Use of Latent Factors and Consumption Patterns for the Construction of a Recommender System

H. Alatrista-Salas, *Member, IEEE,* I. Hoyos, A. Luna, *Member, IEEE,* and M. Nunez-del-Prado, *Member, IEEE*

***Abstract*—In recent years, the recommender systems have become essential tools for companies to offer their products in a personalized way and to improve the user experience. The primary objective of these systems is to propose products or services to the user, according to specific criteria, such as their interests, their preferences, the place where they work or where they live. The problem arises when the system recommends products from an establishment to users who never visited that establishment. Besides, it is known that the order in which users purchase certain products or services can impact on the recommendation. To deal with these two problems, we propose a process that combines two widely used models: latent factors and matrix factorization. Also, to include temporality in our results, we use the *Sequitur* algorithm. In order to test our proposal, we have used a database with approximately 65 million banking transactions. The results obtained show the efficiency of our proposal in terms of average consumption ticket increase.**

***Index Terms*—Recommender system, Latent Factors, Consumption Patterns, Matrix factorization.**

I. Introducción

Debido a la gran cantidad de productos o servicios existentes en los mercados de hoy, resulta difícil para los usuarios encontrar los artículos que realmente necesitan de una manera fácil y rápida. Por otro lado, las empresas son conscientes que los clientes son su posesión más importante. Actualmente, no pueden existir empresas prósperas sin clientes satisfechos, leales y que desarrollen una relación estrecha con la organización. Por tal motivo, una organización debe emplear ciertas estrategias para tratar con sus clientes, los cuales pueden ser descritos por datos personales sobre su actividad comercial, sus preferencias, el lugar donde trabajan, entre otros.

En este contexto, las empresas recopilan gran cantidad de información de sus clientes a fin de que pueda ser utilizada en beneficio de ellos. Estos datos son transformados en información útil que permita a los clientes encontrar el producto de su interés de acuerdo a sus gustos, intereses, necesidades, contexto y preferencias. Estos instrumentos son los que se conocen con el nombre de sistemas de recomendación y ayudan a filtrar la información de manera personalizada y transparente para el usuario.

Por ejemplo, Netflix, la popular plataforma de *streaming* de películas y series, usa un sistema de recomendación que analiza grandes volúmenes de información acerca de qué ve un usuario, a qué hora, durante cuánto tiempo y en qué dispositivo lo hace. Después de analizar los datos, los algoritmos de Netflix recomiendan al usuario qué ver a partir de esta información, presentando las recomendaciones en la página de inicio [1]. El objetivo de los algoritmos de recomendación que utiliza Netflix es compensar el bajo poder de decisión del ser humano ante una gran gama de opciones [2].

Gallego y Huecas [3] describen otro ejemplo de sistemas de recomendación en un contexto financiero y que fue implementado en España. El objetivo es recomendar establecimientos tomando en cuenta las dinámicas espacial y temporal de los clientes. Este sistema busca recomendar entidades donde los clientes del banco puedan pagar con sus tarjetas bancarias. Para tal fin, se hace uso de grandes volúmenes de datos que se tienen acerca de los clientes del banco, de las transacciones realizadas por los clientes y de los establecimientos que visitaron. La recomendación busca que el cliente haga uso de su tarjeta y conozca dónde usarla de acuerdo a la información espacial existente. La principal diferencia de nuestro enfoque con este trabajo es que nosotros agregamos la temporalidad, patrón y orden de las compras.

En ese contexto, el presente artículo describe el proceso de construcción de un sistema de recomendación. El objetivo de nuestra propuesta es fomentar el uso de tarjetas de crédito y de débito de una entidad financiera. En esta entidad financiera, del total de personas naturales registradas en el mes de julio de 2017, se identificaron 3 millones de usuarios de tarjetas de crédito y débito. Como consecuencia directa del fomento de uso se pretende indirectamente aumentar el ticket promedio de compra de los clientes. Este punto en particular es una de las motivaciones de este trabajo, que es generar un sistema de recomendación que promueva el uso de tarjetas en establecimientos que los usuarios nunca visitaron.

El resto del artículo está estructurado como se detalla a continuación. La Sección II muestra el estado del arte y la Sección III describe los fundamentos teóricos de nuestra propuesta. Luego, la Sección IV describe los experimentos y se presentan los resultados obtenidos. En la Sección V se detalla la validación de los resultados y finalmente, en la Sección VI se esbozan las conclusiones y los trabajos futuros.

II. Estado del arte

Existen varios estudios sobre sistemas de recomendación, dentro de los cuales se identifican y diferencian tres enfoques, *basado en contenido*, *filtrado colaborativo*, e *híbrido*. El primero de ellos considera los perfiles de los usuarios y representaciones de los ítems, construidos a partir de términos descriptores de dichos ítems [4]. Sin embargo, la captura de estos últimos requiere reunir información externa que puede no estar disponible o que es difícil de colectar. Además, las

Universidad del Pacífico, Av. Salaverry 2020, Jesús María, Lima, Peru.

The authors are listed in alphabetical order and all contributed equally to the present article.

recomendaciones tienden a sobre-especializarse, de tal manera que se vuelve menos probable obtener una recomendación que sea novedosa y útil. Un ejemplo del uso de enfoque por contenido se encuentra en el Proyecto de Genoma Musical (Genome Musical Project), en el cual se recomiendan canciones basadas en 400 características musicales que capturan la "identidad" musical de una canción [5].

