiento, así como comprender el efecto que tienen distintas variables sobre la misma, proporciona una ventaja estratégica significativa, permitiendo ajustes finos en la producción, promoción y distribución. Además, esta herramienta puede resultar muy valiosa para plataformas de streaming y servicios de música, ya que les brinda la capacidad de ofrecer recomendaciones más personalizadas a sus usuarios, mejorando la experiencia global.

3. Ejecución del proyecto

El proyecto sintetiza el uso de diferentes técnicas de recolección y análisis de datos, así como el uso de machine learning para la creación de modelos predictivos.

En primer lugar, se obtuvo un dataset de la web de Kaggle, en el que había datos sobre 32832 canciones extraídas de Spotify, incluida la popularidad de las mismas.

Para poder abordar el problema que se quería analizar se obtuvieron además los siguientes datos:

- **Idioma de las canciones**: A través de la API de MusixMatch. En base a los datos obtenidos en esta consulta se seleccionaron solo las canciones en inglés del dataset (12890)

- **Letra de las canciones:** Tras eliminar duplicados, se obtuvieron 8072 letras de canciones mediante *web scraping* de la página de genius.com.
- **Popularidad del artista:** A través de la API de Spotify se obtuvo la popularidad de los artistas de las 8072 canciones seleccionadas.

Una vez completado el dataset, se realizó un **análisis exploratorio de los datos (EDA)**, visualizando las variables del problema y eliminando aquellas que desbalanceaban en exceso los datos para predecir la popularidad. Tras este análisis el dataset mantuvo 7637 canciones.

Posteriormente se realizó el **procesamiento del lenguaje natural (NLP)** con las letras de las canciones, obteniendo en primer lugar sus características estructurales gracias a la codificación proporcionada por la web de letras genius.com (estribillos, estrofas, puentes, intro, outro...etc.), convirtiéndolas en variables para incluir en el modelo. En segundo lugar, se decidió dividir el análisis de las letras en dos vertientes: por un lado, se analizaron las letras completas tras una limpieza de las mismas (caracteres indeseados, y stopwords), y por otro lado se extrajeron los estribillos de las mismas para analizarlos por separado. Con las letras completas se realizó en primer lugar un análisis de sentimiento, y posteriormente, debido a la longitud de los textos, se extrajeron las siguientes variables para su procesamiento posterior en el modelo:

- Riqueza léxica con y sin stopwords
- 3 Palabras más y menos frecuentes, y su frecuencia
- Longitud de las canciones
- Uso de lenguaje explícito con el transformer BERT

Con los estribillos se realizó un topic labelling mediante LDA, clasificando a los mismos en 3 temáticas, que fueron añadidas como features para el modelo.

Finalmente, se entrenó un modelo de **Extreme Gradient Boosting (XGboost)** de tipo regresor, para tratar de predecir la popularidad de las canciones en función de las features disponibles. Se utilizó cross validation para evitar el overfitting, así como un tuning de hiperparámetros con grid search. En primer lugar, se realizó un modelo baseline, utilizando solo las variables de las características musicales; posteriormente se incluyeron todas las variables obtenidas en el NLP, y, por último, se eliminaron aquellas que no aportaban información al modelo, con el objetivo de simplificarlo y así optimizar su funcionamiento.

4. Impacto y resultado

Los resultados obtenidos en los modelos entrenados se presentan en la tabla 1, y en los anexos se incluyen las gráficas representando la precisión de los modelos y sus errores.

Se observa que no hay suficiente evidencia para proporcionar un modelo robusto, quizá por el número de instancias, o por el análisis concreto del lenguaje que se ha realizado. También es posible que haya variables extrañas ligadas a la popularidad que no se hayan tenido en cuenta, como por ejemplo la productora musical (que puede influir en la promoción de la canción y el impacto que esta tenga sobre el público) o variables subjetivas propias de los consumidores de Spotify.

