Informe Microproyecto 2

Modelos Avanzados de Análisis de Datos 2

Diciembre 13 de 2020

Camilo Andres Suarez Trillos
200321493

Carlos Francisco Silva Ortiz
201920463

Juan Camilo Florez Caro
201620135

Universidad de los Andes

1. Introducción con la definición del problema y la pregunta de investigación:

Una aplicación de música quiere actualizar su aplicación online para que genere recomendaciones a sus usuarios de nuevos artistas para escuchar. El sistema de recomendación debe tomar en cuenta las preferencias de cada usuario, con el fin de ofrecer recomendaciones automáticas y personalizadas.

Por ello se le pide a usted, en calidad de consultor externo, desarrollar un algoritmo de recomendación de artistas para cada usuario. Implemente el algoritmo a partir del conjunto de datos recibido y evalúe su desempeño con respecto a la métrica apropiada.

Una vez ha desarrollado su primer sistema de recomendación, intente mejorarlo con respecto a la métrica de su elección, considerando además la información que encuentra en la segunda base de datos, donde encuentra información de tipo socio-demográfico por usuario.

2. Metodología Propuesta:

Como primer paso, se realizó un análisis exploratorio de los datos para entender la información recibida. Iniciamos revisando un histograma con la cantidad de reproducciones y la cantidad de usuarios según la cantidad de artistas que escuchan:

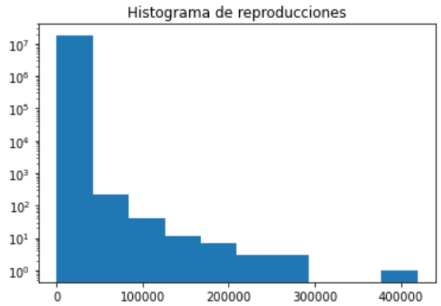

Gráfico 1. Histograma de reproducciones

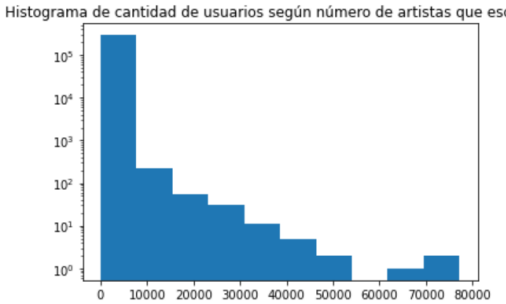

Gráfico 2: Histograma de cantidad de usuarios según cantidad de artistas que escuchan

Observamos que la mayor cantidad de reproducciones se concentran entre 0 - 100.000. Hay algunos pocos registros por debajo de esta cantidad de reproducciones entre artista/usuario. También se evidencia que la mayor cantidad de usuarios (10k aprox) pueden escuchar una gran cantidad de artistas.

Para crear el sistema de recomendación, procederemos a convertir la cantidad de reproducciones de artistas de cada usuario en un puntaje entre 0-5 con decimales, para esto, convertimos por usuario la cantidad de reproducciones en percentiles donde 100 es el artista que más escuchan y 1 el artista que menos escuchan, dejando el 0 libre para catalogar artistas que nunca han escuchado. Esto lo realizamos para cada usuario con el fin de no sobre estimar a aquellos melómanos que escuchan muchísimo un artista vs usuarios ocasionales que escuchan sus artistas preferidos ocasionalmente. Adicionalmente, este “rating” no inicia en 0 sino en 0.83, valor que representa la menor cantidad de veces que un usuario ha escuchado un artista. Posteriormente, debido a la gran cantidad de información (359.347 usuarios y más de 17 millones de observaciones usuario-artista-reproducciones), procedemos a hacer un submuestreo, quedándonos con información de 10.000 usuarios únicamente, esto con el fin de revisar la viabilidad de los modelos y ajustar parámetros que posteriormente servirán para entrenar el modelo con toda la información realizando una validación cruzada de 10 pliegues.

