Fidelización de clientes para MovistarPlus y recomendador para Netflix en el mercado español

LUIS ARDÉVOL MESA, MATÍAS FERNÁNDEZ LAKATOS, RICHARD RÍO JUAN, ANTÍA VARELA REY

La llegada de Netflix a España nos trajo a DataMasters la entrada de varios proyectos con empresas implicadas en este mercado, concretamente, Movistar y Netflix. MovistarPlus busca prevenir la pérdida de clientes antes la creciente competencia de plataformas como Netflix y las ofertas de Vodafone, . Para ello, esta empresa nos proporciona un conjunto de datos históricos para identificar perfiles de clientes propensos a abandonar la empresa, segmentarlos en grupos clave y diseñar una campaña de incentivos personalizada que ayude a evitar la posible fuga de clientes atraídos por la nueva oferta.

Netflix quiere saber qué captación de mercado puede hacer en solitario. Para ello, esta empresa nos contrata para adaptar sus sistemas de recomendación al público español. La empresa es consciente de que los gustos y hábitos de consumo pueden ser diferentes de un país a otro, y sabe que sus sistemas actuales pueden no ser lo suficientemente precisos en el mercado español. Dejan a nuestro cargo el diseño de un nuevo sistema de recomendación y la validación de un sistema de recomendación entrenado con datos "españoles"

CONTENTS

1. ESTRATEGIA DE FIDELIZACIÓN DE CLIENTES PARA MOVISTARPLUS ANTE LA COMPETENCIA DE NETFLIX Y VODAFONE

A. Análisis de perfiles de posibles desertores

Con el objetivo de **identificar el perfil** de los clientes con **mayor probabilidad de abandonar** MovistarPlus, realizamos un análisis basado en técnicas de **aprendizaje supervisado** utilizando un conjunto de datos históricos proporcionado por la empresa. Para aplicar estos modelos de forma efectiva, se llevó a cabo un proceso previo de **preparación y limpieza del dataset**. Para ello:

- **Eliminación de valores nulos**: Se verificó la integridad del conjunto de datos, confirmando que no existían valores ausentes.
- **Codificación de la variable objetivo**: La variable de deserción (Churn) fue transformada a formato numérico, asignando el valor 1 a los clientes que abandonaron (Yes) y 0 a los que permanecieron (No).
- **Transformación de variables categóricas**: Las variables categóricas presentes en el dataset fueron convertidas mediante codificación adecuada para poder ser utilizadas por los modelos de machine learning.
- **Eliminación de identificadores irrelevantes**: La columna customerID, al no aportar información útil para el análisis predictivo, fue eliminada.

Tras esto (pero antes de la transformación de las variables categóricas), realizamos un análisis exploratorio de los datos para tener una imagen completa del problema al que nos enfrentamos. Vemos para cada valor de cada característica, el porcentaje de desertores, con gráficos de barras para las variables categóricas (fig. 1), y *boxplots* para las numéricas, acompañadas de la distribución de valores a través de un histograma (fig. 2).

Con este breve análisis podemos ver la tasa de abandono según los valores de cada una de nuestras variables. El género de nuestros clientes, si tienen o no servicio telefónico, o si tienen contratado con nosotros múltiples líneas telefónicas no influye en gran medida en la tasa de abandono. Según SeniorCitizen, Partner y Dependent, podemos ver que, en proporción, nos abandonan más personas jubiladas, sin pareja y sin personas dependientes en el contrato. Respecto al servicio de internet, los clientes que tienen contratado fibra óptica son los que más nos abandonan (con gran diferencia). Online Security, Online Backup, Device Protection y Tech Support nos indican que la mayor proporción de abandonos la tenemos por parte de clientes sin estos servicios contratados, lo que puede ser indicador de la satisfacción de los clientes que contratan nuestros servicios adicionales. Esto no ocurre con el servicio de Streaming TV o Movies, donde tenemos un porcentaje de abandonos similar para ambos tipos de clientes. Respecto al contrato y método de pago, el grueso de los clientes que nos abandonan parece ser clientes con contrato mensual que pagan de forma electrónica.

Analizando la distribución de los desertores y no desertores en función de los meses de permanencia, vemos que cuanto más tiempo pasan con nosotros, mayor permanencia. La gran mayoría de los clientes que nos abandonan lo hacen antes de cumplir el primer año con nosotros. Respecto a los cargos, no parece haber una diferencia significativa que nos haga discernir características de desertores y no desertores.

Una vez preparado el conjunto de datos, se procedió a su adecuación para ser utilizado con modelos de machine learning. En esta etapa, se consideraron como predictores todas las características disponibles, excluyendo únicamente la columna de deserción codificada (Churn) y su versión categórica original.

Posteriormente, se dividió el conjunto de datos en subconjuntos de **entrenamiento y prueba** utilizando una **partición estratificada**, con el objetivo de mantener la proporción de desertores y no desertores en ambas muestras. Para asegurar una adecuada homogeneización de las variables numéricas, se aplicó una **estandarización** mediante StandardScaler, normalizando las variables a una media de 0 y una desviación estándar de 1.

Para lograr resultados más fiables y comparativos, seleccionamos un conjunto de algoritmos supervisados:

- **Regresión logística con penalización ElasticNet** (L1+L2): combina penalizaciones L1 (que favorece la selección de variables) y L2 (que aporta estabilidad ante multicolinealidad).
- **SVC con kernel lineal**: un modelo robusto frente a valores atípicos, que nos permite capturar relaciones lineales entre las variables y la clase objetivo.