El segundo enfoque es el *filtrado colaborativo* y tiene en cuenta el comportamiento histórico del usuario, tal como las valuaciones realizadas a ítems o adquisiciones previas. Este tratamiento usa las preferencias conocidas de los usuarios para predecir las desconocidas y efectuar una recomendación. Existen dos métodos dentro de este enfoque: 1) el *método de vecindario*, basado la similitud entre los ítems y los usuarios [6]; y 2) el modelo de *factores latentes*, el cual consiste en representar a los usuarios y a los ítems a través de factores o características [7]. Dentro del método de factores latentes se encuentra el trabajo de Koren *et al.* [8], quienes desarrollaron un sistema de recomendación para la competencia de Netflix basado en modelos de factores latentes, factorización de matrices y el uso de mínimos cuadrados alternantes (MCA). Hace aproximadamente un par de años, Covington *et al.* [9] describieron el sistema de recomendación de Youtube usando redes neuronales profundas y métodos de vecindario para procesar las búsquedas. En entornos bancarios, se han desarrollado sistemas de recomendación de establecimientos en los cuales los clientes usan las tarjetas de débito o crédito teniendo en cuenta los contextos espacial y temporal del cliente y su similitud con otros clientes al visitar establecimientos comerciales [3]; así como sistemas de recomendación para servicios bancarios teniendo en cuenta el canal de acceso del cliente [6].

Finalmente, la literatura muestra el enfoque *híbrido*, en el cual se combinan las características del filtro colaborativo y del filtro basado en contenido. Adicionalmente, estos enfoques hacen frente al problema de arranque en frío del filtro colaborativo y a la sobre-especialización de las técnicas basadas en contenido. Burke [10] detalla los distintos métodos para unir los enfoques basados en contenido y filtro colaborativo existentes; por ejemplo: ponderación, intercambio, mezcla, entre otros. Para poder hacer las recomendaciones, el sistema necesita datos que le permitan entender las preferencias de los usuarios, quienes expresan su preferencia hacia un ítem mediante una valuación, la cual puede ser explícita o implícita. La valuación explícita indica claramente la preferencia del usuario mediante una escala, tal como asignarle a un establecimiento de 1 a 5 estrellas para valuarlo. Por otro lado, la valuación implícita se calcula al analizar el comportamiento del usuario en el dominio del sistema; por ejemplo, la cantidad de tiempo destinada a observar una pestaña y el monto gastado en establecimientos. Teniendo en cuenta los pros y contras de los métodos antes descritos, nuestra propuesta se basará en la combinación de los modelos de factores latentes y factorización de matrices, los cuales se escogieron por el gran tamaño y raleza de la matriz a operar. Además, para mejorar la eficacia de nuestra propuesta, hemos implementado el algoritmo *Sequitur* [11], el cual construye secuencias de consumo de los clientes.

III. Proceso de construcción de un sistema de recomendación

La presente propuesta se encuentra divida en seis fases, tal como lo muestra la Figura 1. Primero, se obtiene una base de datos que contiene las transacciones de los clientes efectuadas con tarjeta de crédito o tarjeta de débito. Luego, se filtran aquellas transacciones que no pertenezcan a un establecimiento y se dividen las transacciones según el área (distrito) al que pertenece el establecimiento. Tanto los datos como el proceso de filtrado serán descritos en la sección IV. Luego, usando modelos de factores latentes, factorización de matrices y mínimos cuadrados alternantes, se predicen las valuaciones desconocidas de los clientes. Dichas técnicas son detalladas en las Subsecciones III-A, III-B y III-C, respectivamente. En paralelo, se ejecuta el algoritmo Sequitur, descrito en la Subsección III-D, para calcular los patrones de consumo dentro de las secuencias de transacciones de los clientes. A partir de esto, se obtienen todos los *ratings* de los clientes por establecimiento, así como los establecimientos recomendados. Finalmente, estos últimos son graficados en una cartografía.

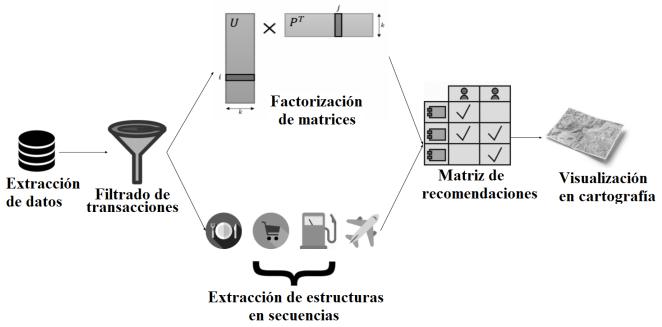


Fig. 1. Esquema de funcionamiento del sistema recomendador.

A. Modelos de Factores Latentes

Los modelos de factores latentes caracterizan a los ítems y a los usuarios mediante factores inferidos de los *ratings*, con el fin de explicar las valuaciones de los ítems [7]. Las técnicas de factorización de matrices son clave para estos modelos, pues hallan los valores de los factores latentes mediante los cuales se calculan las valuaciones desconocidas. Cada usuario y cada ítem está asociado a un vector individual de ciertas características. A partir de ese vector cada usuario y cada ítem, es modelado dentro de un espacio de características. Usando el producto escalar de estos vectores individuales, se puede estimar la valuación de un usuario a un ítem. Para el caso de un ítem, el vector de características indica en qué medida dicho ítem posee un determinado factor como consecuencia de un puntaje asignado. Mientras más alto sea el puntaje, en mayor medida se caracteriza al ítem. Por ejemplo, una película que tenga un puntaje alto en la característica "Drama" será una película de género dramático.