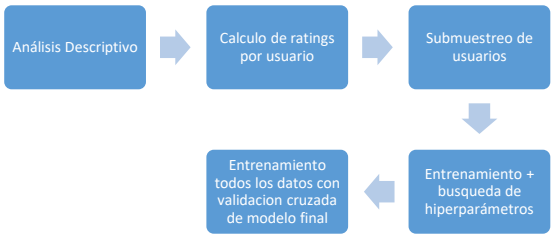

Gráfico 3. Metodología Parte 1

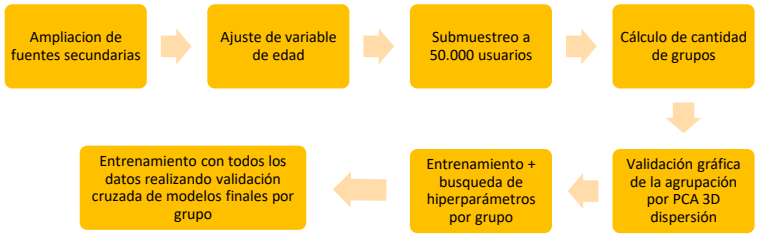

Gráfico 4. Metodología Parte 2

Para la segunda parte, usamos la información demográfica por usuario, la cual contiene información sobre género, edad y país. Gracias a esta última variable, podemos agregar más información como por ejemplo idioma principalmente hablado por país y continente en el cual se encuentra el país del usuario, esto con el fin de ampliar nuestra fuente de secundaria de información. Se revisó el campo de edad y se encontraron valores atípicos como edades extremadamente pequeñas y otras muy elevadas. Se procedió a ajustar las edades por debajo de 14 años y por encima de 86 como valores faltantes para acotar la función de densidad de la variable edad. Se procedió a hacer una agrupación por K medios con un submuestreo de 40 mil usuarios para encontrar una cantidad de grupos ideal mediante la puntuación por siluetas.

Se determinó que 14 grupos son lo suficientemente diversos ya que esta puntuación tuvo una diferencia muy grande vs 15 grupos, el cual sigue cayendo. Una vez determinado el número de grupos, se procedió a realizar la agrupación con toda la base de clientes. Se validó el ejercicio de agrupación realizando un análisis con los 3 primeros componentes principales y graficando esto en un plano de dispersión con 3 ejes para observar gráficamente las agrupaciones. Una vez agrupados todos los clientes, se implementó una búsqueda de grilla de hiperparámetros de los modelos de recomendación para cada uno de los grupos, buscando así minimizar el RMSE del rating. Una vez se obtuvieron los mejores parámetros para cada grupo, se procedió a entrenar nuevamente los modelos pero, esta vez usando toda la información disponible y con validación cruzada con 100 pliegues, evidenciando en estos modelos, una reducción del RMSE vs los modelos anteriores.

3. Resultados:

El modelo usado fue el BaselineOnly, el cual es un algoritmo que estima las predicciones base para cada usuario/item. La principal ventaja de este modelo vs otros modelos para sistemas de recomendación es *la velocidad de entrenamiento* y la facilidad de uso, factores diferenciales en este ejercicio el cual posee una gran cantidad de información y requiere la calibración de múltiples modelos. BaselineOnly requiere 4 parámetros:

- Método: el cual puede ser ALS (Mínimos Cuadrados Alternantes) o SGD (Gradiente Descendiente Estocástico).
- Parámetro de regularización de ítems.
- Parámetro de regularización de usuarios.
- Número de épocas. Parámetro usado solamente si se usa el método ALS ya que es la la cantidad de iteraciones que hará el ALS.

Para el primer modelo, se usaron parámetros usados en ejercicios anteriores que han dado buenos resultados como son: ALS, reg_i=25, reg_u=25 y épocas=25. Entrenando el modelo con 75% de los datos para entrenamiento y dejando un 25% para prueba, se obtuvo un RMSE en prueba de **0.9573** con el submuestreo. Al realizar una búsqueda de grilla de hiperparámetros, estos cambiaron a reg_i=18 y reg_u=2, logrando disminuir el RMSE a **0.9302**. Una leve mejora vs el modelo inicial. Con estos hiperparámetros calibrados, procedemos a realizar un modelo con validación cruzada de 10 pliegues con todos los datos, obteniendo así un RMSE de **0.9254.** Vemos que hemos mejorado el RMSE, no

obstante, no parecen significativas las mejoras vs el modelo inicial. Demostrándonos la complejidad de los datos que requieren agrupaciones y/o modelos más complejos que permitan ajustarse mejor a este tipo de información.