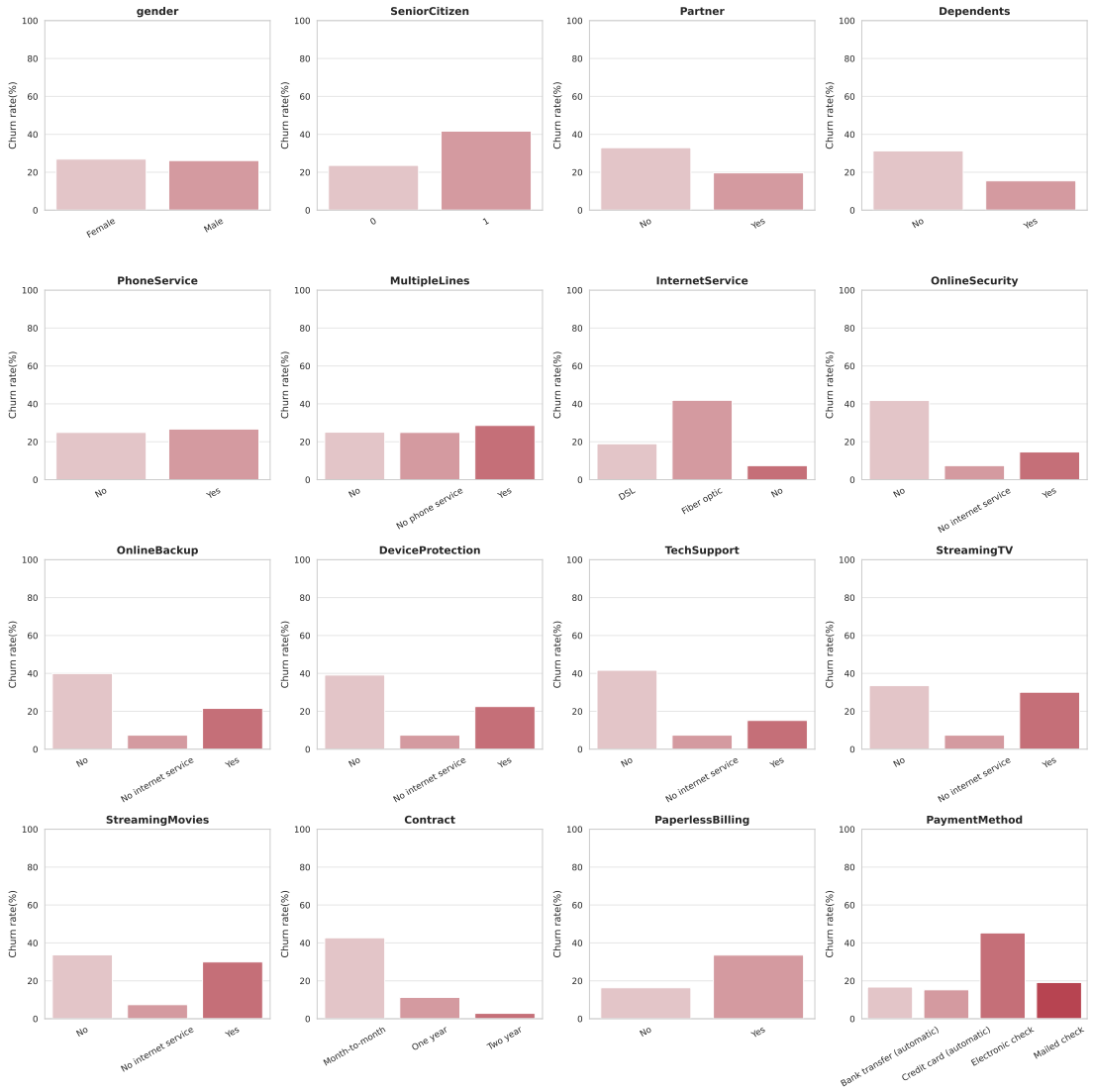


Fig. 1. Análisis exploratorio sobre las variables categóricas. Porcentaje de abandono según cada característica.

- **XGBoost**: aunque menos interpretable, es un modelo de *boosting* potente que nos permite detectar relaciones complejas y no lineales en los datos.
- **CatBoost**: como alternativa más eficiente a XGBoost y con menor sesgo hacia las variables categóricas.
- **MLPClassifier**: usamos una red neuronal multicapa simple, con dos capas ocultas y técnicas de regularización para evitar el sobreajuste. A pesar de su menor interpretabilidad, ofrece la posibilidad de capturar patrones no triviales y altamente no lineales.

Tras entrenar los cinco modelos supervisados anteriores utilizando el conjunto de entrenamiento, se evaluó su capacidad de generalización sobre el conjunto de prueba. Para interpretar las decisiones de cada modelo y entender qué variables influyen más en la predicción de la deserción, recurrimos a los **valores Shapley (SHAP)** como métrica de importancia.

Los valores SHAP proporcionan una medida equitativa de la contribución de cada característica a la predicción, positiva o negativa, en función de si aumenta o reduce la probabilidad de que un cliente sea clasificado como desertor. Estos valores se calculan como la media sobre todas las permutaciones posibles de inclusión de una variable en el modelo, lo que permite capturar su impacto de forma precisa y transparente.

A nivel técnico:

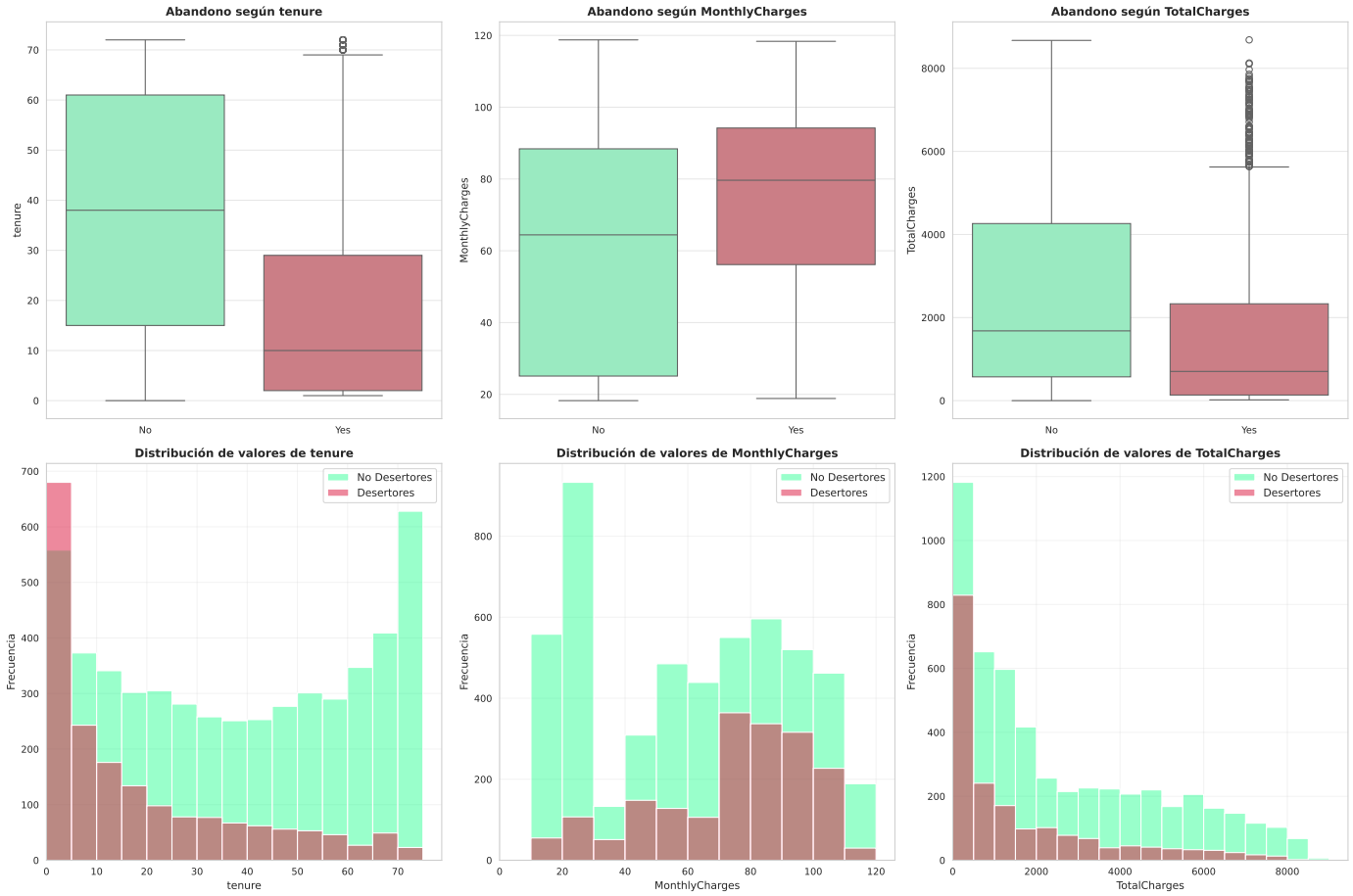


Fig. 2. Análisis exploratorio sobre las variables numéricas. Porcentaje de abandono según cada característica y distribución de valores para los grupos de desertores y no desertores.

1. Se utilizó **SHAP Explainer** para modelos lineales como la regresión logística, SVC y la red neuronal, y **TreeExplainer** para los modelos de boosting (XGBoost y CatBoost).

2. A partir de las matrices de valores SHAP generadas, se extrajo la importancia de cada variable como la media del valor absoluto de sus contribuciones en todas las observaciones.

Para sintetizar una medida global de importancia de las variables, construimos un **ranking conjunto.** En cada modelo, las variables fueron ordenadas por importancia (SHAP), asignando el valor 1 a la más relevante y valores sucesivos en orden descendente. Finalmente, se calculó la **mediana** de cada característica en los cinco rankings, priorizando esta medida frente a la media por su robustez ante posibles discrepancias entre modelos.