B. Factorización de Matrices

Apoyándose en los modelos de factores latentes, la factorización de matrices modela las interacciones entre usuarios e

ítems como el producto escalar entre ambos. Este enfoque es el más preciso al momento de afrontar la dispersión de datos en las matrices de las valuaciones [12]. A continuación, se detalla un modelo básico de factorización de matrices. Dado $u_i \in U$, que representa el vector de características del i-ésimo usuario; y dado $m_j \in M$, que representa el vector de características del j-ésimo ítem, se tiene que el producto interno $u_i^T m_j$ es la interacción entre los respectivos usuarios e ítems, donde u_i^T es la traspuesta de u_i. Debido a que la interacción ítem-usuario se da mediante la valuación del usuario al ítem, la valuación del i-ésimo usuario al j-ésimo ítem se expresa mediante la Ecuación 1.

$$r_{ij} = u_i^T m_j \tag{1}$$

A fin de calcular los valores para las matrices U y M, se debe aproximar la matriz R minimizando una función de pérdida. En ese sentido, dado un par de matrices U y M, la pérdida total del modelo será la suma de todas las pérdidas en todas los valuaciones conocidas, lo cual da como resultado el error cuadrático medio (MSE por sus siglas en inglés, *c.f.*, Ecuación 2).

$$f(R, U, M) = \frac{1}{n} \sum_{i,j} (r_{i,j} - u_i^T \times m_j)^2. \tag{2}$$

Al momento de valuar el error, se debe evitar contabilizar ítems que no tienen valuaciones; por tal motivo, a la Ecuación 2 se le agrega una matriz W de pesos de igual dimensión que la matriz R. Si $r_{i,j}$ existe, entonces $w_{i,j}$ toma el valor de 1, en caso contrario, su valor es 0. Por lo tanto, se puede reescribir la función de pérdida como se muestra en la Ecuación 3.

$$f(R, W, U, M) = \frac{1}{n} \sum_{i,j} w_{i,j} (r_{i,j} - u_i^T \times m_j)^2. \tag{3}$$

Entonces, se puede formular el problema de hallar los valores de U y M que mejor aproximan la matriz R a través de la Ecuación 4.

$$(U, M) = \arg\min_{(U,M)} f(R, W, U, M), \tag{4}$$

C. Mínimos Cuadrados Alternantes

Es un método de optimización para resolver la Ecuación 3. Puesto que se tiene una función no convexa, al desconocer tanto los valores de *U* como los de *M*, se debe resolver la ecuación para cada incógnita por separado [13]. Además, dado que se tienen $(i+j) \times k$ parámetros desconocidos que se deben calcular y una matriz R rala, resolver la Ecuación 3 puede llevar al sobreajuste de los datos. Por ello, a la Ecuación 3 se le agrega un término adicional usando el método de la regularización de Tikhonov, como se observa en la Ecuación 5.

$$f(R, W, U, M) = \sum_{(i,j)} w_{(i,j)} (r_{(i,j)} - u_i^T \times m_j)^2 + \lambda(\sum_i n_{u_i} ||u_i||^2 + \sum_j n_{m_j} ||m_j||^2), \tag{5}$$

Siendo n_{u_i} y n_{m_j} el número de valuaciones que tienen el usuario i sobre el ítem j, respectivamente. Usando la Ecuación 5 como la función a minimizar, la manera de proceder usando mínimos cuadrados alternantes es la siguiente:

- Inicializar *M* aleatoriamente con valores entre 1 y 0
- Fijar *M*, y derivar parcialmente respecto a *U*
- Fijar *U*, y derivar parcialmente respecto a *M*
- Repetir 2do y 3er paso hasta llegar al criterio preestablecido de detención.

El criterio que se fijó fue que la diferencia entre la raíz del *MSE* de la iteración *n* y el de la iteración $n-1$ sea menor a un umbral dado. Al igual que en el estudio de Zhou *et al.* [13], el valor del umbral que utilizamos es de 10^{-4}.

D. Algoritmo Sequitur

Es un algoritmo que infiere una estructura jerárquica de una secuencia de símbolos. Los digramas, o grupos de dos símbolos, que aparezcan más de una vez pueden ser reemplazados por un símbolo no terminal que indica una regla gramatical. Este proceso continúa de manera recursiva, por lo que se obtiene una representación jerárquica de la secuencia original [14]. En el trabajo de Di Clemente *et al.* [11], se usa el algoritmo para hallar estilos de vida de los clientes al analizar las transacciones de sus respectivas tarjetas de crédito. Este algoritmo obedece a dos principios que actúan como restricciones: 1) principio de unicidad, donde cada digrama o grupo de dos símbolos debe ser único en la secuencia; y, 2) principio de utilidad, donde cada regla hallada debe ser usada más de una vez.

En la sección siguiente se describe el proceso mostrado en la Figura 1 cuando se utiliza un conjunto de transacciones bancarias.

