UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DESARROLLO DE UN CLÚSTER DE COMPORTAMIENTO SEGÚN POSICIÓN PARA LAS PRODUCT LISTING PAGES DE UN E-COMMERCE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

PABLO ENRIQUE LASTRA BACHMANN

PROFESORA GUÍA:
CAROLINA SEGOVIA RIQUELME

MIEMBROS DE LA COMISIÓN:
PABLO MARÍN VICUÑA
RODOLFO URRUTIA URIBE

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: **PABLO ENRIQUE LASTRA BACHMANN**
FECHA: 2021
PROF. GUÍA: CAROLINA SEGOVIA R

DESARROLLO DE UN CLÚSTER DE COMPORTAMIENTO SEGÚN POSICIÓN PARA LAS PRODUCT LISTING PAGES DE UN E-COMMERCE

Según el último informe del World Economic Forum, se pronostica un aumento de la penetración del e-commerce en un 40 % para el año 2026 [1]. Además, la pandemia vivida en Chile y el mundo durante 2020 ha provocado estragos en empresas del retail, generando una disminución de sus ingresos en hasta un 25 % [2]. Dado esto, las tiendas por departamento están constantemente invirtiendo recursos para potenciar sus canales online, evaluando que metodologías resultan más beneficiosas para aumentar los ingresos de la compañía.

El trabajo es realizado en la empresa Falabella Retail S.A, y su objetivo es desarrollar una clusterización de las listas de productos (PLP) originadas por las categorías presentes en Falabella.com, según el comportamiento de los elementos que la componen, para evaluar y mejorar el desempeño del modelo utilizado en la optimización de contribución. Dicho modelo, actualmente genera un aumento en la contribución de un 3,61 %, lo que se traduce en ganancias para Falabella.com de aproximadamente $50.000.000.

La metodología utilizada consistirá en realizar un clúster jerárquico con medida de distancia basada en Correlación de Pearson, para agrupar PLP con distribuciones similares en sus KPI, a través de las posiciones de la lista. Luego, se procederá a entrenar un modelo de optimización que recibe como input los grupos de comportamiento similar y entrega como output un orden de productos que maximiza contribución, el cual será comparado de manera offline con el modelo actual, a través de la métrica NDCG (Normalized Discounted Cumulative Gain).

Los resultados de la clusterización dieron origen a 15 clúster para las 4000PLP activas en la página, revelando distribuciones de comportamiento de tipo Pareto, Uniforme y una combinación de las anteriores. Los conglomerados con más venta se identificaron con elementos de la categoría Tecnología, acumulando un 35 % del total de venta y presentando una distribución Pareto. Los clúster con mayor cantidad de unidades vendidas, se identificaron con elementos de las categorías Moda Hombre, Moda Mujer y Moda Niño, acumulando un 40 % del total de unidades vendidas y presentando una distribución más uniforme.

Finalmente, tras la modificación y el entrenamiento del modelo de optimización, los resultados obtenidos fueron disimiles en cada clúster. En específico, el rendimiento promedio para el NDCG de contribución disminuye entre ambas metodologías en un 20,6 %, y el rendimiento promedio para el NDCG de conversión lo hace en un 11,6 %. Además, se concluyó que el nivel de ruido en la distribución no es un factor determinante en el resultado. Sin perjuicio de lo anterior, se estudiaron las posibles causas del bajo rendimiento obtenido, generando 3 hipótesis, con los lineamientos necesarios para un desarrollo futuro.

Agradecimientos

A mis padres, por el apoyo incondicional en cada etapa de mi vida, por entregarme cariño, amor y todo lo que cualquier hijo desearía tener. Gracias por estar siempre presentes, con un abrazo cariñoso en momentos felices o una palabra de ayuda en momentos complejos. Gracias por estar siempre.

A mi hermana, por ser mi ejemplo a seguir, por ser quien me inspira a superarme, a pensar en grande y a ver la vida como un gran desafío que siempre tiene su recompensa. Gracias por tu cariño, tu apoyo y tus retos, aunque aveces son un poco duros, pero se que son de corazón.

A las personas que me permitieron desarrollar el trabajo de título, mis Profesores Guías Carolina y Pablo, gracias por entregarme las herramientas necesarias durante cada clase del año y los consejos precisos en cada feedback que pedí. A Andrés Lara Córdova, Data Scientist, contraparte de la empresa, gracias por el apoyo, los consejos y por compartir toda tu experiencia y conocimiento conmigo, incluso gastando de tu tiempo libre. Gracias también a Aquiles Martínez, ML Engineer, por el apoyo durante el trabajo y por permitirme entrar a la Célula.

A mis amigos, Marco, Benja, Angie, Yerko, Pipe, Nicole, Ani, Jose, Lucas, Agus, Nico, Eybie, Tere, Diego, Nati, por haber hecho que estar en la universidad sea algo que voy a extrañar, pero también por haber hecho que la vida universitaria haya valido totalmente la pena, con cada junta, carrete, cumpleaños y viajes, entre muchos otros.

Por último, pero no menos importante, a todas las personas que de alguna u otra forma, estuvieron presentes en mi paso por la universidad. A mis profesores, a quienes admiro, a esos amigos de plan común, con quienes tomamos caminos distintos y a las organizaciones estudiantiles en las cuales participé, y en algunos casos también lideré.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

En los últimos dos años, el ecommerce y el comercio digital influyen hasta en el 56 % de las ventas en tiendas por departamento [3]. Falabella está consciente de aquello, por lo que constantemente se encuentra innovando en la estructura de su página web, y en particular, en la forma en que sus productos son mostrados dentro de ella. Dado lo anterior, en Falabella.com es posible acceder a los productos ofrecidos por diferentes vías, siendo las más relevantes la pestaña Home y las listas de productos que se despliegan al ingresar a sus categorías.

Estas listas de productos se denominan PLP o *Product Listing Pages* (ver figura 2.1), y son conformadas por todos los productos disponibles para una categoría en particular, visualizados de forma ordenada y separados por página, cuando el navegador así lo requiere. Según esta visualización, cuando un usuario desea buscar productos específicos, navega a través de las PLP, generando interacciones con cada uno de ellos, agregaciones al carro de compra, y en el mejor de los casos tanto para Falabella como para el usuario, la navegación culmina con la venta del producto.

Como consecuencia de esto, el área de Data Science es la encargada de optimizar la visualización de estos productos, a través de la creación, aplicación y mantención de modelos de machine learning, que permitan mejorar la experiencia del usuario y aumentar las métricas claves del negocio, asociadas a la contribución y conversión.

En particular, actualmente existe un modelo global que busca maximizar la contribución (o margen) que se obtiene de la venta de productos a través de las PLP. Para esto, el modelo recoge información relacionada a la navegación en las listas de productos, atributos de estos, características propias de productos y una combinaciones de lo anterior, donde a través de Learning to Rank[1], utiliza el algoritmo XGBClassifier para clasificar los productos en 5 niveles de relevancia.

Posteriormente a través de una estrategia de tipo Pointewise, se asigna un score a cada producto, que se construye en base al nivel de relevancia asignado por el clasificador (0 nada relevante, 4 muy relevante) y la probabilidad de pertenencia de cada producto a cada uno de ellos. Finalmente, el modelo ordena los productos dentro de la PLP en base al score de cada producto, de mayor a menor.

[1] Estrategia algorítmica de ML de aprendizaje supervisado

Es en ese momento, que surge la interrogante acerca de como mejorar el modelo de optimización actualmente utilizado, en pos de aumentar la contribución de los productos vendidos. Una hipótesis, es entregar mejores inputs al modelo, en base a los datos transaccionales y de interacción con el usuario que ya son utilizados, pero agregando de forma explicita el comportamiento de estos datos a través de las posiciones de cada una de las 4000PLP activas en el sitio.

En el desarrollo presentado en los siguientes capítulos, se exponen los pasos metodológicos realizados para probar la hipótesis anterior, a través de la creación de un clúster que agrupe PLP con comportamiento similar a través de sus posiciones, permitiendo entrenar el modelo de optimización con la información transaccional, atributos y características propios de los productos ya presentes, pero separados en grupos de PLP con distribuciones similares de comportamiento, en pos de mejorar el rendimiento del modelo.

.

1.1. Antecedentes generales

1.1.1. Características de la empresa

El trabajo de título es desarrollado en la empresa Falabella Retail S.A. En la actualidad Falabella cuenta con presencia en 6 países de América Latina y se encuentra dividida en 6 unidades de negocio, *Falabella Retail, Mejoramiento del Hogar, Supermercados, Falabella Financiero, Retailment y Linio*

Figura 1.1: Unidades de negocio de Falabella S.A. Fuente: investors.falabella.com

La venta por departamento dentro del Holding, está a cargo de la unidad de negocio de Falabella Retail. Esta unidad se encarga de comercializar productos de diversa índole (vestuario, hogar, equipos electrónicos, útiles de aseo, juguetes, muebles, artículos de decoración, entre otros) a través de tiendas físicas y canales online, acorde a las necesidades de distintos segmentos de clientes.