Tabla 1

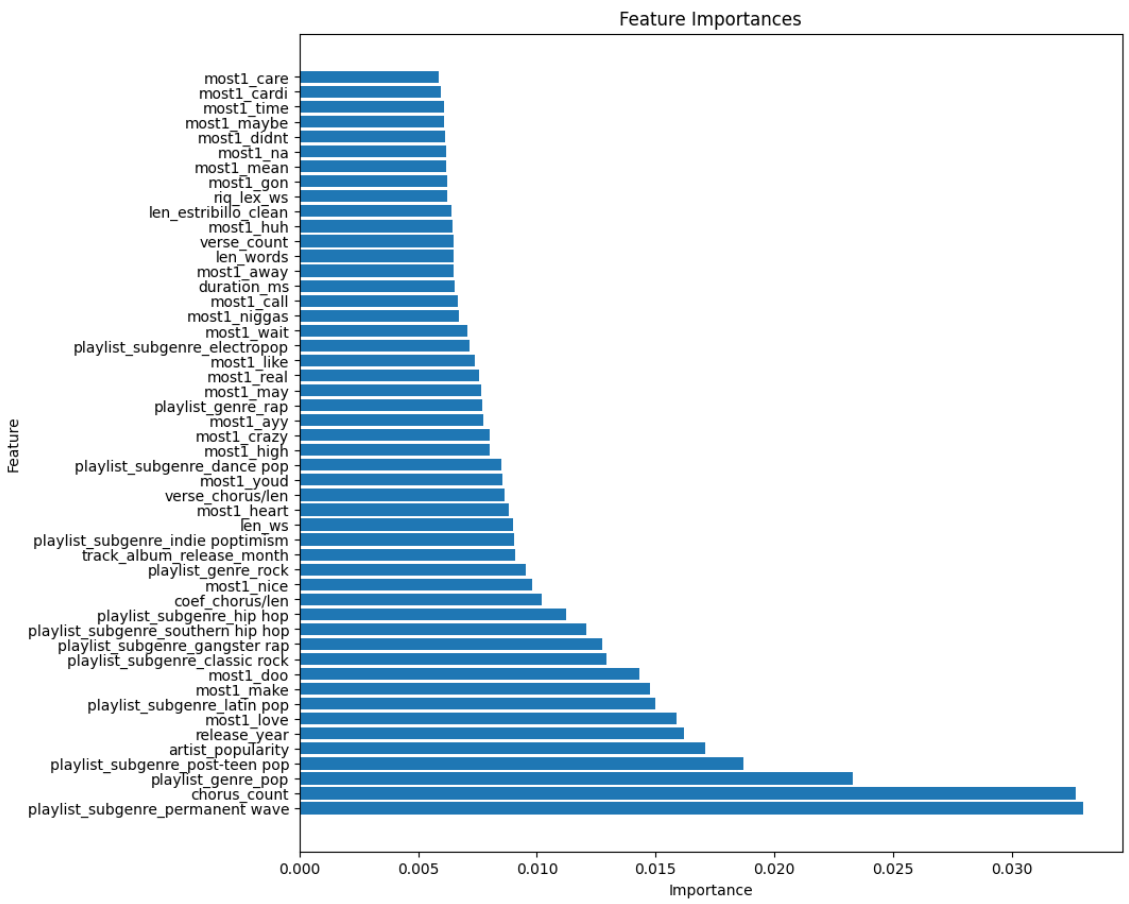
Resultados de los modelos

	MSE	MAE	R2
XGBoost sin lyrics	390.6696	15.5382	0.160184
XGboost con lyrics	369.3508	14.8910	0.206013
XGboost eliminando columnas con feature importance = 0	368.5392	14.8631	0.207758

No obstante, se ha observado que la inclusión de las variables obtenidas a través del procesamiento de las letras de las canciones mejora el modelo, y que, además, estas se encuentran entre las features que más importancia tienen (ver Figura 1), por lo que una exploración más profunda que pudiera ahondar en otras variables podría proporcionar un mayor insight respecto a la relación de las letras y la popularidad de las canciones, así como modelos de machine learning más precisos.

Figura 1

Feature importances del modelo de XGBoost final



5. Líneas de continuidad

En base a los resultados obtenidos no se ha encontrado un modelo que prediga con precisión la popularidad de la canción a través de las variables utilizadas. No obstante, se abren otras vías de exploración que no han podido ser abarcadas en este proyecto, tales como: la utilización de un modelo clasificador en vez de un modelo regresor; la división de las canciones por género musical o por popularidad de los artistas, para dar más peso a las letras sobre otras características; el uso de otras métricas de popularidad como ratings o rankings públicos o de otras aplicaciones; la consideración de las productoras musicales y las campañas de marketing relacionadas con las canciones; o el uso de algoritmos de aprendizaje no supervisado, con el objetivo de encontrar patrones y clústeres no identificados en el análisis exploratorio.