IV. Experimentos y resultados

Los datos que se emplearon en este trabajo están asociados a transacciones bancarias registradas durante un año, específicamente entre los meses de junio de 2016 a julio de 2017. En total, se han contabilizado 65 085 138 transacciones, de las cuales cerca del 80% corresponde a transacciones realizadas con tarjeta de débito, mientras que el el 20% restante corresponde a las realizadas con tarjetas de crédito. A su vez, se tienen 25 variables que proveen información sobre el cliente y dueño de la tarjeta, sobre el establecimiento en el que se realizó la transacción y sobre la transacción *per se*. Para nuestras experimentaciones se seleccionaron las variables *CODCLIENTE*, *CODCOMERCIO*, *CATEGORIA*, y *MONTO_RATING*. Esta última se calcula a partir de la Ecuación 6, que es el cociente entre el valor de la compra realizada por un cliente en determinado establecimiento y el monto total gastado por dicho cliente. Esta variable representa las valuaciones de los clientes hacia los establecimientos.

$$monto\_rating_{i,j} = \frac{monto\_comercio_j}{monto\_total_i} \tag{6}$$

En una primera instancia, los datos son agrupados por regiones (distritos) en los cuales se efectuaron las transacciones comerciales. Por cada región se genera la matriz de

valuaciones R, la cual es una matriz rala. Dado que se tienen muchas valuaciones desconocidas, se estima la matriz R^* mediante factorización de matrices y mínimos cuadrados alternantes, hallando los valores de la matriz de usuarios U y de la matriz de ítems M, asegurándose de que el producto escalar entre ambos se aproxime a R. En R^*, la valuación del cliente i al establecimiento j está representada por r_{ij}. Con el fin de asegurar que se cumpla esta restricción, se fijó que la diferencia de error entre cada iteración sea de 10^{-4}. Para procesar los datos, se usó el lenguaje de programación Python y una instancia de Elastic Compute Cloud (EC2) de Amazon Web Services (AWS). Una vez calculadas las matrices, las filas se normalizan y los n establecimientos con mayores valuaciones son recomendados. Cabe resaltar que se utilizó *EC2*[1] por ser un recurso de calculo bajo demanda que permite pagar solamente por los cálculos efectuados. En lo que respecta a la sensibilidad de los datos, escogimos los servicios de *EC2* de *Amazon Web Service* por sus políticas de privacidad y confidencialidad[2].

En lo que respecta a los patrones de consumo, cada establecimiento está clasificado dentro uno de los 235 rubros comerciales que el banco ha establecido. Luego, aquellos rubros que sean afines son agrupados dentro de categorías comerciales. De esta manera, de los 235 rubros se obtienen 22 categorías. Estas, a su vez, están codificadas por un símbolo, tal como lo muestra el Cuadro I.

Posteriormente, se aplica el algoritmo Sequitur en la secuencia de categorías de los establecimientos en las que un cliente ha efectuado una transacción. Se construyen los patrones para cada cliente, así como la frecuencia de aparición de cada patrón. Tomando R^*, se realiza una ponderación (*c.f.*, Ecuación 7) con el fin de potenciar las recomendaciones de categorías de comercios más frecuentes. Por ejemplo, si Bob siempre toma un café después del almuerzo, cuando el recomendador identifique el pago en un restaurante propondrá con más relevancia comercios que vendan café.

$$r^*_{i,j} = r^*_{i,j} \times frecuencia\_patron \tag{7}$$

El paso siguiente es normalizar, por fila la matriz R^* y finalmente se procede a recomendar los n establecimientos con mayores valuaciones.

A. Análisis y Discusión de Resultados

El distrito que se tomará para análisis es el de Barranca, ubicada en la provincia del mismo nombre y en el departamento de Lima. En la Figura 2, se muestra que tras 5 iteraciones, el MSE obtenido al minimizar la Ecuación 5 es de 0.0000026, con lo que se concluye que la matriz R converge a la matriz R^* rápidamente.

Luego, se muestran las valuaciones predichas en la matriz R^*. Puesto que se tienen 167 883 clientes, 19 949 establecimientos y más de 3 mil millones de valuaciones, el Cuadro II muestra, a modo de ejemplo, las valuaciones de 3 clientes a 3 establecimientos. A partir de esta matriz, se ordenan los

[1]EC2 docs.aws.amazon.com/es_es/AWSEC2/latest/UserGuide/concepts.html
[2]Privacidad AWS: aws.amazon.com/es/compliance/data-privacy-faq/

TABLA I
TABLA DE CATEGORÍAS

NOMBRE CATEGORÍA	COD	NOMBRE CATEGORÍA	COD
PRODSUPER	a	RESTBAR	b
SALUD	c	VEHIDER	d
TIENDEPART	e	ENTRENT	f
ROPMOD	g	BELLEZ	h
TELECOM	i	ALQBIEN	j
FINANC	k	PRODELECT	l
ARTCULT	m	ENSEÑ	n
PRODLOCAL	o	TRANSLI	p
DIVPROP	q	CLUBMA	r
HOGOFIC	s	PROFDIV	t
INFORM	u	RETMAN	v

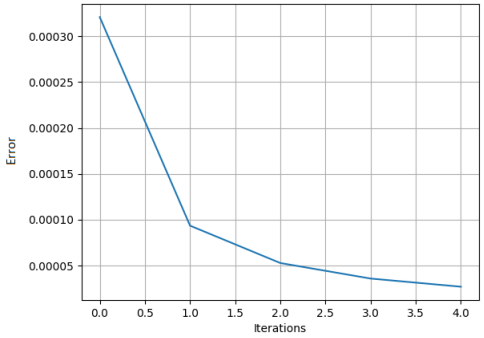


Fig. 2. MSE del distrito de Barranca

ratings de mayor a menor, y los establecimientos a los que correspondan los n mejores valuaciones son recomendados al respectivo cliente.