La plana ejecutiva funciona a través de una estructura jerarquizada, al mando de un Gerente General Corporativo, quien maneja del Holding Latinoamericano en su conjunto, delegando en cada país responsabilidades a un Gerente General, quien a su vez tiene a cargo diferentes gerencias por área.

1.1.2. Desempeño y crecimiento

El desempeño de Falabella Retail Chile ha estado marcado por la presencia de factores externos, tales como la pandemia COVID19 y el pasado estallido social, propiciando un cambio de foco en la compañía hacia el ámbito online [4], y un cierre o disminución de sus canales de venta físicos [5]. Dado lo anterior, tener una mirada histórica de su desempeño permite generar líneas de trabajo, sobre las cuales implementar los cambios estratégicos del futuro. Para esto, se realizó un análisis general de desempeño comercial de los últimos 3 años en la división retail, utilizando datos extraídos directamente de la Gerencia de Inteligencia de Clientes.

Se obtuvo que las **ventas totales para 2017 fueron de $1.682.630 M, para 2018 de $1.655.344 M y para 2019 de $1.602.706 M**[2], mostrando un **decrecimiento en las ventas del 1,6 % para 2018** y un **decrecimiento de 3,2 % para 2019** (ver figura 1.2).

Así mismo, la **cantidad total de clientes para la compañía en 2017 fue de 5.770.791 personas, en 2018 de 5.785.127 personas y en 2019 un total de 5.777.921 personas**[3], constatando un **volumen constante de clientes** a través del tiempo (ver figura 1.3).

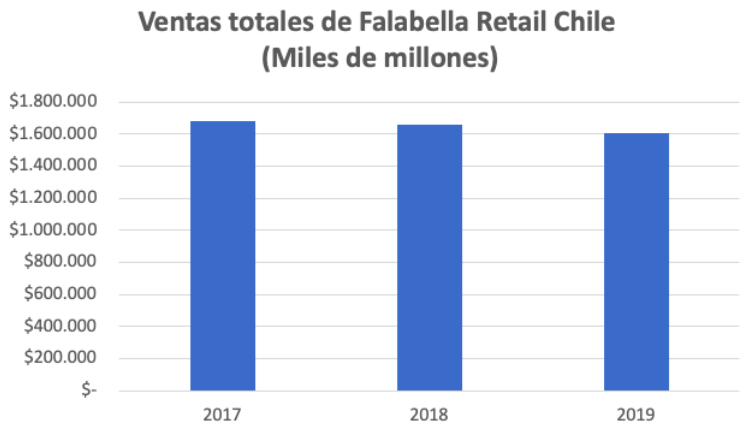


Figura 1.2: Ventas totales Falabella Retail Chile (Miles de millones). Fuente: Elaboración propia

[2] Elaboración propia a través de consultas a la BBDD de Falabella alojada en GCP
[3] Elaboración propia a través de consultas a la BBDD de Falabella alojada en GCP

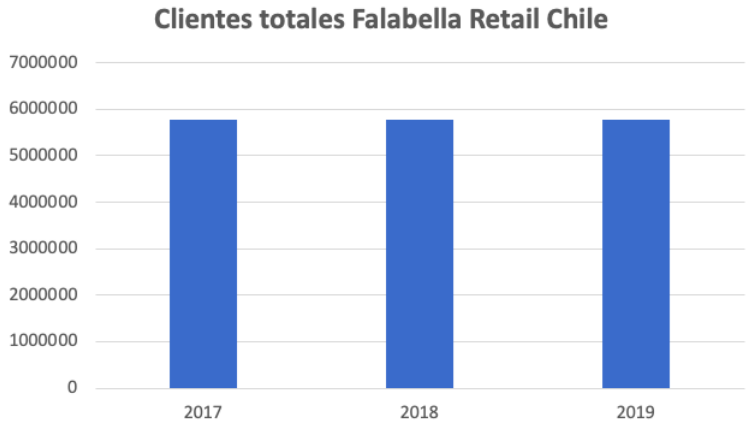


Figura 1.3: Clientes totales Falabella Retail Chile. Fuente: Elaboración propia

Por último, para obtener un mayor nivel de detalle en el desempeño de la compañía, se realizó el mismo análisis anterior, pero desagregado por tipo de canal de venta (Tienda Física o Falabella.com), obteniendo que **para 2018 la venta total correspondiente a Tienda fue de $975.760 M y para 2019 fue de $886.499 M** [4], notando una disminución de ingresos entre ambos años.

Para el caso de Falabella.com, en el año 2018 se registró una venta total de $366.959 M y en 2019 $432.327 M [5], escenario opuesto al anterior, mostrando un aumento en las ventas a través de este canal.

Posterior al análisis, queda de manifiesto que Falabella Retail Chile ha disminuido su nivel de ventas en el periodo estudiado, con foco en un aumento de los canales online, sumado a la mantención de una base constante de clientes. Sin perjuicio de lo anterior, es necesario mencionar que la empresa se encuentra en una etapa de madurez, abarcando el negocio del retail en 4 de los 6 países donde se encuentra presente.

1.1.3. Área de trabajo

El trabajo de título se desarrolló en el área de Data Science de Falabella Retail, específicamente dentro de la Célula de Relevancia PLP, la cual es dependiente de la Gerencia de Business Intelligence de la compañía. El objetivo de la gerencia, y por consiguiente del área, es aumentar la rentabilidad obtenida en la venta de productos, a través de la creación, recolección y transformación de información desestructurada en conocimiento, que permita tomar mejores decisiones de forma conjunta para tiendas físicas y canales online.

Puntualmente la Célula de Relevancia PLP, tiene como objetivo optimizar las métricas asociadas al orden en que se muestran los productos presentes en las categorías de Falabella.com, buscando aumentar su conversión, contribución y venta, a través de modelos de machine learning. El organigrama del área se encuentra en el Anexo A.1.

[4] Elaboración propia a través de consultas a la BBDD de Falabella alojada en GCP
[5] Elaboración propia a través de consultas a la BBDD de Falabella alojada en GCP

Capítulo 2

Planteamiento del problema y justificación

2.1. Oportunidad identificada

Los clientes que ingresan a Falabella.com, son redireccionados directamente a la pestaña Home del sitio, lugar en donde se encuentran las novedades presentes en la tienda, productos en oferta y elementos recomendados según la temporada. Si el usuario requiere un producto en específico, puede buscarlo de dos formas:

1. Seleccionando la categoría del producto que desea encontrar, en el menú desplegable de la página web
2. Buscando directamente el producto que desea visualizar, en el buscador integrado

En el primer ejercicio, el resultado mostrado en pantalla corresponde a una PLP o *Product Listing Page*, con todos los productos que pertenecen a la categoría seleccionada, ordenados según la recomendación que Falabella hace al cliente. (ver figura 2.1)

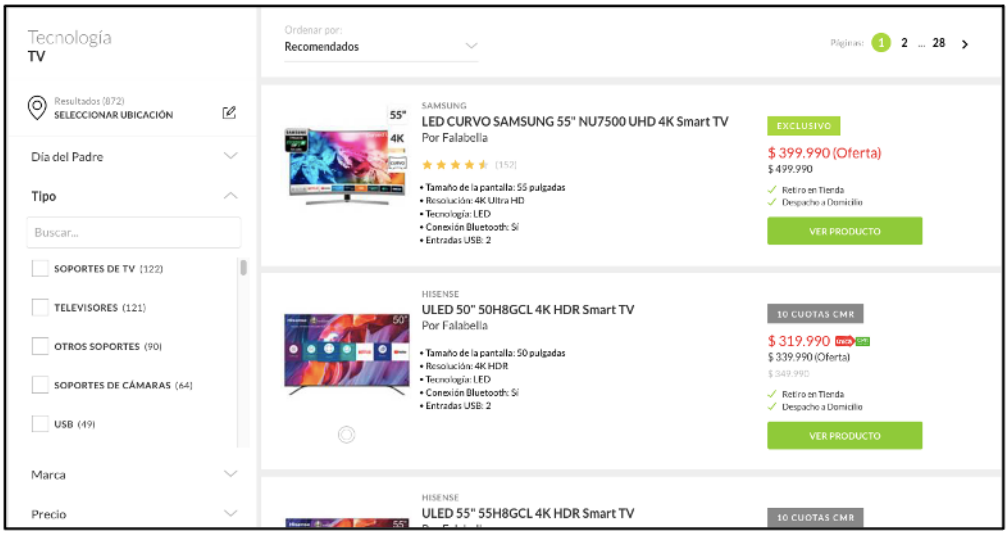


Figura 2.1: PLP de la categoría Smartphone. Fuente: Falabella.com

El segundo ejercicio muestra una SLP o Search Listing Page, con una estructura similar de lista de productos, donde se enumeran todos los que cumplen con las características buscadas.

Un potencial comprador, luego de analizar los productos mostrados en pantalla, realiza un benchmarking de acuerdo a sus preferencias y finalmente decide si concretar o desestimar la compra.

