Informe Intermedio

IIC3633 - Sistemas Recomendadores

Grupo 11

Integrantes:

Vicente Correa

Alberto Maturana

Mariela Zambrano

3 de noviembre, 2025

Progreso en el desarrollo de la solución propuesta

En el informe anterior se indicó que se exploraría la implementación de un modelo que combine el filtrado colaborativo con características de contenido de los juegos (categorías, mecánicas, etc.). De esta manera se implementó un modelo de DeepFM para predecir si a un usuario le gustará un ítem o no, con el objetivo de ver si la información del contenido puede mejorar las predicciones para usuarios con pocas calificaciones. Se binarizaron las calificaciones de tal forma que si la calificación era de 7 o más se tomaba como una interacción positiva (1), sino, era negativa (0).

Experimentación realizada y evaluación intermedia

Se inició con una línea base evaluando el modelo a nivel individual y luego se extendió a un grupo simulado. Para el caso del modelo individual se entrenó durante 10 épocas sobre una muestra de 1 millón de interacciones, para distinguir entre ítems que le gustarían al usuario y los que no. De esta manera, la evaluación intermedia se dio al evaluar el modelo para usuarios individuales con el conjunto de validación. Se generó una lista de recomendación para un top 10 para cada usuario y se calcularon las siguientes métricas de ranking:

Modelo	Precision@10	Recall@10	nDCG@10
DeepFM (individual)	0.4256	0.8421	0.7528
SVD++ (individual)	0.4312	0.8422	0.7600

El valor de nDCG@10 indica que existe una alta calidad en el orden de las recomendaciones.

Para evaluar las recomendaciones grupales, se crearon 1000 grupos, compuestos por 4 usuarios seleccionados aleatoriamente del conjunto de validación. Se aplicaron 3 estrategias de agregación:

- Estrategia de Promedio (Average): Se calculará el rating promedio predicho por el modelo para todos los miembros del grupo.
- Estrategia de Mínima Miseria (Least Misery): La calificación del grupo para un ítem será la calificación mínima predicha entre todos los miembros, buscan evitar ítems que disgusten a alguien.
- Estrategia de Máximo Placer (Most Pleasure): La calificación del grupo será la calificación máxima predicha, buscando maximizar la satisfacción del miembro más entusiasta.

Para evaluar se consideró un ítem como relevante para un grupo solo si todos los miembros del grupo lo habían calificado positivamente.

Análisis preliminar de los resultados obtenidos

La evaluación de las estrategias de agregación grupal arrojó resultados positivos, superando el baseline del modelo individual. Los resultados para el modelo **DeepFM**, **SVD++**, **Most Popular** y **Random** se encuentran en los anexos 2, 3, 4 y 5 respectivamente.

Como se ve en las tablas, casi todos los modelos evaluados en el escenario grupal obtuvieron un nDCG@10 muy superior al rendimiento individual.

Un hallazgo clave es que el baseline simple Most Popular obtuvo el rendimiento más alto, superando marginalmente a SVD++ y DeepFM. Este alto rendimiento general se explica por la estricta definición de "ground truth" utilizada. Al exigir que un ítem sea relevante sólo si todos los miembros del grupo lo calificaron positivamente, la lista de ítems objetivo se reduce a un conjunto muy pequeño de alto consenso. El hecho de que Most Popular gane sugiere que estos ítems de "consenso total" son, casi por definición, juegos universalmente populares y muy bien calificados, haciendo que un modelo de popularidad sea un competidor extremadamente fuerte.

Incluso el modelo Random obtiene un puntaje inesperadamente alto. Esto no indica que un ranking aleatorio sea efectivo, sino que probablemente es un artefacto de la metodología de evaluación: al reordenar una lista de ítems ya calificados por el grupo, la probabilidad de acertar en la pequeña lista de consenso es alta.

Esto demuestra que la solución de agregación es viable, pero también revela que, bajo esta estricta definición de consenso, los modelos avanzados no logran diferenciarse significativamente de un baseline simple como Most Popular.