A.1. Características más relevantes

A partir del análisis global, las variables con mayor capacidad explicativa sobre la deserción fueron (figura 3):

- **tenure**: Número de meses de permanencia
- **Contract**: tipo de contrato (mensual, anual, bianual).
- **MonthlyCharges** y **TotalCharges**: facturación mensual y total acumulada.
- **OnlineSecurity**: contratación del servicio de seguridad online.
- **TechSupport**: Si tiene contratado soporte técnico
- **PaperlessBilling**: Si ha seleccionado facturación sin papel

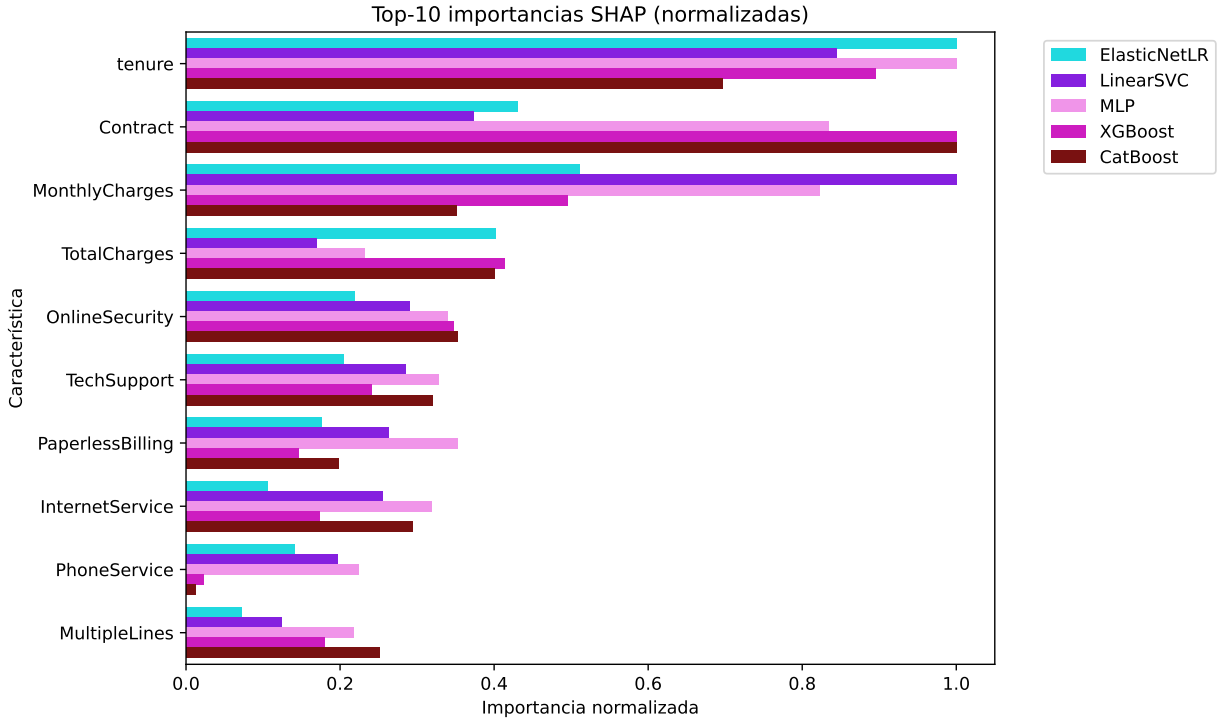


Fig. 3. Top 10 características tras la selección de características.

- **InternetService**: Tipo de servicio de Internet

Tras este análisis y, en conjunto con la exploración inicial de los datos, somos capaces de trazar un perfil acertado de los clientes que nos abandonan y de cuáles prefieren elegirnos año tras año como sus proveedores de internet y/o telefonía móvil. Daremos forma a este perfil en base a las 6 características más importantes [1].

A.2. Perfil del cliente desertor

- Permanencia (*tenure*): baja permanencia, en su gran mayoría menos de un año y, en casi un 90%, menor a 2 años.
- Tipo de contrato (*Contract*): en su gran mayoría son clientes con contratos mensuales (de mes a mes). En este sector la tasa de abandono se eleva hasta el 40%.
- Factura mensual (*MonthlyCharges*): generalmente, el grupo de clientes que nos abandono tiende a pagar más ($\sim$ 80$/mes) que los clientes fieles ($\sim$ 60€/mes). Además, la gran mayoría de los clientes desertores concentran sus contratos entre los 60 y los 100€/mes.
- Seguridad online (*OnlineSecurity*): la gran mayoría de los desertores NO contrata con nosotros la seguridad online. El porcentaje de clientes que nos abandona cumpliendo estas características se eleva más allá del 40%.
- Soporte técnico (*TechSupport*): la gran mayoría de los desertores NO contrata con nosotros el soporte técnico. El porcentaje de clientes que nos abandona cumpliendo estas características se eleva más allá del 40%.
- Tipo de servicio de internet (*InternetService*): la gran mayoría de los desertores tiene contratado fibra óptica. El porcentaje de clientes que nos abandona cumpliendo estas características se eleva más allá del 40%. Esto nos indica posibles problemas con la calidad o precio de nuestra fibra óptica. Otro motivo es que la competencia ofrezca fibra de mejor calidad o a precios más competitivos.

Por tanto, los clientes que nos abandonan son, en su gran mayoría, clientes nuevos con contrato mensual de precio superior a 80€/mes y que no contratan ningún tipo de servicio o soporte técnico con nosotros. La fibra óptica parece un problema a estudiar desde la compañía; sería necesario hacer un estudio de mercado y evaluar el servicio que estamos ofreciendo mediante un análisis DAFO.

[1] No incluiremos los gastos totales ya que es una variable con una correlación directa con la permanencia y el gasto mensual, por lo que no aporta ningún valor adicional. Además, la factura en papel está asociada a nuestros clientes veteranos y personas generalmente mayores, pero la mayoría de contratos actuales no la incluyen.

A.3. Perfil del cliente no desertor

- Permanencia (*tenure*): alta permanencia, en su gran mayoría 36 meses. El grueso de clientes fieles se sitúan por encima de un año de permanencia y por debajo de 5 años.

- Tipo de contrato (*Contract*): en su gran mayoría son clientes con contratos anuales o bianuales. Es muy raro que clientes con uno de los contratos anteriores nos abandone.

- Factura mensual (*MonthlyCharges*): generalmente, el grupo de clientes que nos abandono tiende a pagar aproximadamente 60$/mes. No obstante, los contratos de nuestros clientes files abarcan un rango muy amplio, desde los más baratos de los que disponemos, sobre los 20$/mes, hasta los de mejor oferta y con más servicios incluidos, en torno a los 90$/mes.

- Seguridad online (*OnlineSecurity*): la gran mayoría de nuestros clientes más fieles no tienen contratado un servicio de internet, por lo que no les aplica la seguridad online. No obstante, de los clientes con servicio de internet contratado, entre los que sí disponen de servicio de seguridad online, solo nos abandona un poco más del 10%, lo que indica alta fidelidad también entre los clientes con seguridad online contratada.

- Soporte técnico (*TechSupport*): la gran mayoría de nuestros clientes más fieles no tienen contratado un servicio de internet, por lo que no les aplica el soporte técnico. No obstante, de los clientes con servicio de internet contratado, entre los que sí disponen de servicio de soporte técnico, solo nos abandona un poco más del 10%, lo que indica alta fidelidad también entre los clientes con soporte técnico contratado.

- Tipo de servicio de internet (*InternetService*): de nuevo, nuestros clientes más fieles se caracterizan por no tener servicio de internet contratado. No obstante, entre los que sí lo tienen contratado, los que tienen DSL tienen una tasa de permanencia mucho mayor, sobre el 80%, lo que nos indica buena satisfacción con este servicio.