TABLA II
MATRIZ DE VALUACIONES PREDICHAS: DISTRITO DE BARRANCA

	100070157	100070158	101087053
Bob	0.8535	0.5671	0.6053
Alice	0.9306	0.8505	0.4561
Eve	0.7923	0.3194	0.4622

Luego, se calculan los patrones de consumo mediante el algoritmo Sequitur. El Cuadro III presenta una muestra de los patrones hallados para algunos clientes a partir la secuencia de las categorías de los establecimientos donde se ha realizado una transacción. La columna "Secuencia Transacciones" posee las categorías de dichos establecimientos. A partir de ésta, se obtiene la columna "Secuencia Sequitur", donde se observa dicha secuencia calculada después de la aplicación del algoritmo Sequitur y con los patrones reemplazados por números. Finalmente, la columna "Patrones" representa las categorías que conforman los patrones hallados a partir de "Secuencias Transacciones".

Posteriormente, se pondera la matriz R^* con la ayuda de la frecuencia de los patrones obtenidos con Sequitur. Finalmente, esta matriz es normalizada por filas y los resultados se muestran en el Cuadro IV.

Como se puede ver, al aplicar la ponderación y normalización, las valuaciones de los establecimientos varían. Debido

TABLA III
PATRONES SEQUITUR

Cliente	Secuencia Transacciones	Secuencia Sequitur	Patrones
Bob	a b b g d b b	a **1** b g d l	b b
Alice	f a a b b b g g g g c e d	f a a b b b **1 1** c e d	g g
Eve	f a a a a b b b c c c e	f **1 1** b b b c c c e	a a
Charlie	q q q q f l a a a b b b b g e	**1 1** f l a a a **2 2** g e	q q , b b

TABLA IV
MATRIZ PONDERADA DE VALUACIONES PREDICHOS - BARRANCA

	100070157	100070158	101087053
Bob	0.8535	0.5671	0.6053
Alice	0.5931	0.5450	0.5474
Eve	0.7923	0.3194	0.4622

a ello, los n establecimientos recomendados utilizando la ponderación pueden ser diferentes a los n establecimientos recomendados sin ponderación. Las valuaciones obtenidas al ponderar resultan mejores que aquellas que no la incluyen debido a que se está tomando en cuenta el contexto temporal de las transacciones, es decir, el orden en que fueron realizadas por el cliente.

La Figura 3 muestra los lugares visitados previamente y los lugares recomendados para un cliente. Los establecimientos están señalados por marcadores y sus referencias se detallan en la figura.

Fig. 3. Visualización de establecimientos I

V. VALIDACIÓN DEL SISTEMA DE RECOMENDACIÓN

Para la verificación y validación del recomendador se realizó una prueba preliminar y una validación en muestras correspondientes a algunos distritos de Lima.

A. Prueba Preliminar

Para la prueba del sistema de recomendación se realizaron encuestas sobre la aceptación de los resultados de las recomendaciones a 10 usuarios de tarjetas de crédito o débito de un distrito en particular. Los establecimientos recomendados para este distrito fueron calculados al predecir las valuaciones de $149,362$ clientes. Estos establecimientos fueron divididos en dos tipos: 1) recurrentes, lugares que el cliente ya había visitado antes, y 2) recomendados, lugares novedosos para el cliente. Mediante las encuestas se mide la aceptación de la recomendación por parte de los clientes con respecto a los establecimientos y rubros comerciales. En resumen, la encuesta fue personalizada para cada usuario en función de sus consumos históricos y recomendación de nuevos establecimientos.

En el Cuadro V se observa el porcentaje de recomendaciones de establecimientos y rubros, tanto recurrentes como novedosos, en los que el cliente aceptaría realizar algún consumo.

TABLA V
RESULTADOS DE LA PRUEBA

	Recurrente	Novedoso
Rubros	77.0%	76.0%
Comercios	84.5%	70.0%

En el caso de los rubros comerciales, la aceptación es de 77% y 76% para los establecimiento recurrentes y los establecimientos novedosos, respectivamente. Para el caso de los comercios *per se*, la aceptación es de 84.5% para los establecimientos recurrentes y de 70% para los establecimientos novedosos (*c.f.*, Cuadro V). Si bien el tamaño de la muestra tomada para la prueba es pequeño, este nos permitió probar la congruencia del recomendador en una prueba piloto. Podemos apreciar, a partir de los resultados, que el motor de recomendación es pertinente en un 76% en cuanto a los nuevos rubros que sugiere y un 70% con respecto a los nuevos establecimientos.