Considerando lo anterior, Falabella presenta la necesidad de ordenar los productos adecuadamente, buscando que los usuarios puedan encontrar el mejor producto acorde a sus intereses, y a la vez, que el producto vendido deje la mayor contribución (o margen)[1] a la compañía. Actualmente **las PLP concentran entre un 20 % y 30 % de las ventas totales de Falabella.com, lo cual se tradujo para julio de 2020 en aproximadamente $30.000.000.000 vendidos**.

Dada esta necesidad, la célula de relevancia PLP es la encargada de, a través de modelos de machine learning, generar, mantener y optimizar el orden de los productos dentro de Falabella.com. En específico, el objetivo de la célula es maximizar la contribución (o margen) que obtiene la compañía, al vender productos a través de las PLP. Para cumplir esta tarea, actualmente existe un modelo de optimización, cuyo target es la contribución, y opera sobre totalidad de PLP disponibles en el sitio.

Por otra parte, cuando los clientes navegan a través de las PLP, interactúan de forma constante con los productos presentes en ellas, permitiendo a Falabella registrar todas sus interacciones, junto con datos transaccionales e historial de comportamiento, en caso de completar una compra. El detalle de los datos registrados se encuentran en el anexo B.1

Al analizar los datos registrados, y en particular, realizar un análisis sobre el comportamiento de estos datos a través de las posiciones que ocupan los productos dentro de las PLP, se encontró que existen diferentes patrones de comportamiento, para diferentes listas de productos.

Para ilustrar de forma gráfica lo anterior, en la figura 2.2 se adjuntan las distribuciones presentes en las PLP originadas por las categorías de Moda Mujer y Ropa de Cama, para el KPI “Unidades vendidas”, medido a través de la posición en la que se encuentran los productos dentro de la lista, desde el 1 de agosto al 30 de agosto de 2020.

[1] Se define contribución como el dinero neto que resulta de restar el precio de venta menos los costos del producto

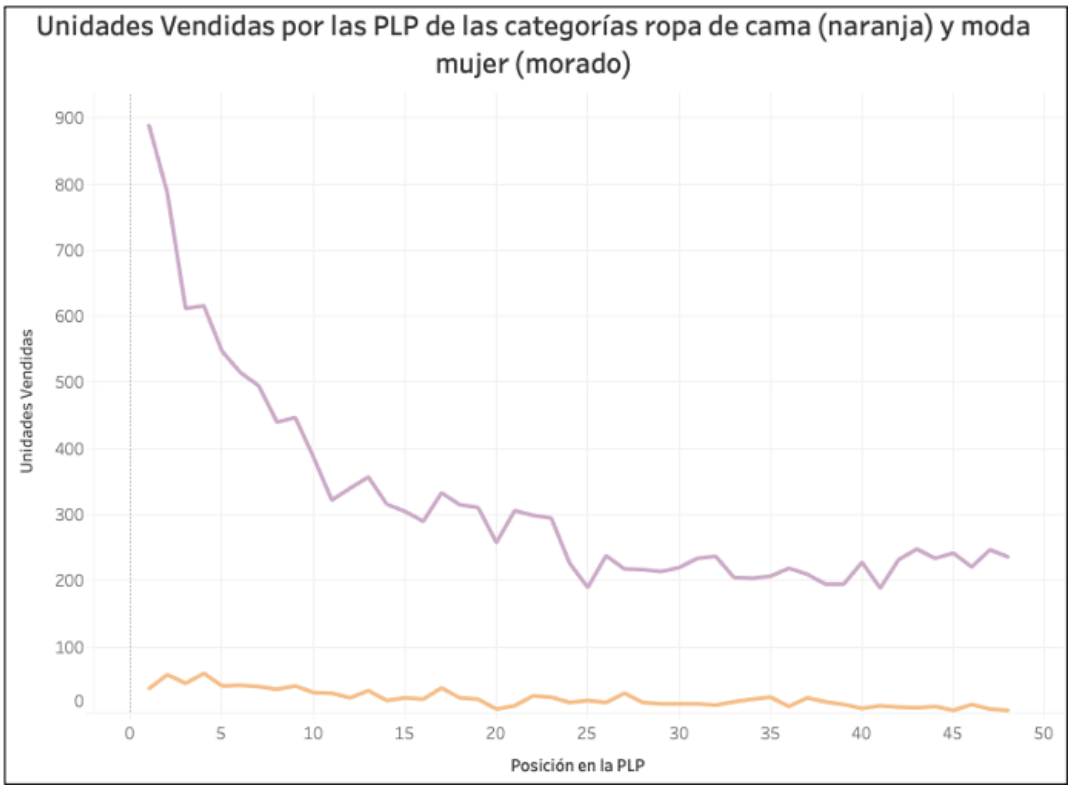


Figura 2.2: Unidades vendidas por las PLP de las categorías Ropa de cama (naranja) y Moda Mujer (morado). Fuente: Elaboración propia

Es en este punto, donde se encuentra la **oportunidad de mejora identificada, ya que el modelo de optimización actualmente utilizado, no considera las distribuciones de comportamiento por posición de forma explicita**, es decir, no considera este input de información para ordenar los productos dentro las PLP.

En consecuencia, bajo la hipótesis de aumentar el rendimiento del modelo actual, incorporando nuevos inputs, el objetivo de este trabajo de título es realizar una clusterización de PLP que presenten comportamiento similar a través de sus posiciones, para entregar de forma explicita esta información al modelo de optimización y evaluar como cambia su rendimiento en contribución, conversión y venta, a través de la métrica NDCG.

2.2. Propuesta de valor de la solución planteada

El impacto de la propuesta planteada para la oportunidad encontrada, es medible a través de métricas claves del negocio. Actualmente el modelo global de ordenamiento es monitoreado mediante un lift de contribución.[2]

El lift del modelo es calculado como el porcentaje de crecimiento o decrecimiento de una métrica, para usuarios que reciben una nueva campaña versus un grupo de control. En este caso, la métrica estudiada es la contribución, y la nueva campaña, es un nuevo orden de productos dentro de las PLP.

La métrica es medida comparando el orden del modelo actual versus el orden por defecto que presentaría la PLP sin ninguna optimización, a través de un experimento online (A/B test). En este escenario, el modelo actual, para el mes de julio del año 2020, generó un lift de contribución positivo de un 3,6 %, implicando una ganancia en ventas de aproximadamente $52.000.000 para Falabella.com [3]

La propuesta planteada, al generar un aumento en el lift de contribución de un 1 %, a través de la modificación e incorporación de clúster de comportamiento por posición en el modelo de optimización, generaría ganancias de aproximadamente $14.000.000 extras al modelo actual.

[2] Lift Analysis es una forma de medir como una campaña/cambio afecta a una métrica clave
[3] Datos proporcionados por la Célula de Relevancia PLP

Capítulo 3

Objetivos y alcance

3.1. Objetivo general

El objetivo general para el trabajo de título es: "Desarrollar una clusterización de las listas de productos presentes en Falabella.com, según el comportamiento de los elementos que la componen, para evaluar el impacto de la inclusión de estos conglomerados al modelo de optimización actualmente utilizado"

3.2. Objetivos específicos

Los objetivos específicos para el trabajo de título se dividen en:

1. Definir KPI's relevantes para la clusterización de PLP, en base a la interacción con los clientes e información transaccional de los productos
2. Determinar distribuciones de comportamiento según posición y clusterizar PLP en función de KPI's previamente definidos
3. Evaluar el impacto de incluir clusterizaciones en el modelo global de optimización

3.3. Alcances y resultados esperados

Se define el alcance del trabajo de título, como la realización de un clúster de PLP's según su comportamiento por posición, para las 4000 listas de productos activas en la página de Falabella.com, utilizando los KPI's transaccionales y de interacción con el usuario descritos en el anexo B1.

Luego, se busca modificar y entrenar el modelo global de optimización, con las categorías clusterizadas disponibles en el periodo agosto - septiembre y finalmente comparar resultados entre ambos modelos, con predicciones realizadas para el mes de octubre del año 2020.

Se espera generar conglomerados de PLP cuyos KPI's tengan un comportamiento a través de las posiciones que representen distribuciones definidas y similares, cumpliendo criterios del negocio, basados en que la distribución del porcentaje de ventas de cada conglomerado

no supere el 40 % para cada uno y criterios de ajuste matemático, basados en el coeficiente de la silueta mayor a 0,5.

Luego, al utilizar estos conglomerados como nuevo input de entrenamiento para el modelo global que optimiza contribución, y realizar predicciones para el mes de octubre 2020, se espera alcanzar un rendimiento en la métrica NDCG de contribución mayor a 0,58.

Capítulo 4

Marco conceptual

A continuación, se realiza un estudio del estado del arte en el área de Clustering, con énfasis en las metodologías que existen para generar conglomerados de datos, las métricas usadas en el proceso y los últimos descubrimientos de la academia. Además se presentan los fundamentos del modelo de optimización utilizado en la Célula de Relevancia PLP.