Por tanto, los clientes que permanecen con nosotros se caracterizan por contratos anuales (al menos) de un precio variable en torno a 60$/mes y que, de forma general, no tienen servicio de internet contratado. De los que tienen servicio de internet contratado, la contratación de servicios y soporte técnico la asociamos con una mayor fidelidad. Además, el factor clave es la permanencia de estos clientes: los datos demuestran que cuanto más tiempo pasan con nosotros, menos probabilidad de abandono existe.

B. Segmentación de clientes

Con el objetivo de diseñar campañas de incentivos más personalizadas y efectivas, se procedió a la **segmentación de los clientes** en función de sus atributos más relevantes. Esta segmentación se realizó por separado para los desertores y los no desertores, dado que sus motivaciones y necesidades pueden diferir significativamente.

Se decidió usar un algoritmo **K-Means** ya que, debido a la dimensionalidad (número de características) de los datos, podemos disponer a los clientes como puntos en un espacio de alta dimensión. Esto hace que los conjuntos sean, con gran probabilidad, convexos y bien diferenciados entre ellos, por lo que un **K-Means** resulta un método muy fiable. Para determinar el **número óptimo de clusters** (o grupos en los que segmentar los clientes), se utilizó el **método del codo**, que mostró un punto de inflexión claro en 5 clusters para los no desertores y en 4 para los desertores (ver figura 4). En las tablas 1 y 2 se muestran los resultados del *clustering* sobre los grupos de desertores y no desertores, respectivamente.

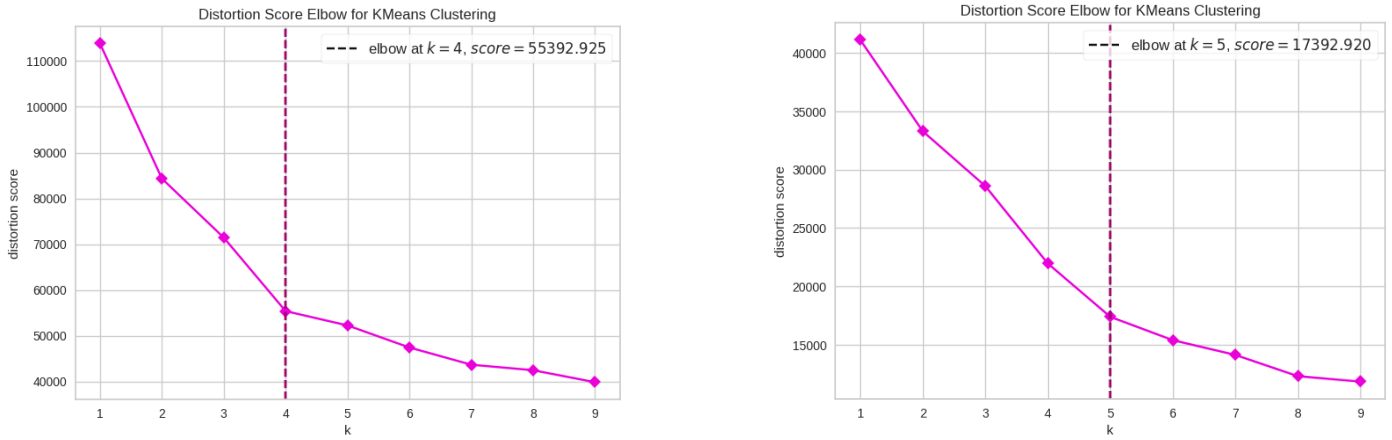
B.1. Segmentación de clientes desertores

1. El primer grupo que identificamos está formado por usuarios de alrededor de dos años de antigüedad que gastan unos 80€/mes. Contratan internet (tanto fibra como DSL) y teléfono, pero tienen contratos mensuales (mes a mes), sin añadir casi ningún tipo de servicio o soporte adicional.

2. El segundo grupo lleva más de dos años con nosotros, pero sin internet contratado ni servicios extra. Tienen un consumo bajo, de unos 20€ al mes y muchos prefieren firmar con nosotros por uno o dos años.

3. El tercer grupo que identificamos en los desertores van en contra de los anteriores: clientes veteranos (más de 5 años con nosotros), con facturas mensuales muy elevadas, por encima de 80€ y varios servicios y soporte contratados. Estos clientes nos interesa mantenerlos, ya que son de alto valor.

4. El último grupo que identificamos aquí son clientes con solo DSL contratado y algún servicio contratado, pero facturas mensuales de unos 40€.



(a) Búsqueda sobre el grupo de desertores.

(b) Búsqueda sobre el grupo de no desertores.

Fig. 4. Búsqueda del número óptimo de grupos.

B.2. Segmentación de clientes no desertores

1. El primer grupo de clientes fieles que encontramos son clientes que llevan con nosotros unos 18 meses, pagando unos 38€/mes. Solo contratan servicio DSL de internet, pero ningún tipo de extra.

2. El segundo grupo son clientes con un gasto mensual elevado, de unos 90€/mes con contratos mes a mes. La gran mayoría tienen contratada fibra óptica pero sin servicios adicionales contratados.

3. El tercer grupo que identificamos son nuestros clientes más valiosos: clientes de factura elevada (sobre los 95€/mes) que llevan con nosotros un largo periodo de tiempo (más de cuatro años), con lo cual nos han dejado una facturación total elevada. Los contratos de estos clientes son, al menos, anuales, con una variedad de servicios adicionales contratados.

4. Un cuarto grupo de clientes existente sería el opuesto al anterior: clientes nuevos (de menos de un año), con factura mensual extremadamente baja (20€/mes) y sin servicio de internet, por lo que tampoco disfrutan de nuestros servicios adicionales.

5. Por último, identificamos un grupo que supone un compromiso entre el segundo y el cuarto: nuevos clientes (menos de un año) con contrato mes a mes y de facturación relativamente elevada (unos 72€/mes). Estos clientes recurren a nuestro servicio de internet por fibra óptica, pero no al soporte técnico o seguridad online que ofrecemos para complementarlo.

C. Diseño de la campaña de incentivos

En vista del perfil de los clientes, así como de la segmentación trazada sobre cada perfil (desertores y no desertores), debemos diseñar una campaña de incentivos adecuada para mantener a los clientes que actualmente deciden dejarnos, pero sin olvidarnos también de cuidar los clientes que deciden seguir con nosotros mes a mes o año tras año. Para ello, esbozamos a continuación una serie de incentivos para cada grupo de los anteriormente mencionados, destacando los grupos más valiosos a mantener, de forma que la dirección de la empresa pueda iniciar las campañas que crea necesarias para conservar el/los grupos que considere de mayor valor.

C.1. Campaña de incentivos para desertores

1. El primer grupo de clientes se mueven según la oferta, lo que se puede inferir por su falta de compromiso (contrato mes a mes); debido a esto, son propensos a abandonarnos si la competencia les muestra una mejor oferta. Lo ideal para evitar la deserción de estos clientes sería ofrecer descuentos en los contratos anuales o bianuales, haciendo que su precio sea menor que 12$\times$ precio/mes.

2. El segundo grupo se trata de clientes relativamente veteranos pero sin internet contratado con nosotros. Su marcha nos hace indicar que la competencia puede ofrecerles algún paquete con internet incluido a un precio más atractivo que nosotros. Para evitar perder a estos clientes, recomendaríamos ofrecer nuestro servicio de internet, ya sea fibra o DSL, por un precio menor durante un periodo introductorio de, al menos, 6 meses.