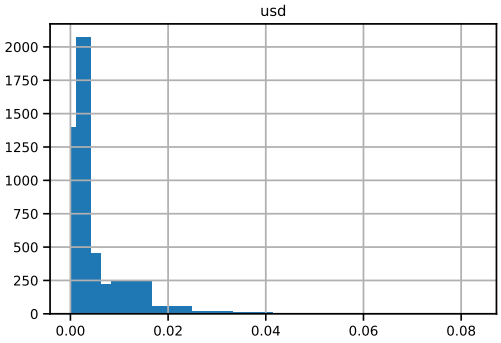
Esta primera prueba se realizó para verificar la coherencia de las recomendaciones ratificadas por los usuarios encuestados. Sin embargo, para una validación estadísticamente rigurosa se debe tener en cuenta el tamaño de la muestra y otros factores que se presentan a continuación.

B. Validación

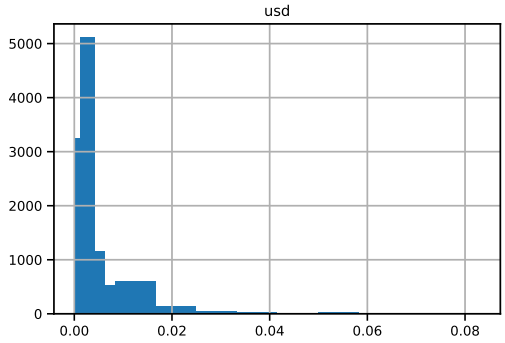
Para la validación se escogieron cuatro distritos de Lima: San Isidro, San Borja, Miraflores e Independencia. Para cada uno de ellos se tomó una muestra de clientes aleatoria, n = 4500. Por motivos de confidencialidad no podemos revelar el número exacto de clientes por distrito. Sin embargo, para dar una idea del orden de magnitud de clientes, en el Cuadro VI vemos que San Isidro, San Borja, Miraflores, e Independencia representan tres, cuatro, cinco y dos veces aproximadamente el tamaño de la muestra.

TABLA VI
RESUMEN DE RESULTADOS DE LA VALIDACIÓN DE LA MUESTRA

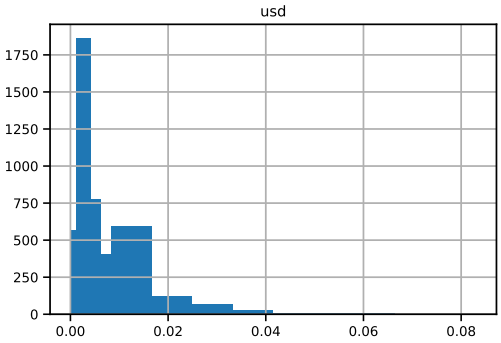
Distrito	San Isidro	San Borja	Miraflores	Independencia
Tamaño	3n	4n	5n	2n
Media de la muestra	0.0043	0.0071	0.0080	0.004
Media de la población	0.0042	0.0072	0.0079	0.003
Desviación estándar (s)	$< 1\%$	$< 1\%$	$< 1\%$	$< 7\%$
Test-t	-0.55	0.25	-0.36	1.72
KLD	0.0009	0.001	0.0018	0.0009
Tamaño mínimo de muestra	410	422	415	400



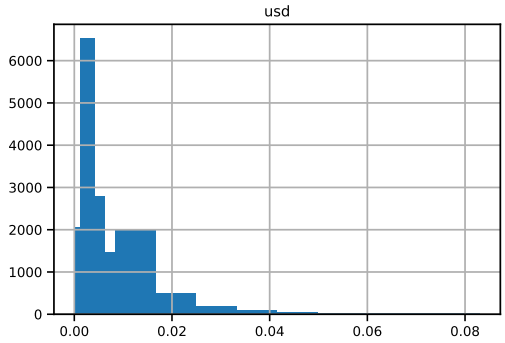
a) Distribución de la muestra de San Isidro



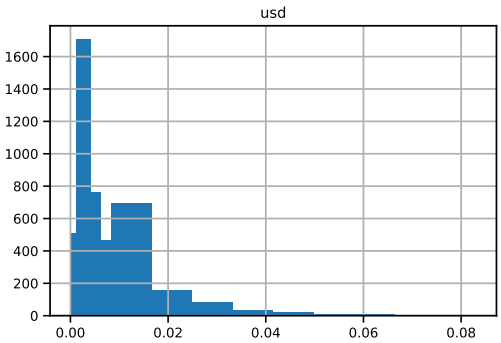
b) Distribución de la población de San Isidro



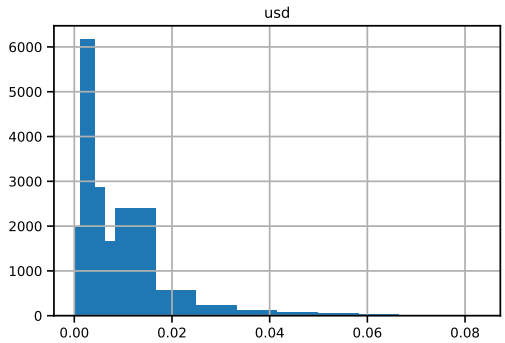
c) Distribución de la muestra de San Borja



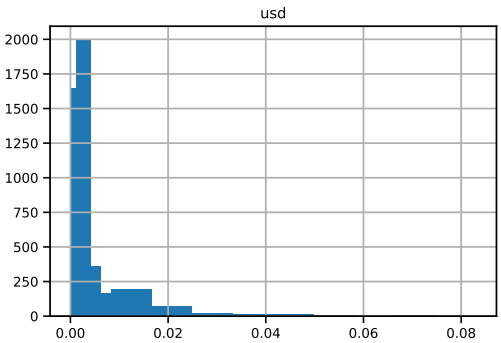
d) Distribución de la población de San Borja