Adicionalmente, mostraremos distintos casos de usuarios, grupos de usuarios y sus recomendaciones sobre cómo funciona los modelos DeepFM y SVD++.

Por ejemplo para el usuario 235 se muestran las primeras 10 recomendaciones en el Anexo 6 y 7 según su puntaje asignado. Para un grupo de usuarios (ids: [235, 236, 237, 238]) son las siguientes recomendaciones. Vemos que al haber tanta variedad de items, los puntajes tienen diferencias mínimas. Aún así, es curioso ver cómo los juegos más populares (según los autores por lo que ellos conocen de su limitado tiempo jugando juegos de mesa) son los que se terminan recomendando el modelo por más que no sea un modelo Most Popular.

Problemas identificados durante el proceso

No se pudieron procesar todos los datos del dataset porque al tener 12 millones de interacciones, no se puede procesar y entrenar en “Google Colab”. Es por esto que se optó por trabajar con una muestra aleatoria de 1 millón.

Además, se puede ver que hubo un sobreajuste, ya que si bien el rendimiento mejoraba para el entrenamiento, para el caso del conjunto de validación hubo una baja en el rendimiento, lo que indica que el modelo comenzaba a memorizar los datos del entrenamiento, en vez de generalizar, lo cual podría afectar en la calidad de las predicciones.

También, se consideraron usuarios con 10 o más interacciones en el conjunto de validación. Esto excluye a los usuarios “cold-start” o “menos activos”, lo que implica que no se mide cómo se comporta con grupos que incluyan este tipo de usuarios que ayuden a simular condiciones del mundo real.

Revisión del plan propuesto en etapa anterior y justificación de ajustes

1. Implementación de Modelos Avanzados:

Se cumplió. Se implementó el modelo híbrido DeepFM, tal como estaba planificado, usando tanto características colaborativas como de contenido. Se lograron buenos resultados individuales (nDCG@10 ≈ 0.75).

2. Desarrollo de Estrategias de Agregación Grupal:

Se implementaron las tres estrategias propuestas —Average, Least Misery y Most Pleasure— y las evaluamos en 1000 grupos simulados. Los resultados fueron excelentes (nDCG@10 > 0.95), superando el rendimiento individual.

3. Diseño de la Evaluación Grupal:

Ejecutado según el plan. Se crearon grupos sintéticos a partir del conjunto de validación y se usaron métricas de ranking (Precision@10, Recall@10, nDCG@10) como estaba previsto.

4. Cronograma

No se cumplió el cronograma y se debieron tomar ciertos ajustes necesarios para cumplir con las fechas definidas. Sin embargo, la predicción y asignación de tiempos era acertada.

5. Ajustes

En general, el desarrollo avanzó de acuerdo con lo planificado en la propuesta, logrando implementar el modelo híbrido DeepFM y las estrategias de agregación grupal con resultados sobresalientes. No obstante, se identificó la necesidad de incorporar modelos existentes en la literatura de recomendación grupal, con el fin de contar con baselines más sólidos y evaluar con mayor rigor el desempeño de nuestra propuesta. Esta mejora permitirá, en etapas posteriores, comparar el modelo desarrollado con enfoques consolidados y determinar con mayor precisión si efectivamente ofrece beneficios frente a soluciones de referencia. Además, los resultados obtenidos en esta entrega revelaron que la definición actual de *ground truth* tiende a favorecer baselines simples, como **Most Popular**. Por ello, para la etapa final se ampliará el esquema de evaluación, incorporando una definición de relevancia más flexible que permita un análisis más robusto sobre la capacidad de los modelos para recomendar ítems de nicho y medir la satisfacción grupal desde múltiples objetivos.