4.1. Clustering

Una definición rudimentaria de clustering puede considerarse "El proceso de organización de objetos en grupos, cuyos miembros son similares en algún aspecto". Esta idea pensada por el ser humano desde hace miles de años, se relaciona con la idea primitiva de asociar sombras en la pared, para luego ordenarlas según la similitud de sus características y posteriormente asociarles un nombre [6].

Junto a esto, los filósofos griegos en el sigo V a.c reflexionaban sobre la función cerebral de "agrupamiento" a través de preguntas sobre el conocimiento de la realidad, y la forma en como la percibimos, para luego describirla mediante características.

Luego, en la década 70 y 80, con la aparición y popularización del estudio del machine learning, se lleva a cabo una profundización estadística de este proceso de agrupamiento, generando distintos métodos que permitan distintos objetivos, y algoritmos para su realización automática. Dentro del Machine Learning, los procesos de agrupamiento o clustering quedan bajo la rama de aprendizajes no supervisados, es decir aprendizajes que buscan clasificar objetos sin ningún conocimiento a priori sobre el grupo al cual pertenecen. [7]

En la actualidad existen diferentes aplicaciones para esto tipos de agrupamiento, que van desde medicina (reconocimiento de patrones en imágenes para agrupar células cancerígenas), economía (evaluar grupos de riesgo en acciones de la bolsa), marketing (segmentación de clientes) hasta ciencias sociales agrupando perfiles psicológicos. Sin embargo, también existen desventajas en estas formas de agrupamientos, ya que el análisis de clúster es una metodología descriptiva, ateórica y no inferencial, es decir, no tiene bases estadísticas sobre las cuales deducir inferencias estadísticas para una población a partir de una muestra. Es un método basado en criterios geométricos y se utiliza fundamentalmente como una técnica exploratoria y descriptiva. [8]

4.1.1. Métodos de clustering

Dado lo anterior, al poseer un conjunto de datos con valores y comportamiento desconocidos, un algoritmo de clustering puede descubrir subgrupos presentes dentro de el, como se muestra en la figura 4.1:

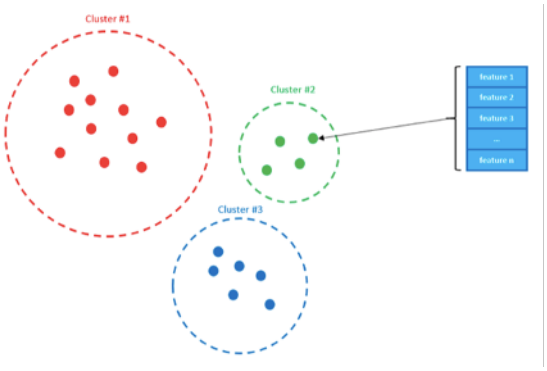


Figura 4.1: Ejemplo de un cluster genérico. Fuente: Practical Guide to Cluster in R [8]

Para esto, los métodos de clustering se dividen en 2 grandes grupos [9]:

- Algoritmos de Partición: Método de dividir el conjunto de observaciones en k clúster, en donde k lo define inicialmente el usuario. (Ver anexo C.1)
- Algoritmos jerárquicos: Método que entrega una jerarquía de las divisiones del conjunto de elementos, en clúster.

En ambos grupos de clúster, los pasos a seguir para realizar una correcta aplicación de los algoritmos es la siguiente: [9]

1. Elección de variables
2. Elección de la medida de asociación
3. Elección de técnica de clúster
4. Validación de los resultados

4.1.2. Algoritmos jerárquicos

Este grupo de métodos tiene su origen en el estudio de las taxonomías, basándose principalmente en la noción de aglomeración y división de conjuntos. Por ejemplo, un método jerárquico de división se inicia con un sólo grupo, que incluye todas las observaciones presentes en los datos, para luego en cada paso buscar el mejor corte que separa al grupo en dos subgrupos, y luego este proceso se repite para los subgrupos creados.

De manera opuesta, los métodos aglomerativos parten con un grupo en cada una de las observaciones de los datos, y luego en cada paso se van juntando pares de grupos que se consideren cercanos. Estos métodos son apropiados cuando se intuye o requiere una estructura jerárquica en la agrupación de los datos, ya que el resultado entrega la información completa

de la jerarquía resultante, la cual típicamente se visualiza mediante un dendrograma.

Un dendrograma se define como una representación gráfica o diagrama en forma de árbol, que organiza datos que se van dividiendo en subcategorías, hasta llegar al nivel de detalle deseado (similar a un árbol). Este tipo de diagrama permite apreciar claramente las relaciones de agrupación entre los datos.

Algunas aplicaciones típicas de este método son la categorización de especies animales, proteínas y ADN, como también la generación de índices temáticos de sitios web por categorías, e.g., DMOZ y Yahoo. [7]

4.1.3. Clustering jerárquico

Una de las desventajas de K-means, es el requisito previo de un número definido de K clúster a encontrar. Una alternativa para esto, es la utilización de un clúster jerárquico, que no requiere ningún número particular de K, y que además entrega los resultados en forma de árbol (dendrograma), sacando ventaja del algoritmo de partición K-means.

El algoritmo de un clúster jerárquico parte con la elección de una medida de distancia (usualmente la distancia euclidiana) y luego desde el fondo del dendrograma, cada una de las n observaciones es tratada bajo su propio clúster. Luego, los dos clúster más similares entre ellos (de acuerdo a la medida de distancia) se fusionan, por lo que ahora hay n-1 clúster. Luego, los dos clúster más similares se vuelven a fusionar, quedando ahora n-2 clúster. El algoritmo continúa hasta que todas las observaciones pertenecen a un solo clúster, y el dendrograma se encuentra terminado. [8] Ver figura 4.2

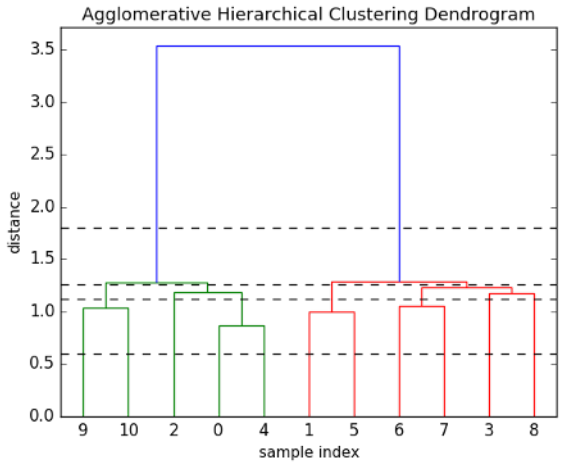


Figura 4.2: Dendrograma resultante luego de ejecutar un clustering jerárquico. Fuente: Practical Guide to Cluster Analysis in R. [8]

Para poder realizar las uniones mencionadas entre cada una de las observaciones, se requiere definir un método de unión de clúster, cuya finalidad es especificar como aplicar la medida de distancia entre los elementos de cada sub clúster.

Este método de unión, o también llamado método de linkage, presenta las siguientes po-

sibilidades:

Método complete: Se caracteriza por formar clústers cuya distancia entre clúster está definida utilizando la máxima distancia entre los miembros de un clúster y el siguiente.

$$D_{AB} = max(d(u_i, v_j)) \ \forall u \in A, v \in B \tag{4.1}$$

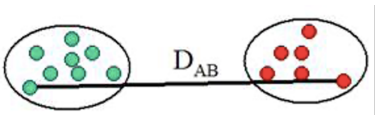


Figura 4.3: Linkage complete. Fuente: Elaboración propia

Método average: Se caracteriza por formar clúster cuya distancia entre conglomerados se define utilizando la distancia promedio de todos sus miembros.

$$D_{AB} = max(\frac{1}{(N_a N_b)} \sum \sum (d(u_i, v_j))) \ \forall u \in A, v \in B \tag{4.2}$$

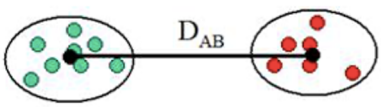


Figura 4.4: Linkage single. Fuente: Elaboración propia

Método single: se caracteriza por formar clúster cuya distancia entre clúster está definida utilizando la mínima distancia entre los miembros de un clúster y el siguiente.

$$D_{AB} = min(d(u_i, v_j)) \ \forall u \in A, v \in B \tag{4.3}$$

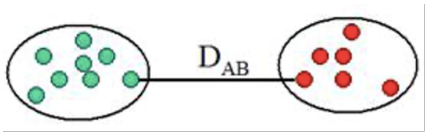


Figura 4.5: Linkage single. Fuente: Elaboración propia

4.2. Elección de medidas de distancia

La clasificación de observaciones en grupos requiere de métodos para calcular la distancia de similitud o disimilitud entre cada par de observaciones. La elección de una correcta medida de distancia es muy importante, ya que tiene una influencia directa en el tipo de algoritmo a utilizar y los resultados obtenidos. Las medidas de distancia se dividen en dos grupos, el primero basado en medir la distancia geométrica de dos observaciones (ver anexo C.2), y el segundo basado en su correlación. [8]