	Grupo 1	Grupo 2	Grupo 3	Grupo 4
tenure	22.187886	32.331210	57.541386	37.017578
MonthlyCharges	73.686434	21.136058	89.278418	43.520996
TotalCharges	1732.513195	701.685740	5208.500000	1764.628906
Contract_Month-to-month	0.874536	0.300778	0.096260	0.435547
Contract_One year	0.114957	0.251238	0.389332	0.255859
Contract_Two year	0.010507	0.447983	0.514408	0.308594
OnlineSecurity_No	0.775649	0.000000	0.321275	0.503906
OnlineSecurity_No internet service	0.000000	1.000000	0.000000	0.000000
OnlineSecurity_Yes	0.224351	0.000000	0.678725	0.496094
TechSupport_No	0.792336	0.000000	0.302882	0.490234
TechSupport_No internet service	0.000000	1.000000	0.000000	0.000000
TechSupport_Yes	0.207664	0.000000	0.697118	0.509766
PaperlessBilling_No	0.343016	0.714084	0.365420	0.474609
PaperlessBilling_Yes	0.656984	0.285916	0.634580	0.525391
InternetService_DSL	0.433869	0.000000	0.365420	1.000000
InternetService_Fiber optic	0.566131	0.000000	0.541386	0.000000
InternetService_No	0.000000	1.000000	0.000000	0.000000
PhoneService_No	0.000000	0.000000	0.000000	1.000000
PhoneService_Yes	1.000000	1.000000	1.000000	0.000000
MultipleLines_No	0.623609	0.765039	0.276517	0.000000
MultipleLines_No phone service	0.000000	0.000000	0.000000	1.000000
MultipleLines_Yes	0.376391	0.234961	0.723483	0.000000

Table 1. Resultados del clustering sobre el grupo de desertores

3. El tercer grupo es el más valioso entre los desertores, y su permanencia debe ser una **prioridad**: los veteranos con contratos mensuales elevados. Su abandono tras un largo periodo de tiempo puede sugerir problemas puntuales en algún servicio o un precio elevado en comparación a paquetes similares ofrecidos por la competencia. Para favorecer la permanencia de este grupo, deberíamos lanzar un programa Gold o VIP para este tipo de clientes, ofreciendo un soporte 24h, ventajas exclusivas o descuentos en dispositivos. Las propuestas anticipadas de renovación, con condiciones preferentes, también puede ser un buen incentivo para mantener a estos clientes con nosotros.

4. Este grupo solo dispone de DSL contratado, pero de contratos de un precio mensual medio, sobre los 40€, y su abandono puede sugerir problemas o falta de satisfacción con la calidad del servicio. Para frenar su marcha, se podría ofrecer la posibilidad de migrar a fibra óptica (sin costes de instalación) si su región lo permite.

C.2. Campaña de incentivos para no desertores

1. El primer grupo de clientes se mantiene únicamente con el contrato de DSL. Ya identificamos un grupo de desertores similar que termina abandonando la compañía, lo que vimos podía sugerir problemas con el servicio. Por tanto, para este grupo se podría implementar una solución similar: ofrecer la posibilidad de migrar a fibra óptica (sin costes de instalación) si su región lo permite.

2. El segundo grupo de los no desertores se trata de los clientes con gasto mensual elevado, fibra, pero sin servicios adicionales. La idea para este grupo sería introducirles a las ventajas que suponen esos servicios de los que carecen.

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
tenure	15.835294	20.969871	50.810811	8.238938	8.471591
MonthlyCharges	37.533824	89.119010	94.409189	20.368142	72.254048
TotalCharges	687.087941	1950.335366	4910.098108	173.919912	651.586790
Contract_Month-to-month	0.900000	1.000000	0.021622	0.876106	0.997159
Contract_One year	0.082353	0.000000	0.762162	0.079646	0.002841
Contract_Two year	0.017647	0.000000	0.216216	0.044248	0.000000
OnlineSecurity_No	0.829412	0.879484	0.556757	0.000000	0.857955
OnlineSecurity_No internet service	0.000000	0.000000	0.000000	1.000000	0.000000
OnlineSecurity_Yes	0.170588	0.120516	0.443243	0.000000	0.142045
TechSupport_No	0.835294	0.883788	0.437838	0.000000	0.862216
TechSupport_No internet service	0.000000	0.000000	0.000000	1.000000	0.000000
TechSupport_Yes	0.164706	0.116212	0.562162	0.000000	0.137784
PaperlessBilling_No	0.305882	0.126255	0.259459	0.628319	0.298295
PaperlessBilling_Yes	0.694118	0.873745	0.740541	0.371681	0.701705
InternetService_DSL	1.000000	0.045911	0.254054	0.000000	0.298295
InternetService_Fiber optic	0.000000	0.954089	0.745946	0.000000	0.701705
InternetService_No	0.000000	0.000000	0.000000	1.000000	0.000000
PhoneService_No	1.000000	0.000000	0.000000	0.000000	0.000000
PhoneService_Yes	0.000000	1.000000	1.000000	1.000000	1.000000
MultipleLines_No	0.000000	0.001435	0.318919	0.911504	0.974432
MultipleLines_No phone service	1.000000	0.000000	0.000000	0.000000	0.000000
MultipleLines_Yes	0.000000	0.998565	0.681081	0.088496	0.025568

Table 2. Resultados del clustering sobre el grupo de no desertores

Para ello, puede ser conveniente ofrecer un paquete introductorio a servicios adicionales y soporte, de forma que dispongan de ellos gratuitamente durante un periodo de tiempo corto pero prudente.

3. El tercer grupo consta de nuestros clientes más valiosos, los que debemos mantener: clientes veteranos con un contrato de precio elevado. Aquí, al igual que ocurría con los clientes valiosos que deciden abandonarnos, un programa VIP con soporte continuo y ventajas exclusivas creemos que es lo óptimo para favorecer la permanencia; son clientes que debemos cuidar.

4. Para el grupo de clientes nuevos con menos facturación mensual, deberíamos incluir ofertas de tarifas con internet, ya sea DSL o fibra, a un precio competitivo y sugerente al comienzo del contrato, de modo que no se vean atraídos por ofertas de la competencia.

5. Para los nuevos clientes con contratos elevados, pero sin servicios contratados, se debería incluir algún tipo de oferta introductoria a estos servicios, de modo que aumente su nivel de satisfacción y decidan permanecer con nosotros.

2. SISTEMAS DE RECOMENDACIÓN PARA NETFLIX EN EL MERCADO ESPAÑOL

En el contexto del desembarco de Netflix en España, se hace necesario adaptar los sistemas de recomendación a los gustos y hábitos del público local. A continuación, describimos el diseño e implementación de un sistema de recomendación

basado en factorización matricial, comparándolo con alternativas clásicas como métodos basados en popularidad, filtrado colaborativo por usuario y por ítem.

Además de los enfoques tradicionales y de factorización, se exploró un modelo de recomendación basado en contenido, que considera las características explícitas de los ítems (por ejemplo, géneros) y las preferencias aprendidas de los usuarios mediante un modelo de regresión Support Vector (SVR) con núcleo lineal. Este modelo fue implementado en un entorno interactivo y permite analizar recomendaciones personalizadas a partir de perfiles explícitos de usuario.

A. Fundamento Teórico

En el corazón del *data mining* yace una idea fundamental: representar fenómenos complejos de la manera más simple posible. Esta idea está profundamente conectada con el concepto de compresión y representación estructurada: cuanto mejor podamos modelar la información relevante, mejor podremos entenderla y utilizarla para hacer predicciones. En el contexto de los sistemas de recomendación, esto puede lograrse mediante distintas estrategias. Por un lado, la factorización matricial permite extraer patrones latentes que explican las preferencias implícitas de los usuarios. Por otro, los modelos basados en contenido aprovechan características explícitas de los ítems, como sus géneros, para construir perfiles personalizados de los usuarios. Ambos enfoques, aunque distintos, comparten el objetivo de ofrecer recomendaciones eficaces a partir de representaciones compactas e informativas.