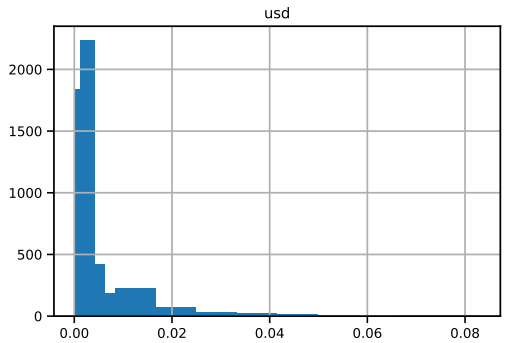
e) Distribución de la muestra de Miraflores



f) Distribución de la población de Miraflores



g) Distribución de la muestra de Independencia



h) Distribución de la población de Independencia

Fig. 4. Distribución de las muestras y poblaciones de diferente distritos.

Para comprobar que la muestra es representativa de la población realizamos tres pruebas diferentes. Primero, verificamos que las muestras tengan la misma media que sus respectivas poblaciones mediante la prueba de *One Sample t Test* [15]. Para efectuar esta prueba, se tomó una confianza de 5% (*i.e.*, α=0.05) con 4499 grados de libertad y un valor T de 1.9605 para una prueba de dos colas. La hipótesis nula que se planteó es que las medias de los consumos de la muestras y de la población son iguales $H_0 : \ \mu_0 - \mu = 0$. La hipótesis alternativa es que estas medias sean diferentes $H_{alt} : \ \mu_0 - \mu \neq 0$. Los valores calculados de la prueba t se encuentran en el Cuadro VI y fueron calculados usando la Ecuación 8.

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{8}$$

Donde $\overline{x}$ es la media de la población, μ es la media de muestra, s es la desviación estándar, y n es el tamaño de la muestra. Como resultado de la prueba, se acepta la hipótesis nula que señala que la media del consumo de la muestra y la población son iguales. Cabe resaltar que los valores de la desviación estándar son bajos y están expresados en porcentaje en relación a la media de la muestra por razones de confidencialidad.

El segundo método utiliza la divergencia de *Kullback-Leibler* para mostrar que a partir de la distribución de la muestra p, podemos reconstruir la distribución de la población q [16]. Para ello utilizamos la Ecuación 9. Las distribuciones que se comparan son la que se muestran en la Figura 4.

$$KLD = \sum_{k=1}^{n} p_k \times ln(\frac{(p_k)^2}{q_k}) \tag{9}$$

Así el Cuadro VI muestra los valores obtenidos para la divergencia. Nótese que los valores cercanos a cero significan que no hay divergencia, es decir, las distribuciones son iguales.

Finalmente, se calculó el tamaño de la muestra ($n*$) necesaria para tener una representación estadísticamente significativa. Al conocer el número de clientes en cada uno de los distritos, empleamos la fórmula de la Ecuación 10 para calcular el valor de $n*$.

$$n* = \frac{N \times (Z)^2 \times p \times q}{(e)^2 \times (N-1) + (Z)^2 \times p \times q} \tag{10}$$

En donde, N es el tamaño de la población, Z es la desviación del valor medio que aceptamos para lograr el nivel de confianza deseado (95%), en nuestro caso Z = 1.96, p es la probabilidad de éxito o proporción que esperamos encontrar (p=50%), q representa la probabilidad de fracaso (*i.e.*, $q = 1 - p$) y e es el margen de error máximo que admitimos (5%). Para preservar la confidencialidad del número exacto de clientes se completó el Cuadro VI sobrestimando los valores del tamaño de la muestra $n*$. En todos los casos, los resultados obtenidos son un orden de magnitud menor al valor real tomado para cada distrito.

Una vez demostrado que las muestras tomadas son estadísticamente representativas de sus respectivas poblaciones, pasamos a describir los resultados del motor de recomendación. La validación del motor se realizó utilizando los clientes seleccionados como muestra, a quienes se les envió ofertas personalizadas en dos rubros retail y restaurantes durante las tres primeras semanas del mes de marzo de 2018. Al final de dicho mes se compararon los promedio de los tickets de compra de la muestra y de la población sin los usuarios en la muestra y se observó que hubo un aumento del ticket de consumo promedio en retailers y restaurantes de 10% y 22%, respectivamente. Esto significó un incremento en gastos de 33.6 USD a 37 USD en retailers y de 27.88 USD a 34.24 USD en restaurantes. Estos resultados nos muestran la pertinencia del motor de recomendación y su impacto potencial en los establecimientos.

VI. Conclusiones y trabajos futuros

En este trabajo se desarrolló un sistema de recomendación para una entidad financiera. Se observó que el monto gastado en un comercio puede ser usado como una valuación implícita. Así, el monto refleja la preferencia del cliente por los establecimientos, de manera que a mayor monto gastado, mayor preferencia se tiene por dicha tienda o casa de comidas. De modo implícito, se tiene en cuenta el contexto geográfico del cliente al generar la agrupación por área (distrito), identificando dónde se realizaron las transacciones. La división geográfica distrital mantiene las recomendaciones dentro de un marco geográfico coherente evitando recomendar establecimientos en distritos distantes o lugares que se visitaron por única vez, por ejemplo, durante el periodo de vacaciones. Las recomendaciones logran ser novedosas, aunque no se descarta que establecimientos anteriormente visitados estén dentro de la lista de establecimientos recomendados.