4.2.1. Medidas de distancia basadas en correlación

Las medidas de distancia basadas en correlación consideran que dos objetos son similares, si es que sus valores están altamente correlacionados, a pesar de que estos estén lejos en términos de distancia geométrica (la distancia entre 2 objetos es 0 cuando están perfectamente correlacionados). Si lo que se busca es identificar clúster de observaciones con un perfil similar, independiente de sus magnitudes, la medida de distancia basada en correlación es la correcta.[10]

Distancia de Correlación de Pearson: La distancia de Correlación de Pearson mide la relación lineal entre dos perfiles, (ver fórmula 4.4) y también es conocida como una relación paramétrica, ya que depende de la distribución de los datos. La distancia está basada en el coeficiente de Correlación de Pearson, tomando valores entre -1 (correlación total negativa) y 1 (correlación total positiva).[8]

$$d_corr : 1 - \frac{\sum(x_i - x)(y_i - y)}{\sqrt{\sum(x_i - x)^2 \sum(y_i - y)^2}} \tag{4.4}$$

Definición de la medida de correlación de Pearson. Fuente: An Introduction to Estatistical Learning [10]

Existen ciertas variaciones de esta distancia, dependiendo del valor del coeficiente de correlación. La Distancia Absoluta de Correlación de Pearson, es la medida de distancia usada cuando los valores del coeficiente de correlación se mueven entre 0 y 1. La Distancia de Correlación Centrada, ocurre cuando el coeficiente de correlación toma valores entre 0 y 2. En ambos casos el método de funcionamiento es el mismo a la distancia de referencia. [10]

Distancia de Correlación de Spearman & Kendall: La correlación de Spearman entre dos variables es igual a la correlación de Pearson entre el rango de valores de esas variables. La correlación de Pearson evalúa las relaciones lineales, mientras que la correlación de Spearman evalúa las relaciones monótonas (sean lineales o no). Intuitivamente la correlación de Spearman entre dos variables será alta, cuando las observaciones tengan un rango similar, y baja cuando las observaciones tengan un rango diferente.

La distancia de Kendall es una métrica que cuenta el número de diferencias entre dos listas de ranking, cuanto mayor sea la distancia, mas grande va a ser la diferencia entre las dos listas, la distancia de kendall también es llamada la distancia de bubble-sort, ya que es equivalente al número de swaps que el algoritmo de bubble-sort usaría para colocar una lista en el mismo orden de la otra lista. [8]

4.3. Aplicación de una medida de distancia basada en correlación

En particular, al utilizar un algoritmo del tipo jerárquico, la elección de una correcta medida de similaridad es muy importante, ya que presenta un fuerte efecto en el resultado de un dendrograma. Por ejemplo, si se considera un retail online, interesado en clusterizar compradores basándose en su historia pasada de compra, en este caso particular el objetivo sería identificar subgrupos de compradores similares, para que a los compradores de un mismo subgrupo se les pueda mostrar publicidad dirigida según sus intereses.

Asumiendo que los datos se pueden modelar en forma de matriz, donde las filas son los compradores, y las columnas son los elementos disponibles para ser comprados, los elementos de la matriz corresponden a las veces que los compradores compraron el elemento correspondiente (0 si nunca compró, 1 si lo compró una vez, 2 si lo compró 2 veces, etc..). Si se escoge una medida de distancia euclidiana, entonces los compradores que hayan comprado pocos elementos (es decir compradores infrecuentes del ecommerce) serán clusterizados en un mismo grupo.

Por otro lado, si se usa una medida de distancia basada en correlación, los compradores que tengan preferencias similares (es decir los compradores que hayan comprado el ítem A y B pero no el C y D) serán clusterizados en el mismo grupo, independiente si estos compradores compraron mayor cantidad de elementos que el resto, dentro del mismo grupo.

Para ilustrar gráficamente lo anterior, en la figura 4.6 se muestran 3 observaciones con mediciones para 20 variables cada una. Las observaciones 1 y 3 tienen similares valores para cada variable, por lo tanto, presentan una pequeña distancia euclidiana entre ellas, pero están débilmente correlacionadas, por lo que tienen una gran distancia basada en correlación. Por otro lado, las observaciones 1 y 2 tienen una gran diferencia de valores para cada variable, por lo tanto, tienen una gran distancia euclidiana entre ellas, pero están altamente correlacionadas, por lo que tienen una pequeña distancia basada en correlación.

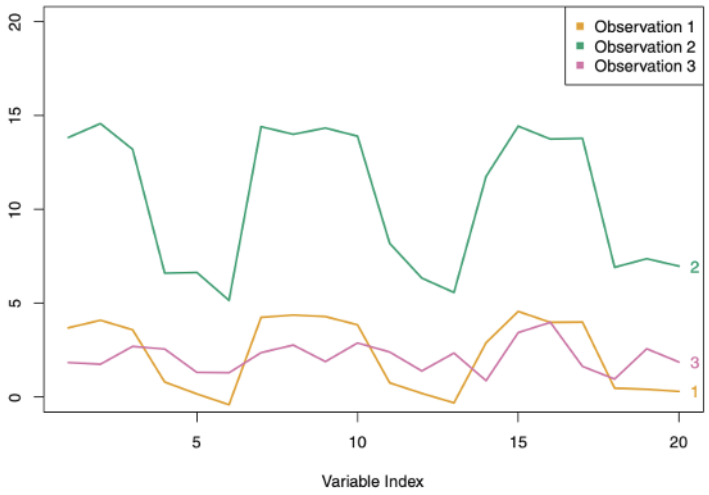


Figura 4.6: Clusterización de distribuciones, según distancia de correlación.
Fuente: An Introduction to Satistical Learning [10]

4.4. Medidas de evaluación de clúster

Para evaluar el rendimiento de un algoritmo no supervisado, en específico de un clúster, la dificultad radica en poder determinar si lo que considera "cerca" o "lejos" el algoritmo desarrollado, es realmente lo que se busca determinar. En específico, se debe buscar una forma de medir cuantitativamente si los elementos ubicados dentro de un clúster pertenecen o pertenecen a dicho conjunto, o más general aún, si dicho clúster tiene sentido.

La literatura habla de dos formas diferentes para medir lo anterior, la primera es utilizando datos etiquetados que no se han usado para la clusterización, o también llamada evaluación externa. En caso de no poseer estos datos, se plantea usar el mismo modelo para evaluar (evaluación interna).[12]

4.4.1. Coeficiente de silueta

El coeficiente de la silueta es una métrica de medida interna que tiene como objetivo encontrar el número óptimo de clúster. Para obtener el coeficiente, de ahora en adelante S(i), se debe tener como requisito dos componentes:

1. Los grupos obtenidos luego de implementar un algoritmo de clustering
2. La distancia entre los objetos de cada grupo

Luego se toma como referencia cualquier objeto i del conjunto de datos, y se calculan los siguientes valores: [13]

- a(i): Distancia media entre el objeto y todos los otros objetos de la misma clase
- b(i): Distancia media entre el objeto y todos los objetos del clúster más cercano

Entonces el valor de S(i) puede ser obtenido como una combinación de los valores de a(i) y b(i):

$$S(i) = 1 - \frac{a(i)}{b(i)} \Leftrightarrow a(i) < b(i) \tag{4.5}$$

$$S(i) = 0 \Leftrightarrow a(i) = b(i) \tag{4.6}$$

$$S(i) = \frac{b(i)}{a(i)} - 1 \Leftrightarrow a(i) > b(i) \tag{4.7}$$

casos posibles para el coeficiente de silueta. Fuente:[13]

Los valores para S(i) fluctúan entre -1 y 1. Mientras más cercano a 1, quiere decir que existe una distancia inter-clúster pequeña y una distancia intra-clúster más grande, lo que permite afirmar que el algoritmo de aprendizaje no supervisado funciona de manera correcta.[13]

El coeficiente de la silueta se puede representar de manera gráfica, dibujando la silueta de cada clúster en orden descendiente, para facilitar su visualización. (ver figura 4.7) Un ancho

de silueta amplio, indica mayor cantidad de elementos clusterizados, mientras que la altura de la silueta indica valores más altos de S(i), y por lo tanto una mejor clusterización.

Este método de evaluación permite ver de forma gráfica y analítica cuales elementos se encuentran bien agrupados dentro de un clúster y cuales se encuentran de manera forzada. [17]

En la figura 4.7, al lado izquierdo, se observan 3 conglomerados de datos bien agrupados, donde el ancho de cada grupo (color diferente) es similar, lo que indica una cantidad de elementos bien distribuida. Además se observa un alto promedio marcado por la línea punteada que indica el coeficiente de silueta para la totalidad de datos. Si existiese un conjunto de datos de tamaño muy pequeño en comparación al resto, o si la línea punteada se ubica cerca del 0, esto indicaría elementos agrupados de manera forzada.