A.1. Descomposición Matricial

Dado un conjunto de valoraciones $R \in \mathbb{R}^{m \times n}$, donde m es el número de usuarios y n el número de películas (ítems), se busca aproximar esta matriz como el producto de dos matrices de menor rango:

$$R \approx P \cdot Q^\top$$

donde:

- $P \in \mathbb{R}^{m \times k}$: matriz de factores latentes de usuarios.
- $Q \in \mathbb{R}^{n \times k}$: matriz de factores latentes de películas.
- $k \ll \min(m, n)$: número de factores latentes.

Cada fila de P representa las preferencias latentes de un usuario, mientras que cada fila de Q representa las características latentes de una película. La predicción de una valoración $\hat{r}_{ui}$ se realiza mediante el producto escalar entre las respectivas filas:

$$\hat{r}_{ui} = P_u \cdot Q_i^\top$$

Inspiración en la Descomposición en Valores Singulares (SVD)

Una de las técnicas clásicas para entender matrices es la *Descomposición en Valores Singulares* (SVD), que plantea que una matriz A se puede escribir como:

$$A = USV^\top$$

donde:

- $U \in \mathbb{R}^{m \times m}$: matriz ortogonal de usuarios.
- $S \in \mathbb{R}^{m \times n}$: matriz diagonal con valores singulares.
- $V \in \mathbb{R}^{n \times n}$: matriz ortogonal de ítems.

Los valores ubicados en la diagonal de la matriz S corresponden a los valores singulares de la matriz A, ordenados de mayor a menor. Estos valores pueden interpretarse, de forma análoga al Análisis de Componentes Principales (PCA), como la cantidad de varianza explicada por cada una de las componentes latentes del sistema.

En este contexto, la **varianza truncada** hace referencia a la práctica de conservar únicamente los primeros k valores singulares más grandes, ignorando los restantes. Esto permite construir una *aproximación de bajo rango* de la matriz original A, que captura la mayor parte de la estructura o información relevante, mientras descarta ruido o detalles menos significativos.

Este procedimiento es conceptualmente similar a PCA, donde también se retienen las primeras k componentes principales (aquellas que explican mayor varianza) para reducir la dimensionalidad de los datos. En ambos casos, la idea es representar los datos en un espacio de menor dimensión que conserva la mayor parte de la varianza o información original, facilitando tareas como compresión, visualización o predicción.

En la práctica, esta idea se reescribe como:

$$A \approx X\Theta^\top$$

donde:

- $X \in \mathbb{R}^{m\times k}$: representa usuarios en un espacio latente.
- $\Theta \in \mathbb{R}^{n\times k}$: representa ítems en el mismo espacio latente.

Los factores latentes k podrían asociarse con dimensiones como "acción", "romance", o incluso factores geográficos (e.g. "NYC" vs. "Dallas"), aunque estos no siempre sean interpretables de forma directa. Por ello, este enfoque es conocido como **factorización matricial con factores latentes** (*Latent Factor Matrix Factorization*).

A.2. Función de Costo Regularizada

El modelo se entrena minimizando una función de costo que solo considera los pares observados:

$$J(P,Q) = \frac{1}{2}\sum_{(u,i)\in\mathcal{K}} (r_{ui} - P_u \cdot Q_i^\top)^2 + \frac{\lambda}{2}\left(\|P_u\|^2 + \|Q_i\|^2\right)$$

donde:

- $\mathcal{K}$: conjunto de pares (usuario, ítem) con valoración conocida.
- λ: parámetro de regularización que evita el sobreajuste.

En este enfoque, los datos disponibles corresponden a un subconjunto de valoraciones observadas, representado por el conjunto $\mathcal{K}$, que contiene los pares (u,i) donde el usuario u ha valorado el ítem i, con sus respectivas puntuaciones r_{ui} almacenadas en la matriz Y. La matriz binaria R, del mismo tamaño que Y, indica las posiciones observadas ($R_{ui} = 1$) y no observadas ($R_{ui} = 0$). Por otro lado, los parámetros que se entrenan son las matrices X y Θ: X contiene los vectores de características latentes de los usuarios, mientras que Θ representa los vectores de características latentes de los ítems. El proceso de entrenamiento consiste en ajustar X y Θ de modo que la función de costo regularizada se minimice sobre las entradas observadas, permitiendo así predecir valoraciones no observadas a partir del producto escalar entre los factores latentes correspondientes.

De manera matricial, esto puede expresarse como:

$$J = \frac{1}{2}R \odot (Y - X\Theta^\top)^2 + \frac{\lambda}{2}\left(\|X\|^2 + \|\Theta\|^2\right)$$

donde R es una matriz binaria del mismo tamaño que Y, con 1 en las posiciones observadas y 0 en las no observadas. El operador $\odot$ indica multiplicación elemento a elemento.

A.3. Optimización

Los parámetros X y Θ se actualizan mediante gradiente descendente, utilizando las derivadas parciales de la función de costo:

$$\frac{\partial J}{\partial X} = R \odot (Y - X\Theta^\top)\Theta + \lambda X$$

$$\frac{\partial J}{\partial \Theta} = \left(R \odot (Y - X\Theta^\top)\right)^\top X + \lambda\Theta$$

Estas expresiones permiten aplicar algoritmos de optimización eficientes, como `L-BFGS-B`, ampliamente disponibles en entornos de programación como `R` o `Python`.

A.4. Recomendación Basada en Contenido

El modelo de recomendación basado en contenido parte de la idea de que las preferencias de un usuario pueden inferirse directamente a partir de las características de los ítems que ha valorado, sin necesidad de recurrir a las valoraciones de otros usuarios. En este caso, cada película se representa mediante un vector binario que codifica la presencia o ausencia de distintos géneros cinematográficos (por ejemplo, acción, comedia, drama, etc.). Este vector constituye una descripción explícita y estructurada del contenido de la película.

Para capturar las preferencias individuales de cada usuario, se entrena un modelo de regresión Support Vector Regression (SVR) con núcleo lineal, tomando como variables independientes los géneros de las películas, y como variable dependiente, las valoraciones otorgadas por el usuario. Este tipo de modelo ofrece buena capacidad de generalización y es adecuado para contextos con pocas observaciones o alta colinealidad entre características, ya que busca un equilibrio entre margen de error y complejidad del modelo.

Este enfoque no requiere información sobre otros usuarios, lo que lo hace especialmente útil en escenarios de inicio en frío o cuando se dispone de metadatos ricos sobre los ítems. El modelo genera un perfil de preferencias para cada usuario que puede interpretarse directamente a través de los coeficientes asociados a cada género, facilitando tanto la personalización como la interpretación de las recomendaciones generadas.

Resumen

Los sistemas de recomendación pueden abordarse desde múltiples enfoques complementarios. La factorización de matrices con factores latentes es un método robusto y ampliamente utilizado, que permite capturar patrones complejos en las interacciones usuario-ítem, generando recomendaciones personalizadas y escalables a partir de la descomposición de la matriz de valoraciones. Esta técnica facilita una compresión eficiente de la información, ofreciendo representaciones implícitas de las preferencias de los usuarios y las características de los ítems.

Por otro lado, los modelos de recomendación basados en contenido utilizan atributos explícitos de los ítems —como los géneros cinematográficos— para construir perfiles personalizados a nivel de usuario. Este enfoque permite interpretar directamente las preferencias individuales y resulta especialmente útil cuando se dispone de metadatos ricos o en escenarios con usuarios nuevos. En conjunto, ambos métodos proporcionan soluciones eficaces y complementarias dentro del diseño de sistemas de recomendación.

Otros Recomendadores

Los sistemas de recomendación colaborativos se basan en el principio de que los usuarios que han coincidido en sus valoraciones en el pasado, probablemente coincidan en el futuro. Existen varias variantes de estos modelos, entre las que destacan:

Modelo Popular

Este modelo recomienda los ítems más populares a todos los usuarios, es decir, aquellos con mayor puntuación promedio o mayor cantidad de valoraciones recibidas.