Parámetros\Modelo	Baseline undersampling 75/25	Baseline undersampling 75/25 gridsearch	Baseline full data + 10 Fold CV
RMSE	0.9573	0.9002	0.8854
FCP	0.5372	0.5608	0.5758
Parámetros	ALS Reg Items: 25 Reg Users: 20 Epochs: 20	ALS Reg Items: 18 Reg Users: 2 Epochs: 20	ALS Reg Items: 18 Reg Users: 2 Epochs: 20

Para la segunda parte, no se especificaba que NO se podía agregar más información a partir de la información existente, por lo que se agregó el detalle del idioma oficial más hablado por país como continente. También se realizó un ajuste a los datos de edad, los cuales contenían valores extremos atípicos ligeramente frecuentes, pudiéndose afectar la generación de los grupos

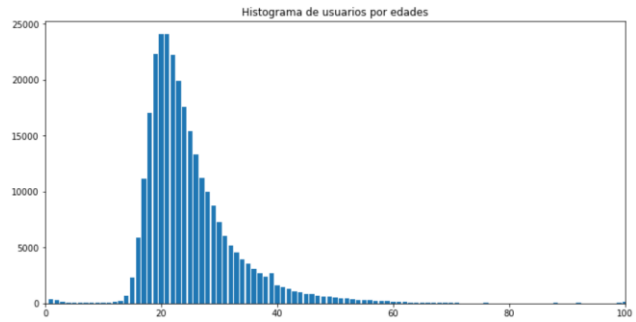

Gráfico 5. Histograma de edades

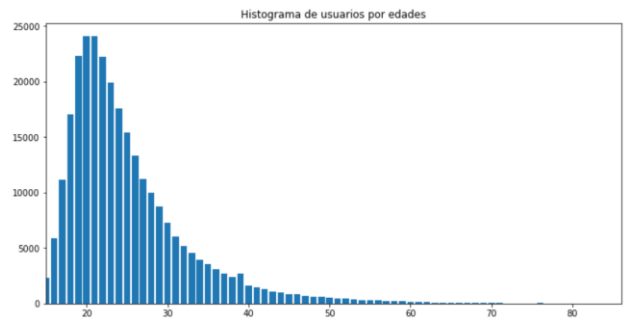

Gráfico 6. Histograma de edades sin valores extremos

Al realizar el undersampling para revisar los gráficos de siluetas por grupos, se evidencia que el grafico de 11 siluetas es el que mejor separa los grupos:

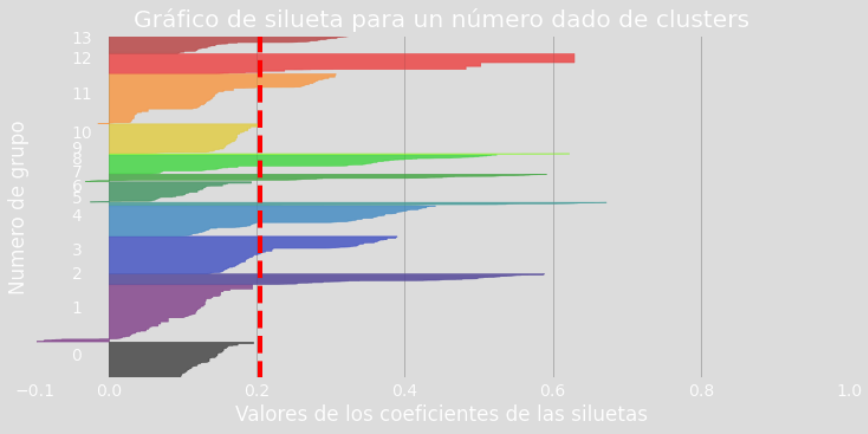

Gráfico 7. Diagrama de siluetas de 11 grupos

Al realizar esta agrupación, validamos que se evidencie una separación entre los usuarios, para lo cual, tomamos toda la base de usuarios, la agrupamos y luego generamos un análisis de componentes principales, usando únicamente los 3 primeros componentes para graficar los usuarios en un plano 3d de dispersión, obteniendo:

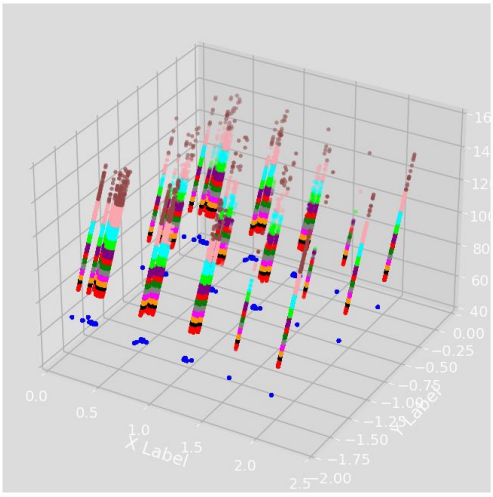

Gráfico 8: Validación gráfica de la agrupación por K medias usando PCA

Se evidencia una separación de los grupos exitosa. Procedemos entonces a realizar un gridsearch de épocas, parámetros de regularización de ítems y de usuarios por grupos con una partición 75% entrenamiento, 25% prueba. Validando la precisión de estos modelos según RMSE, tenemos 2 versiones de esta: RMSE promedio de **0.7144** y RMSE promedio ponderado por cantidad de usuarios por grupos de **0.7312**, logrando así una mejora significativa vs el modelo individual. Con los parámetros óptimos que reducen el RMSE por grupo ya definidos, procedemos a hacer una validación cruzada de 100 pliegues por cada grupo, obteniendo un RMSE promedio de **0.6315** y un RMSE promedio ponderado por cantidad de usuarios por grupo de **0.6472**, el cual es mucho mejor que el **0.7312** obtenido sin la validación cruzada.

Parámetros\Modelo	Baseline con gridsearch por grupo	Baseline con gridsearch y CV por grupo
RMSE Promedio	0.7144	0.6315
RMSE Promedio Ponderado	0.7312	0.6472
FCP Promedio Ponderado	0.6212	0.6285

Se logra mejorar el desempeño de los modelos usando agrupaciones, búsqueda de grilla de hiperparámetros y validación cruzada. Cabe resaltar que, aunque pareciera que los modelos sin la validación cruzada pudiesen tener un ligero mejor desempeño vs los modelos con validación cruzada, este efecto de mejor RMSE puede ser atribuible únicamente a las particiones de prueba, siendo mas robusto el RMSE de validación cruzada.

4. Conclusiones:
   - De acuerdo con ambos puntos, realizar modelos por agrupaciones mejora el desempeño vs un modelo individual con toda la información. Se recomienda a la empresa suministrar más información de los usuarios como, por ejemplo, perfiles de redes sociales asociadas a la cuenta para poder segmentar con más granularidad los usuarios.
   - Obtener una fuente de información adicional con datos complementarios de los artistas como año de inicio de carrera, álbumes lanzados, géneros principales, lugares donde ha realizado conciertos, con el fin de complementar la información inicial. También se sugiere agregar el álbum escuchado por artista, esto debido a que múltiples artistas van cambiando con el tiempo su género interpretado, lo que podría también asociarse a gustos de los usuarios por determinados tipos de géneros musicales.
   - Debido al gran volumen de información, se recomienda diseñar e implementar estos modelos en servidores en la nube especializados para este tipo de análisis (por ejemplo, AWS ML), debido al gran tiempo requerido para calibración y entrenamiento de estos modelos. Con mejores dispositivos, se podrían implementar modelos mas robustos que tal vez mejoren los resultados de los modelos desarrollados en este micro proyecto.

- Se sugiere también suministrar información de reproducción de playlists por usuario que es uno de los usos mas comunes en aplicaciones de streaming, esto con el fin de usar también estas preferencias como insumo para mejorar los modelos actuales.
- Con las agrupaciones realizadas, se podrían realizar sugerencias iniciales a aquellos usuarios recién registrados a partir de los amigos en común que usen la aplicación como también, de la información sociodemográfica registrada. Una vez hayan hecho alguna cantidad de reproducciones iniciales, se pueden incluir en la siguiente calibración regular de los modelos para sugerencias más personalizadas.