Bibliografía

Honda, H., Kagawa, R., & Shirasuna, M. (2022). On the round number bias and wisdom of crowds in different response formats for numerical estimation. Scientific reports, 12(1), 8167. https://doi.org/10.1038/s41598-022-11900-7

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. Communications of the ACM, 52(10), 144–147. https://doi.org/10.1145/1562764.1562800

Wadkins, J. (2021). *Board Games Database from BoardGameGeek* [Data set]. Kaggle.https://www.kaggle.com/datasets/threnjen/board-games-database-from-boardgamegeek

Chen, T., Yin, H., Long, J., Nguyen, Q. V. H., Wang, Y., & Wang, M. (2022). Thinking inside The Box: Learning Hypercube Representations for Group Recommendation. En *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery. https://arxiv.org/pdf/2204.02592

Wu, X., Xiong, Y., Wu, Y., Chen, Y., Yu, P. S., & Lai, J.-H. (2023). ConsRec: Learning Consensus Behind Interactions for Group Recommendation. En *Proceedings of the ACM Web Conference 2023 (WWW '23)*. Association for Computing Machinery. https://arxiv.org/pdf/2302.03555

Anexos

Se investigaron modelos del estado del arte diseñados para recomendación grupal, tales como CubeRec (Chen, 2022) y ConsRec (Wu, 2023).

El estudio de CubeRec enuncia que representar las preferencias de un grupo como un único punto (embedding) es un cuello de botella y no refleja el "intervalo" que existe para negociar entre los miembros del grupo. Para solucionar esto, propone una forma más realista de negociación de preferencias. Dada esta prometedora línea de investigación, para la entrega final del proyecto se planea implementar un modelo inspirado en las ideas de CubeRec, buscando modelar la preferencia grupal como un hipercubo en lugar de un punto, para compararlo con nuestras estrategias de agregación actuales.

Por otro lado, ConsRec se enfoca en el consenso que alcanza un grupo al final de un acuerdo, y para ello integran perspectivas de miembros, ítems y del grupo para modelar.

Anexo 1: Investigación del estado del arte.

Estrategia	Precision@10	Recall@10	nDCG@10
Average	0.8872	0.1420	0.9609
Least misery	0.8879	0.1421	0.9609
Most Pleasure	0.8745	0.1399	0.9555

Anexo 2: Resultados grupales para DeepFM

Estrategia	Precision@10	Recall@10	nDCG@10
Average	0.9039	0.142668	0.965189
Least misery	0.9033	0.142596	0.965215
Most Pleasure	0.9065	0.143025	0.967000

Anexo 3: Resultados grupales para SVD++

Estrategia	Precision@10	Recall@10	nDCG@10
Average	0.9017	0.143909	0.967632
Least misery	0.9017	0.143909	0.967632
Most Pleasure	0.9017	0.143909	0.967632

Anexo 4: Resultados grupales para Most Popular.

Estrategia	Precision@10	Recall@10	nDCG@10
Average	0.4644	0.187787	0.734357

Least misery	0.4641	0.187695	0.734323
Most Pleasure	0.4641	0.187562	0.730359

Anexo 5: Resultados grupales para Random.

userID	itemID	name	score
235	1123	Catan	0.9842
235	5061	Carcassonne	0.9725
235	3180	Ticket to Ride	0.9698
235	879	7 Wonders	0.9621
235	1407	Dominion	0.9593
235	10456	Azul	0.9548
235	421	Terraforming Mars	0.9507
235	6612	Pandemic	0.9475
235	9031	Splendor	0.9440
235	2332	Dixit	0.9416

Anexo 6: Recomendación individual para usuario (DeepFM)

itemID	name	score_grupo
5061	Carcassonne	0.9712
3180	Ticket to Ride	0.9658
879	7 Wonders	0.9630

10456	Azul	0.9585
1407	Dominion	0.9573
2332	Dixit	0.9549
6612	Pandemic	0.9520
421	Terraforming Mars	0.9504
1033	Codenames	0.9462
2078	Everdell	0.9447

Anexo 7: Recomendación para un grupo (DeepFM)