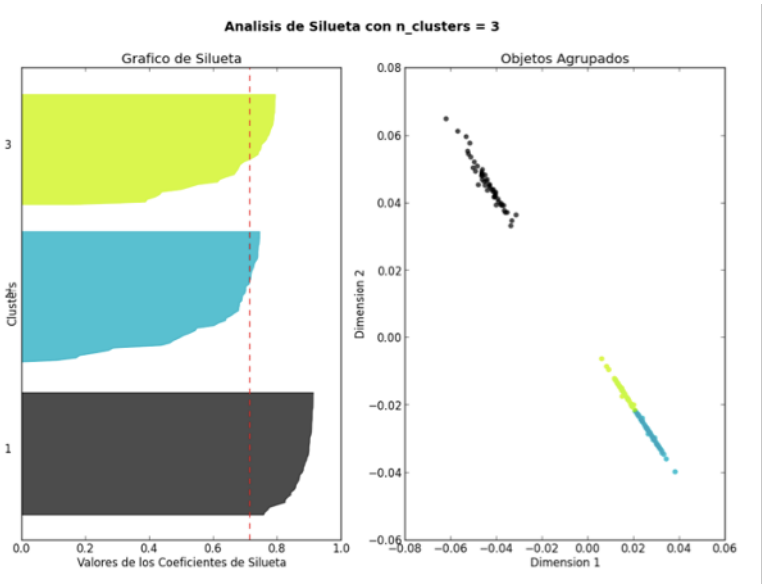


Figura 4.7: Diagrama de silueta de un dataset, para K-medioid con K=3.
Fuente: [13]

4.4.2. Índices de validación externos

La agrupación externa compara el algoritmo generado por el modelo de clustering, con otro que es considerado como "el mejor agrupamiento". Para poder realizar una validación externa, es necesario contar con estos datos de comparación, además de presentar una buena calidad de estos.

Este trabajo de título busca desarrollar un modelo de clustering que no cuenta con elementos etiquetados a priori, por lo que no es posible comparar con un grupo de control. Sin perjuicio de lo anterior, se dejan expuestas los nombres de las técnicas utilizadas para este tipo de validación, de acuerdo a la literatura: [13]

1. Adjusted Rand Index (ARI)
2. Normalized Mutual Information (NMI)

3. Adjusted Mutual Information (AMI)

4. Índices de validación relativos

4.5. Modelo global de optimización

El problema que pretende resolver el modelo global de optimización que opera actualmente en Falabella.com, es generar un ranking de productos para maximizar la contribución que deja cada uno de ellos a la compañía.

Para modelar lo anterior, el área utiliza una estrategia algorítmica de machine learning llamada Learning to Rank (LTR), la cual consiste en un problema de aprendizaje supervisado, donde la función objetivo del algoritmo, es una función de ranking compuesta por niveles de relevancia, la cual luego de su optimización, da origen a un modelo de ranking.

En específico, la estrategia utilizada asigna a cada nivel de relevancia una jerarquización distinta, es decir, si la función objetivo se divide en 5 niveles, el nivel 4 será caracterizado como *muy relevante* y el nivel 0 será caracterizado como *sin relevancia.*[15] Luego, el modelo utiliza un clasificador multiclase a través del algoritmo matemático XGBClassifier, para obtener la probabilidad de pertenencia de cada elemento a cada clase.

Para el caso particular de este problema se ocupa una estrategia LTR de tipo pointwise, ver figura 4.8, la cual consiste en asignar un score a los productos de forma individual, que luego permite ordenarlos en función del mismo.

Dicha distinción se basa en el desarrollo acerca de las aplicaciones del Learning to Rank en las búsquedas de ecommerce, presentado por Shubhra Kanti Karmaker Santu, Parikshit Sondhi, ChengXiang Zhai (2019) y Ping Li, Christopher J.C. Burges, Qiang Wu (2007)

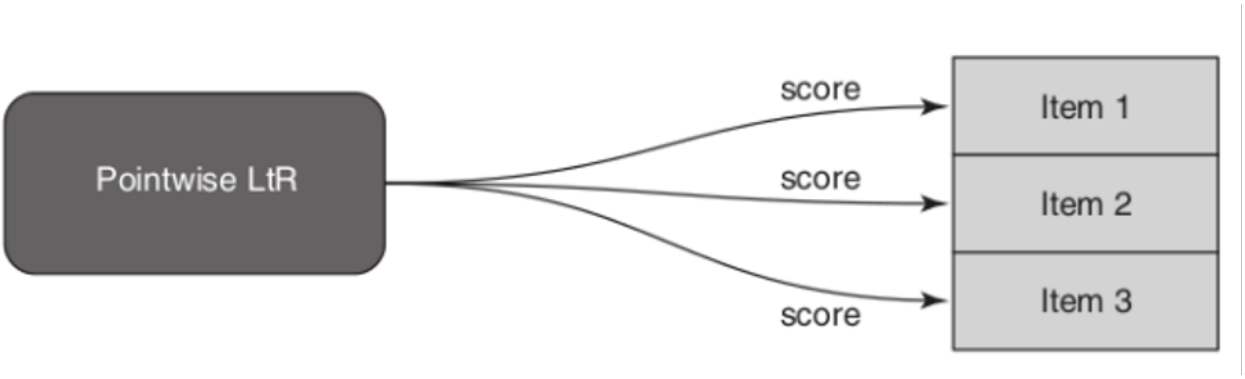


Figura 4.8: Estrategia PointWise. Fuente: [15]

Posterior a la clasificación, el score de cada producto es calculado sumando la probabilidad de pertenencia de cada clase multiplicada por su nivel de relevancia.[16]

4.5.1. Métricas de evaluación para modelo global de optimización

La evaluación de un sistema de recomendación o ranking es un problema fundamental y no trivial a la hora de reportar mejoras en modelos de machine learning. Esta necesidad surge en ámbitos de sistemas que generan recomendaciones a usuarios, listas de productos para ecommerce, sistemas de recuperación de información, entre otros, y cuyo objetivo es cuantificar que tan bien se encuentran ordenados o "rankeados" los elementos respecto a una métrica.

En caso de realizar una evaluación offline para sistemas de recomendación o ranking, la literatura provee 2 grupos de evaluaciones según las métricas que se deseen estudiar:

1. Métricas de evaluación no dependientes del ranking de elementos
2. Métricas dependientes del ranking de elementos

Para el primer grupo se encuentran las medidas de exactitud ("accuracy") que incluyen el MAE (mean absolue error) y RMSE (root mean square error), cuyo objetivo es indicar que tan "cerca o lejos" se encuentra una predicción versus sus valores esperado. Dentro de este grupo de métricas también se encuentran la precisión, recall y F1-score.

EL segundo grupo de métricas de evaluación, además de centrarse en evaluar elementos, también busca mejorar el ranking, para así mostrar elementos importantes, pero en el top de las recomendaciones. Para esto la literatura presenta 3 métricas:

1. MRR: Mean Reciprocal Rank
2. MAP: Mean Average Precision
3. NDCG: Normalized Discount Cumulative Gain

MRR es la métrica más simple de las tres (ver fórmula 4.8), cuyo objetivo es medir ¿Dónde está el primer ítem relevante?

$$MRR(O,U) = \frac{1}{U}\sum_{u}\frac{1}{k_u} \tag{4.8}$$

Fórmula Mean Reciprocal Rank. Fuente:[12]

Para cada usuario final U, genera una lista de recomendaciones, encuentra un ranking de la primera recomendación relevante, y por último calcula el recíproco de aquel ranking. Las ganancias de este método es que es simple de calcular e interpretar, pero tiene como deficiencia que no evalúa el resto de la lista recomendada, solo se centra en un único elemento.

MAP (Mean average precision) utiliza un promedio de la precisión (métrica descrita anteriormente) a través de sucesivas sub listas, con el objetivo de penalizar los errores al principio de la lista, con mayor grado que al final de esta. Tiene la ventaja de poder manejar ranking de listas de recomendaciones, pero solo en el caso de tener valoraciones binarias (relevante / no relevante).

La última y más destacada métrica, corresponde al NDCG, cuyo objetivo es medir documentos relevantes en el top del ranking generado. Es una métrica altamente recomendada

para evaluar listas de recomendación, ya que se basa en saber que existen documentos más "importantes" que otros, y estos deben ser mostrados al principio de la recomendación. La forma en que se calcula el NDCG para un ranking de elementos (ver anexo C3.1) es la siguiente:

Primero es calculado el Cumulative Gain, midiendo la relevancia de los ítem dentro de la recomendación, sin tomar en cuenta su posición dentro de esta. Como el objetivo es medir que tan importante son estos documentos dentro de un ranking, es necesario tomar en cuenta la posición donde se encuentran. Para esto se calcula el factor Discount Cumulative Gain, que corresponde al CG pero siendo penalizado por el logaritmo en base 2 de la posición aumentada en 1 (para encontrarse dentro del dominio de la función). Se utiliza la base logarítmica para obtener una curva suave.

Existe otra variante del factor DCG que corresponde a la misma métrica, pero dando mayor énfasis a la relevancia de los documentos a través del uso de una potencia en base 2 para su cuantificación.