Filtrado Colaborativo Basado en Usuarios (User-Based)

Este método predice la valoración de un usuario u sobre un ítem i en base a las valoraciones de usuarios similares:

$$\hat{r}_{u,i} = \bar{r}_u + \frac{\sum_{v \in N_u(i)} \text{sim}(u,v) \cdot (r_{v,i} - \bar{r}_v)}{\sum_{v \in N_u(i)} |\text{sim}(u,v)|}$$

donde $N_u(i)$ es el conjunto de vecinos del usuario u que han valorado el ítem i, $\text{sim}(u,v)$ es la similitud entre usuarios, y $\bar{r}_u$ es la media de valoraciones del usuario u.

Filtrado Colaborativo Basado en Ítems (Item-Based)

En este caso, se estima la valoración en base a la similitud entre ítems:

$$\hat{r}_{u,i} = \frac{\sum_{j \in N_i(u)} \text{sim}(i,j) \cdot r_{u,j}}{\sum_{j \in N_i(u)} |\text{sim}(i,j)|}$$

donde $N_i(u)$ representa el conjunto de ítems similares a i que el usuario u ha valorado.

B. Implementación del modelo

Se implementaron tres enfoques complementarios para construir sistemas de recomendación, todos desarrollados en Python. El primero se basa en regresión lineal regularizada (`Ridge`), el segundo en factorización matricial mediante la biblioteca `Surprise`, y el tercero en una regresión por *Support Vector Machines* con núcleo lineal (`LinearSVR`), centrado en las características de contenido de las películas.

B.1. Regresión Ridge para preferencias de usuario

Este enfoque tiene como objetivo capturar las preferencias individuales de un usuario representativo, a partir de los géneros de las películas que ha valorado.

1. Carga y preprocesamiento de los datos.
2. Transformación de los géneros a un formato numérico binario.
3. Selección de un usuario con suficientes valoraciones para asegurar estabilidad en el modelo.
4. Entrenamiento de un modelo `Ridge` (regresión lineal con regularización L2), lo que permite controlar el sobreajuste en presencia de multicolinealidad.
5. Evaluación del modelo sobre un conjunto de validación utilizando la métrica RMSE.
6. Interpretación de los coeficientes del modelo para analizar las preferencias del usuario en términos de géneros.

B.2. Factorización matricial con `Surprise`

Para capturar patrones generales de interacción entre usuarios e ítems, se utilizó el algoritmo `SVD` provisto por la biblioteca `Surprise`, especializada en sistemas de recomendación.

1. Los datos se cargaron desde el conjunto MovieLens 10M y se adaptaron al formato requerido por `Surprise`.
2. El conjunto se dividió en datos de entrenamiento y test (usualmente en una proporción 80/20).
3. Se entrenó el modelo `SVD` configurando hiperparámetros como el número de factores latentes, la tasa de aprendizaje y el número de épocas.
4. Se evaluó el rendimiento del modelo mediante las métricas RMSE y MAE.
5. Finalmente, se generaron recomendaciones personalizadas para un subconjunto de usuarios.

B.3. Recomendación basada en contenido con regresión SVR

Este enfoque tiene como objetivo capturar las preferencias individuales de un usuario representativo a partir de las características de las películas que ha valorado, en particular, sus géneros.

1. **Carga y preprocesamiento de los datos.**
2. **Transformación de los géneros a un formato numérico binario**, en el que cada género se representa como una variable indicadora.
3. **Selección de un usuario con suficientes valoraciones** para asegurar estabilidad en el modelo.
4. **Entrenamiento de un modelo de regresión SVR (Support Vector Regression)** con núcleo lineal. Este modelo busca una función que aproxime las valoraciones del usuario minimizando el error dentro de un margen definido y penalizando desviaciones mayores, lo que permite una buena generalización sin asumir una relación estrictamente lineal.
5. **Evaluación del modelo** sobre un conjunto de validación utilizando la métrica RMSE.
6. **Interpretación de los coeficientes del modelo**, que reflejan la influencia de cada género en las valoraciones del usuario, permitiendo un análisis personalizado de sus preferencias.

Este modelo fue implementado y explorado en un entorno interactivo de *notebook*, donde se visualizan tanto los coeficientes generados por el modelo como ejemplos de predicción para películas no vistas por el usuario. Esta visualización facilita la comprensión de cómo las preferencias aprendidas influyen en las recomendaciones producidas.

C. Resultados

En esta sección se presentan los resultados obtenidos por los distintos modelos de recomendación evaluados. Se incluyen métricas de error como RMSE y MAE, así como medidas de desempeño basadas en precisión y cobertura.

C.1. Errores de predicción

En la Tabla 3 se puede observar el RMSE promedio obtenido por cada algoritmo, incluyendo un modelo base aleatorio. Estos valores permiten contextualizar el rendimiento relativo de los enfoques más sofisticados respecto a estrategias simples.

Algoritmo	RMSE promedio
Random Items	1.5207
Popular Items	0.9471
User-based CF	0.9882
Matrix Factorization	**0.9462**
LinearSVR (usuario 59269)	0.613
LinearSVR (agrupado)	0.998 – 1.086

Table 3. RMSE promedio por algoritmo de recomendación

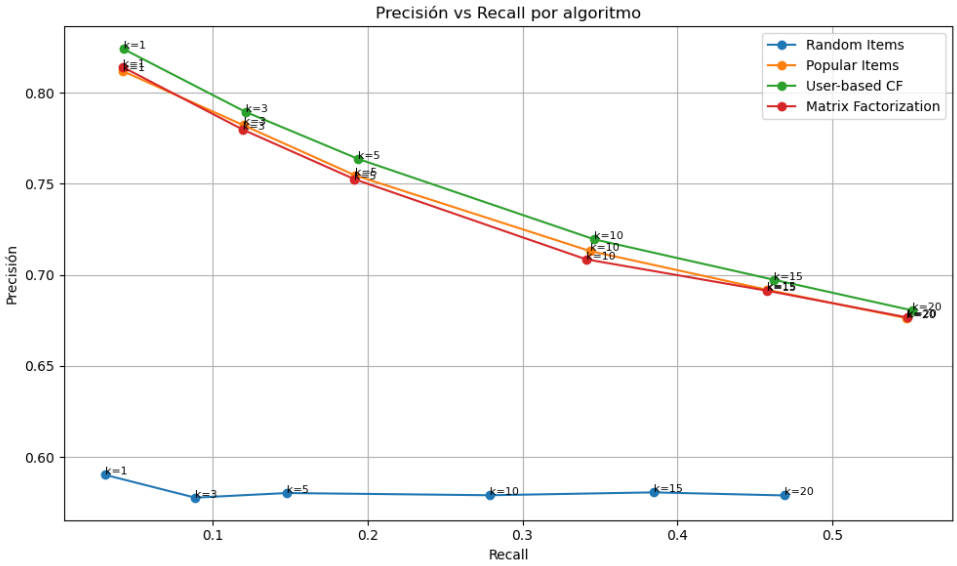


Fig. 5. Curvas de precisión vs. recall para distintos algoritmos y valores de k.

C.2. Precisión y cobertura

Modelos de recomendación basados en algoritmos

Además de las métricas de error, se evaluaron las recomendaciones generadas utilizando curvas de precisión vs. recall y curvas ROC para distintos valores de k (número de ítems recomendados). Las Figuras 5 y 6 muestran estas curvas para los diferentes algoritmos comparados.

Estos resultados refuerzan que, si bien el modelo de factorización matricial logra los menores errores de predicción, los métodos basados en vecinos (User-based e Item-based) tienden a ofrecer mejor cobertura en valores bajos de k, lo cual puede ser útil en sistemas que prioricen diversidad o exploración.