6. Conclusiones

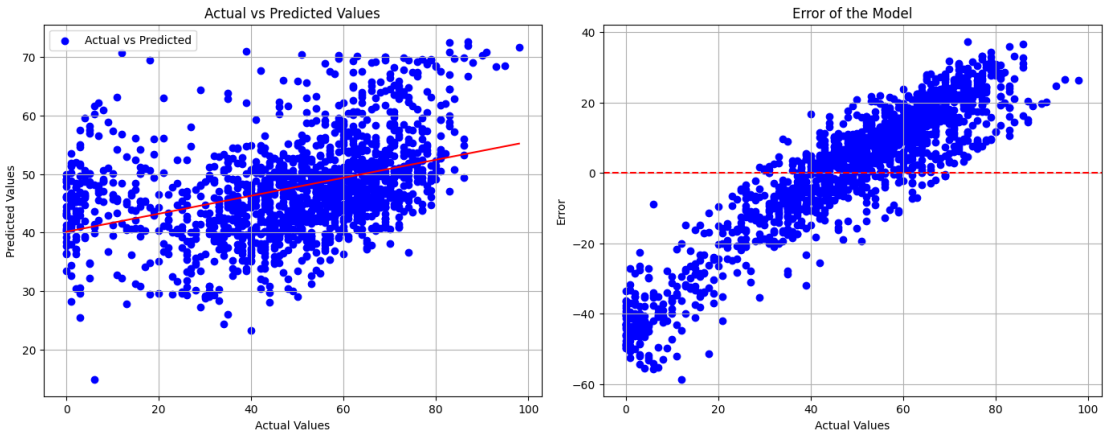
En conclusión, este proyecto representa un desafiante y revelador punto de partida como jóvenes profesionales en el ámbito del análisis de datos y aprendizaje automático. A pesar de los esfuerzos dedicados y la aplicación de metodologías rigurosas, es importante reconocer que el modelo de predicción de popularidad de canciones no ha alcanzado el éxito deseado. Este resultado, lejos de ser motivo de desánimo, proporciona valiosas lecciones sobre las complejidades inherentes a la relación entre las letras y las características musicales y la preferencia del público.

El proceso de construcción y evaluación del modelo ha permitido identificar áreas de mejora y ha planteado preguntas estimulantes sobre la naturaleza de la popularidad musical. Estos desafíos no solo resaltan la realidad intrínseca de la experimentación en ciencia de datos, sino que también subrayan la importancia de la iteración y la adaptación continua. Este proyecto, aunque no haya alcanzado sus metas iniciales, representa un valioso ejercicio de aprendizaje que contribuirá al crecimiento y desarrollo del campo. Además, ofrece una perspectiva realista sobre los desafíos que enfrentan los científicos de datos

7. Anexos

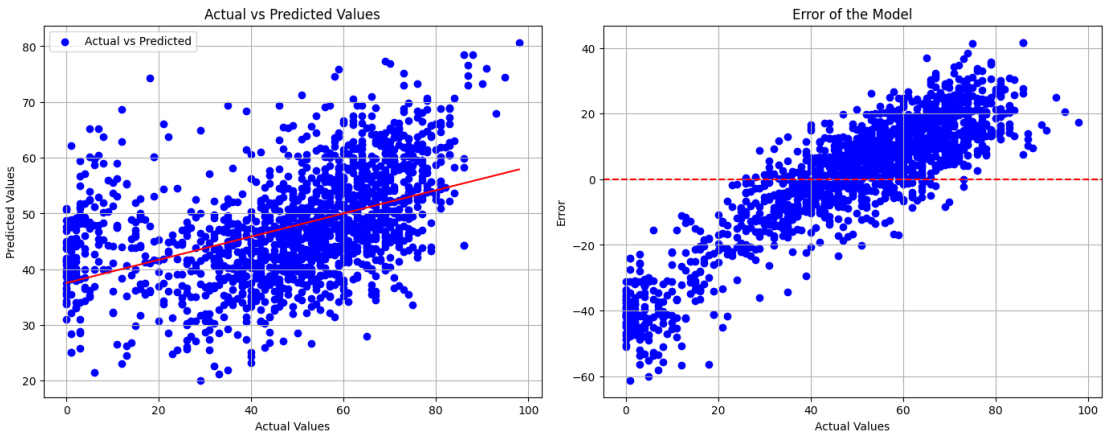
Anexo 1

Precisión y error del primer modelo: XGBoost sin lyrics



Anexo 2

Precisión y error del segundo modelo: XGBoost con lyrics



Anexo 3

Precisión y error del tercer modelo: XGboost eliminando features