Por último, para obtener que tan correcto es el orden entregado por el modelo estudiado, se calcula el DCG del ranking en un orden ideal, es decir ordenando sus relevancias de mayor a menor a través de sus posiciones. Finalmente se calcula métrica Normalized Discount Cumulative Gain, que corresponde a una proporción entre los resultados obtenidos por el modelo estudiado y el orden ideal, cuyo valores más cercanos a 1 indican un orden perfecto, y valores más cercanos a 0 solo ruido[1]. Un ejemplo práctico de lo anterior, corresponde a:

Suponiendo que se tiene una lista de 6 documentos

$$D_1, D_2, D_3, D_4, D_5, D_6 \tag{4.9}$$

Con una escala de relevancia que va de 0 a 3 para cada documento (0 es no relevante y 3 es muy relevante) donde el documento 1 tiene relevancia 3, el documento 2 tiene relevancia 2, etc.

$$3, 2, 3, 0, 1, 2 \tag{4.10}$$

El CG para este conjunto de documentos es:

$$CG_6 = \sum_{i=1}^{6} Rel_i = 3 + 2 + 3 + 0 + 1 + 2 = 11 \tag{4.11}$$

Luego, como se quiere enfatizar que los documentos importantes aparezcan al principio de la lista, se calcula el DCG, penalizando a través de una escala logarítmica

[1] NDCG al estar normalizado, presenta valores entre 0 y 1

Tabla 4.1: Construcción DCG. Fuente: Elaboración propia

i	**Rel_i**	**log_2(i+1)**	**(Rel_i)/log_2(i+1)**
1	3	1	3
2	2	1,585	1,262
3	3	2	1,5
4	0	2,322	0
5	1	2,585	0,387
6	2	2,807	0,712

$$DCG_6 = \sum_{i=1}^{6} \frac{Rel_i}{log_2(i+1)} = 3 + 1,262 + 1,5 + 0 + 0,387 + 0,712 = 6,861 \tag{4.12}$$

Por último, para calcular el NDCG, es necesario calcular el DCG del orden ideal, es decir, repetir el mismo proceso anterior, considerando los documentos ordenados según relevancia (3,3,2,2,1,0), obteniendo:

$$iDCG_6 = 7,141 \tag{4.13}$$

$$NDCG_6 = \frac{DCG_6}{iDCG_6} = \frac{6,861}{7,141} = 0,961 \tag{4.14}$$

Las ventajas de utlizar NDCG como métrica de evaluación están relacionadas con medir la importancia o relevancia de los atributos de una lista junto con el orden o ranking en que se muestran, penalizando por elementos poco importantes mostrados primero o elementos importantes mostrados en las últimas posiciones (ventaja comparativa respecto a la métrica MAP descrita anteriormente).

4.6. Cross Industry Standard Process For Data Mining (CRISP-DM)

Para cumplir con los objetivos planteados en este trabajo de título, se utilizará la metodología aplicada a proyectos de Minería de Datos CRISP-DM (Cross Industry Standard Process for Data Mining). Esta metodología surge a partir del término KDD, utilizado para extraer conocimiento de los datos, pero manteniendo un enfoque aplicado en el negocio. La metodología contempla como eje central el análisis de datos y presenta un estructura moldeable acorde a las necesidades del proyecto.

CRISP DM cuenta con las siguientes 6 etapas genéricas:

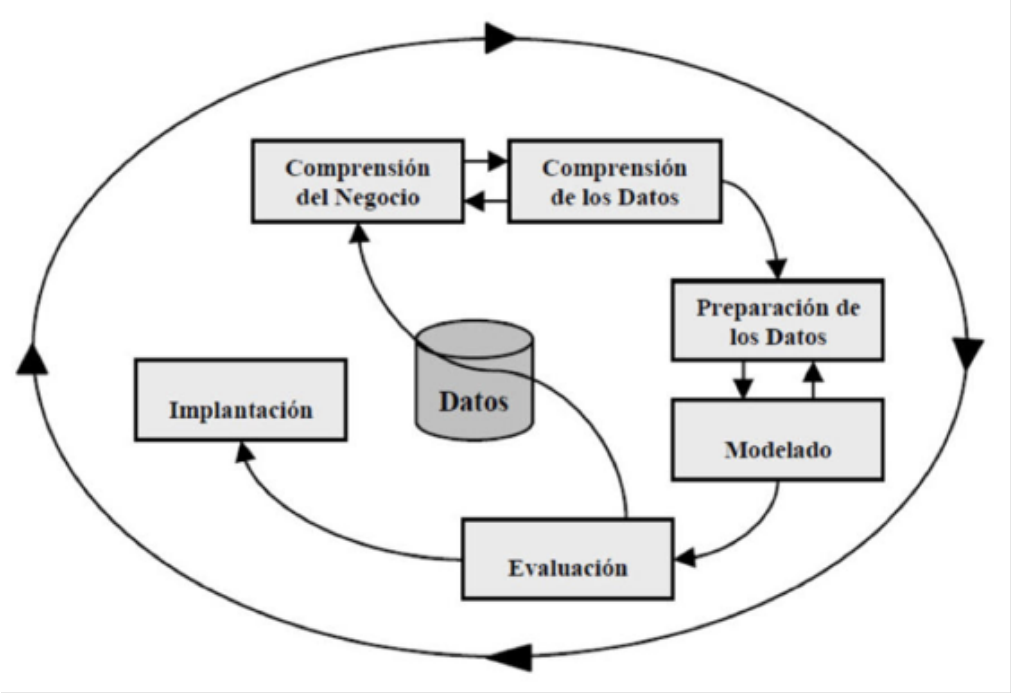


Figura 4.9: Metodología CRISP-DM para el desarrollo de proyecto de Data Minning. Fuente: [14]

1º Etapa, Comprensión del negocio: Fase inicial que se enfoca en la fijación y comprensión de los objetivos del proyecto, evaluación de la situación actual y generación del plan general de trabajo.

2º Etapa, Comprensión de los datos: Etapa que busca familiarizarse con los datos, a través de su recopilación inicial, descripción y exploración, buscando verificar calidad y usabilidad de ellos.

3º Etapa, Preparación de los datos: El objetivo es construir el conjunto final de datos, que posteriormente será utilizado en las herramientas de modelado, para ellos se deben seleccionar solo los relevantes, limpiarlos, integrarlos, formatearlos etc.. según corresponda.

4º Etapa, Modelado: En esta etapa se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema, abarcando el diseño, construcción y evaluación del modelo

5º Etapa, Evaluación: Se busca evaluar los resultados obtenidos en la etapa anterior, revisando si son útiles para las necesidades del negocio, revisando el proceso y estableciendo los siguientes pasos u acciones a seguir.

6º Etapa, Despliegue: Finalmente, se busca explorar la utilidad de los modelos, integrandolos a la toma de decisiones de la organización, dando fin al proyecto.

Capítulo 5

Metodología

5.1. CRISP-DM aplicado al caso de estudio

En base a las etapas de la metodología CRISP-DM, se propone a continuación su desarrollo, aplicado al contexto del trabajo de título.

Como se indica en la figura 4.9, esta metodología no es rígida, y se basa en la constante interacción entre los datos, la comprensión del negocio y el aporte de un juicio experto, en pos de lograr la creación de modelos con resultados satisfactorios. En la figura 5.1, se describen cada uno de estos pasos, acorde al objetivo planteado.

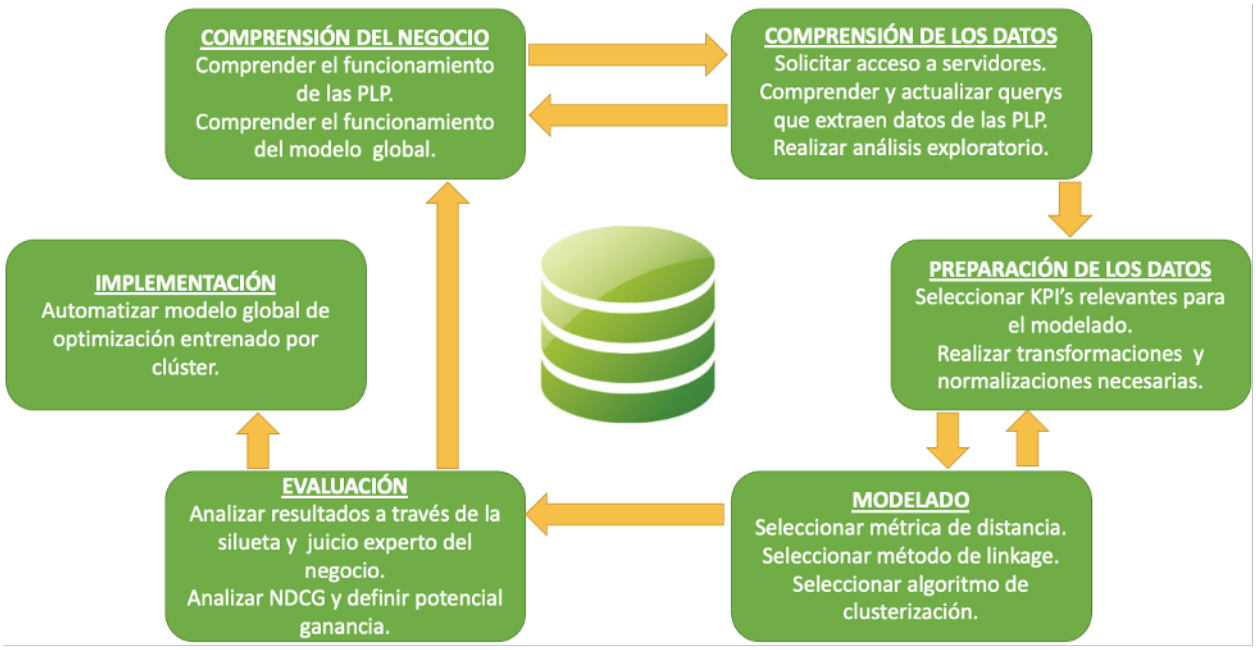


Figura 5.1: Metodología CRISP-DM aplicada en el trabajo de título. Fuente: Elaboración propia

A continuación se describe el detalle de cada uno de ellos.