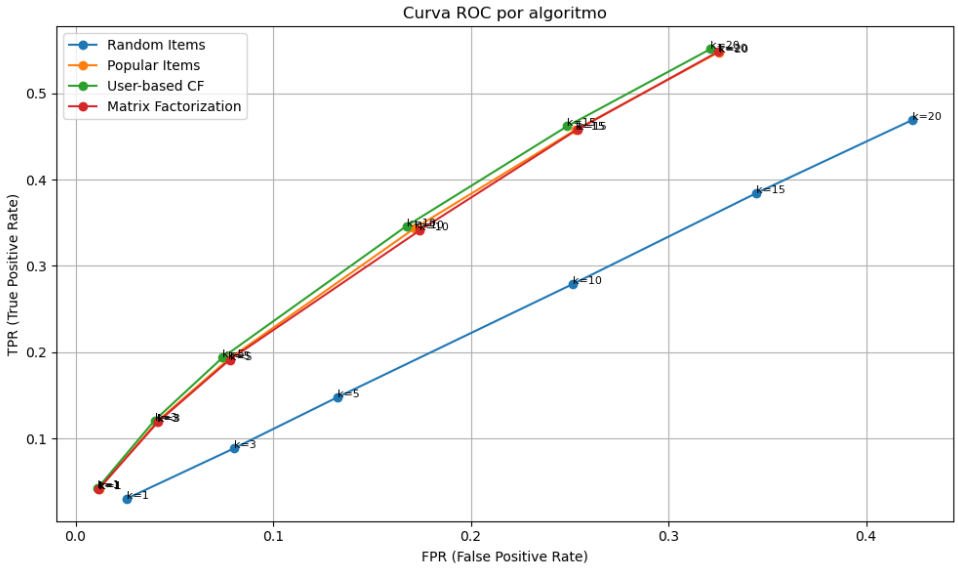
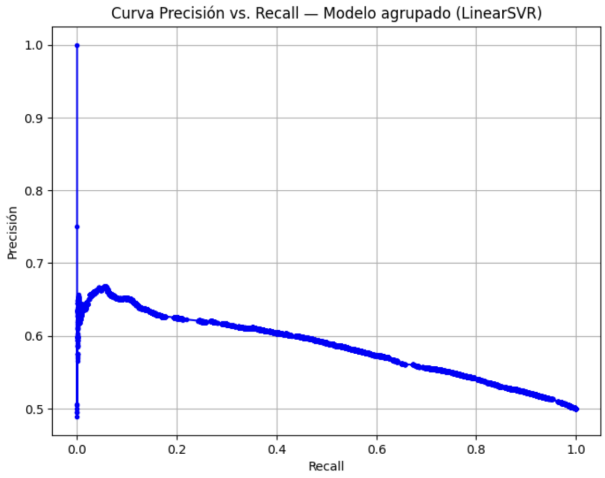


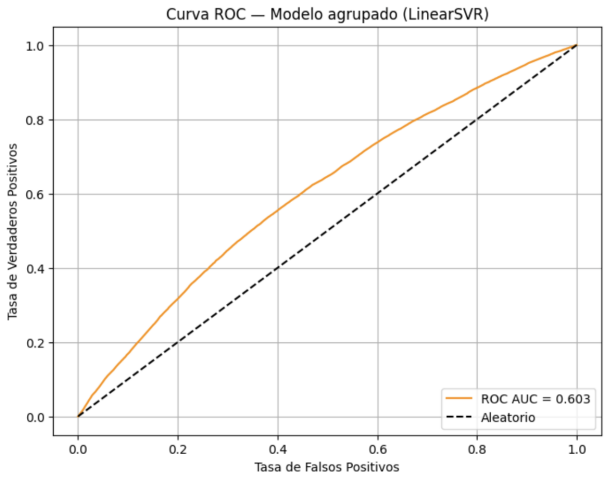
Fig. 6. Curvas ROC superpuestas para distintos algoritmos de recomendación.

Modelo de recomendación basado en contenido

Además de las métricas de error, se evaluaron las recomendaciones generadas utilizando curvas de precisión vs. recall y curvas ROC para el modelo de recomendación basado en contenido agrupado. Las Figuras 7a y 7b muestran estas curvas para los diferentes algoritmos comparados.



(a) Curvas de precisión vs. recall para el modelo basado en contenido de usuarios agrupados.



(b) Curva ROC para el modelo basado en contenido de usuarios agrupados.

Fig. 7. Precisión y cobertura en modelos de recomendación basados en contenido.

El análisis gráfico complementa la evaluación cuantitativa del modelo LinearSVR agrupado. En la Figura 7b, se observa que la curva ROC presenta un área bajo la curva (AUC) de 0.603, lo cual indica que el modelo tiene una capacidad de discriminación apenas superior a la del azar. Por otro lado, la Figura 7a muestra una caída progresiva en la precisión a medida que aumenta el recall, lo que sugiere que el modelo tiende a hacer más predicciones positivas, pero a costa de una mayor proporción de falsos positivos.

Ambas gráficas reflejan la limitación de los modelos agrupados para capturar adecuadamente las preferencias individuales de los usuarios, reforzando la conclusión previa de que los modelos personalizados —aunque más costosos computacionalmente— resultan más precisos y útiles para sistemas de recomendación centrados en el usuario.

D. Conclusión

Los resultados obtenidos muestran que el modelo basado en factorización matricial (SVD) logra el mejor rendimiento global entre los enfoques evaluados. Como se observa en la Tabla 3, este modelo obtiene el menor error cuadrático medio (RMSE = 0.9462), superando tanto a los modelos clásicos basados en popularidad como a los enfoques colaborativos tradicionales. Esto confirma su capacidad para capturar patrones latentes complejos en las interacciones usuario-ítem que otros métodos no alcanzan a modelar.

Además de su precisión, el modelo SVD presenta buenas propiedades de escalabilidad y permite representar tanto a usuarios como a ítems en un espacio latente compacto, facilitando la interpretación de preferencias implícitas. Si bien su entrenamiento puede ser más costoso, una vez optimizado, genera recomendaciones de manera eficiente mediante productos escalares en un espacio de baja dimensión.

Por su parte, los métodos basados en vecinos, como el filtrado colaborativo por usuario, tienden a ofrecer mejor cobertura en valores bajos de k, lo que puede ser útil en sistemas que prioricen diversidad, aunque lo hacen con menor precisión y mayores requerimientos de cómputo en tiempo de consulta.

En cuanto al modelo de recomendación basado en contenido implementado mediante `LinearSVR`, se observa que ofrece una alternativa valiosa cuando se busca interpretabilidad y personalización explícita. Al entrenar un modelo por usuario o grupo de usuarios, este enfoque permite identificar directamente la influencia de cada género sobre las valoraciones. El análisis mostró que los modelos personalizados presentan un mejor RMSE (por ejemplo, 0.613 para un usuario concreto), mientras que los modelos agrupados obtienen un RMSE promedio en el rango de 0.998 a 1.086.

Este rendimiento inferior en los modelos agrupados sugiere que, aunque generalizables, no capturan adecuadamente las particularidades individuales. No obstante, su capacidad para generar recomendaciones interpretables y su utilidad en escenarios de inicio en frío o sin historial colaborativo los convierten en un complemento útil a los enfoques latentes.

En conclusión, la factorización matricial constituye la opción más eficaz cuando se dispone de datos abundantes y se prioriza la precisión, mientras que el modelo basado en contenido con `LinearSVR` aporta valor añadido en términos de interpretabilidad, transparencia y adaptabilidad en contextos con menor densidad de datos o mayor necesidad de explicabilidad. La combinación de ambos enfoques, en un sistema híbrido, podría representar una línea prometedora para trabajos futuros.