Además, se ha usado el algoritmo Sequitur para hallar patrones secuenciales en la lista de transacciones del cliente, y con ellos, ponderar valuaciones para que los establecimientos cuyas categorías estén dentro de los patrones tengan mayor opción de ser recomendados. Dicho enfoque le añade conocimiento e información a las valuaciones, pues se beneficia a aquellos establecimientos que pertenezcan a una categoría en la cual es frecuente hacer transacciones. De esta manera, no sólo se tiene en cuenta el monto gastado en un establecimiento, sino el patrón de visitas del usuario a comercios similares.

En relación con los trabajos futuros, dado el alto volumen de datos con los que se trabaja, se planea usar el *framework* de Apache Spark integrado para implementar análisis avanzados, como por ejemplo procesar los datos de manera paralela y trabajar en un entorno compartido. Esto puede ser ejecutado en *clusters* de Elastic Map Reduce (EMR) usando PySpark o Scala. También, se debe tener en consideración la privacidad de los usuarios y de sus datos, con los que se ejecutan las recomendaciones. En consecuencia, se planea implementar una capa adicional al motor de recomendación para poder garantizar un cierto nivel de privacidad. Finalmente, se piensa realizar un piloto a gran escala con diez mil usuarios para validar el sistema recomendador y tener una mejor medida de precisión real y no subjetiva.

AGRADECIMIENTOS

Los autores agradecen el financiamiento del Vicerectorado de investigación de la Universidad del Pacífico en la subvención del proyecto PY-ESP-0210013216.

REFERENCIAS

[1] Carlos A. Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.*, 6(4):13:1–13:19, December 2015.

[2] Antti Oulasvirta, Janne P. Hukkinen, and Barry Schwartz. When more is less: The paradox of choice in search engine use. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 516–523, New York, NY, USA, 2009.

[3] Daniel Gallego and Gabriel Huecas. An empirical case of a context-aware mobile recommender system in a banking environment. In *Proceedings of the 2012 Third FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing*, MUSIC '12, pages 13–20, Washington, DC, USA, 2012.

[4] Charu C. Aggarwal. *Recommender Systems: The Textbook*. 1st edition, 2016.

[5] Michael Castelluccio. The music genome project. *Strategic Finance*, 88(6):57–58, 12 2006.

[6] H. Abdollahpouri and A. Abdollahpouri. An approach for personalization of banking services in multi-channel environment using memory-based collaborative filtering. In *The 5th Conference on Information and Knowledge Technology*, pages 208–213, May 2013.

[7] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, January 2009.

[8] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

[9] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 191–198, New York, NY, USA, 2016.

[10] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002.

[11] Riccardo Di Clemente, Miguel Luengo-Oroz, Matias Travizano, Bapu Vaitla, and Marta C. Gonzalez. Sequence of purchases in credit card data reveal life styles in urban populations. *Nature communications*, 8:1–33, 2017.

[12] Dheeraj Bokde, Sheetal Girase, and Debajyoti Mukhopadhyay. Matrix factorization model in collaborative filtering algorithms: A survey. *Procedia Computer Science*, 49(Supplement C):136 – 146, 2015. Proceedings of 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15).

[13] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. In *Proceedings of the 4th International Conference on Algorithmic Aspects in Information and Management*, AAIM '08, pages 337–348, 2008.

[14] Craig G. Nevill-Manning and Ian H. Witten. Identifying hierarchical structure in sequences: A linear-time algorithm. *J. Artif. Int. Res.*, 7(1):67–82, September 1997.

[15] Amanda Ross and Victor L Willson. *Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures*. Springer, 2018.

[16] James M Joyce. Kullback-Leibler divergence. *International encyclopedia of statistical science*, pages 720–722, 2011.

Hugo Alatrista-Salas es doctor en Ciencias de la Computación de la Universidad de Montpellier en Francia. Se interesa en la Minería de Datos espacio-temporal con aplicación en el medio ambiente y la salud pública. Además tiene una maestría en Calculabilidad, Algorítmica, Seguridad y Administración de Redes de la misma Universidad. Actualmente, es profesor investigador en la Universidad del Pacífico y vice decano del programa de Ingeniería de la Información.

Isaías Hoyos fue asistente de investigación en la Universidad del Pacífico. Es Ingeniero de la Información en la Universidad del Pacífico y actualmente trabaja como Científico de datos en Prestamype.

Ana Luna es profesora en la Universidad del Pacífico. Realizó su doctorado y su licenciatura en Ciencias Físicas en la Universidad de Buenos Aires (UBA), Argentina. Se especializó en el área de fotónica y en la actualidad está incursionando en técnicas estándar de recomendación y en la privacidad de datos.

Miguel Nunez-del-Prado doctor en informática por la Universidad de Toulouse. Obtuvo este título por su trabajo sobre ataques de inferencia en datos geolocalizados y su impacto en la privacidad de los usuarios en el LAAS-CNRS Francia.Es Ingeniero en Computación, Redes y Telecomunicaciones. Tiene dos maestrías, una en Informática y Telecomunicaciones y otra en Gestión estratégica de la Innovación. Trabajó como científico de datos en el Grupo INTERSEC (París, Francia).