5.1.1. Comprensión del negocio

Para la comprensión del negocio, se realizará un análisis de la pagina web de Falabella, buscando entender el funcionamiento de las PLP, su comportamiento actual en base a la interacción con el usuario, y además comprender la forma en que se aplica el modelo global para optimizar el orden de las listas de productos.

1. Comprender el funcionamiento de las PLP, desde el punto de vista del cliente
2. Analizar el cambio de orden dentro de los productos vistos en la página, para comprender el funcionamiento del modelo global de optimización
3. Para cumplir con esta etapa es fundamental la comunicación con el área de Data Science, quienes aportarán el conocimiento del negocio necesario para lograr un entendimiento completo del proyecto

5.1.2. Comprensión de los datos

Para esta etapa, se solicitará acceso a cada lista de productos de Falabella.com y las querys utilizadas para extraer los datos que generan, con el fin de obtener toda la información disponible asociada a cada PLP activa en la página, incluyendo interacción con el cliente, datos transaccionales y detalle de cada producto. Todo lo anterior a una de granularidad de posición y página. Posteriormente se consolidará y almacenará la información obtenida.

Con la información recopilada, se procederá a una comprensión de los datos, a través de un análisis exploratorio, utilizando herramientas adecuadas para su visualización.

1. Solicitar acceso a los servidores de Falabella, alojados en Google Cloud Plataform.
2. Se Solicitará acceso al repositorio Gitlab de la célula
3. Realizar una comprensión de la arquitectura bajo la cual se encuentran alojados los datos
4. Realizar modificaciones a las querys, para obtener la totalidad de información con la granularidad deseada
5. Generar consultas a la base de datos, extrayendo y almacenando la información obtenida
6. Por último, realizar un análisis exploratorio de los datos, utilizando la herramienta Tableau Desktop y la licencia otorgada por la Universidad de Chile

5.1.3. Preparación de los datos

Con la información necesaria ya recolectada y analizada, se prepararán los datos para ser posteriormente utilizados en la aplicación de modelos de machine learning no supervisados. Esta preparación se dividirá en dos etapas, siendo primero necesario filtrar la información disponible para las listas de productos, seleccionando solo los datos que aportarán valor al desarrollo del modelo final, y luego, modificar, adecuar y generar la estructura necesaria de estos, para que sea interpretable por la clusterización.

1. Seleccionar KPI's que aporten valor al desarrollo del modelo final
2. Realizar transformaciones a los datos, para que sean interpretables por la clusterización

5.1.4. Modelado

Para determinar grupos de PLP con comportamiento similar se implementará un modelo de machine learning no supervisado, de las listas de productos disponibles en el sitio. Primero se definirá, dentro de los modelos disponibles, cual es el que se ajusta mejor a los datos recolectados y los resultados esperados, luego se definirán los parámetros que el modelo requiera, junto con la medida de similitud que se utilizará.

1. Seleccionar el modelo de machine learning no supervisado más adecuado
2. Definir la medida de distancia o similitud que se utilizará
3. Definir los parámetros que el modelo requiera para la unión de conglomerados
4. Por último, se procederá a ejecutar el modelo

5.1.5. Evaluación

La evaluación se divide en dos ejes dependientes y consecutivos. Primero se evaluará el resultado de la clusterización, a través de la interpretación de cada conglomerado, analizando su pertinencia desde el punto de vista matemático y del negocio, a través de métricas de evaluación internas y análisis de KPI's respectivamente.

Luego, haciendo uso de las facultades otorgadas por la metodología utilizada, y según se muestra en las figuras 5.1 y 4.9 se volverá cíclicamente a realizar el proceso descrito en la figura 5.1 hasta determinar que los datos seleccionados, su transformación y posterior clusterización, dan origen a conglomerados válidos matemáticamente y relevantes para el negocio.

Posteriormente, con la clusterización realizada, se procederá a modificar y evaluar el modelo de optimización que se encuentra actualmente funcionando. Para lo anterior, se utilizará como base el modelo desarrollado en la Célula de Relevancia PLP, el cual se modificará para ser implementado, en forma paralela, en cada uno de los clúster ya definidos. Se considerará como set de entrenamiento las PLP de las categorías pertenecientes a cada conglomerado, con features relacionadas a información transaccional, información de interacción con el usuario, datos de los productos, y combinaciones matemáticas de lo anterior. Se utilizará como periodo de entrenamiento el intervalo agosto - septiembre y como periodo de predicciones, octubre.

Finalmente, para conocer el ajuste de cada conglomerado con el modelo, se evaluará el rendimiento matemático de este, a través las métricas de AUC, accuracy, recall y F1-score, además de un análisis de Feature Importance[1] en cada set de entrenamiento. Nuevamente se hará uso de las facultades que presenta la metodología CRISP-DM, para iterar sobre este análisis, hasta obtener el mejor ajuste posible.

Sin embargo, la evaluación final del rendimiento para cada uno de los clúster, cuyo resultado busca confirmar o rechazar la hipótesis planteada al inicio de este trabajo, se realizará a través de la comparación entre el modelo actual y los modelos para cada clúster, utilizando la métrica de NDCG. Este análisis offline, permitirá definir si el modelo separado, entrenado e implementado por clúster, generaría mayores ganancias para la compañía en términos de conversión, contribución y venta.

5.1.6. Implementación

Finalmente, según la metodología CRISP-DM, salir del ciclo descrito en la figura 5.1 se debe a que se obtuvieron los mejores resultados posibles para el trabajo propuesto.

En esta etapa, se implementará de manera online el modelo de optimización separado por clúster, si y solo si, los resultados de la evaluación a través del NDCG muestran una clara ventaja competitiva de esta aproximación versus el modelo actual. La implementación se realizará a través de la automatización del modelo, entrenando con los dos últimos meses de datos disponibles, y realizando predicciones de forma diaria, para obtener nuevos órdenes de productos cada día.

La automatización del modelo será realizada a través de la plataforma Google Cloud Plataform, que posee herramientas para orquestar procesos, llamadas DAG.

[1] Feature Importance se utiliza para describir cuales son los atributos más importantes en la predicción de las clases de un modelo de machine learning

Capítulo 6

Desarrollo metodológico

Acorde a la metodología planteada en el capitulo 4, se presenta a continuación el desarrollo realizado para cada uno de los pasos asociado al CRISP-DM, aplicado al caso de estudio.

6.1. Obtención y comprensión de los datos

Luego de interactuar y comprender el funcionamiento de las PLP, desde el punto de vista del cliente (ver figura (2.1)) se solicitó acceso a los servidores de Falabella.com, para recopilar los datos generados como resultado de la interacción entre las listas de productos y las personas. Estos datos se encontraban alojados en los servidores de almacenamiento y procesamiento de Google (Google Cloud Plataform).

Se tuvo acceso a la base de datos originada a partir del funnel de compra desarrollado e implementado para listas de productos (ver anexo D.1). Esta base, compuesta por 7.5TB de datos, contenía información de interacción entre las PLP y los clientes, información transaccional, información proveniente del funnel de compra e información de cada producto, para las 4000 PLP activas en la página. Cada fila de esta base de datos, correspondía al registro de un evento[1] a un nivel de granularidad de fecha, categoría, número de página, posición dentro de la PLP e id del producto.

Posteriormente, para poder trabajar con estos datos, se solicitó acceso al repositorio Gitlab de la célula, donde se encontraban almacenadas las querys que dan origen a esta y otras bases. En el repositorio, se seleccionó y modificó la consulta asociada a esta base de datos, para crear una tabla con información adecuada para la posterior clusterización.

La tabla resultante, contenía en sus columnas información asociada a:

- **Categoría:** Campo que describe la categoría a la que pertenece cada producto
- **Fecha:** Fecha en la que se registra el evento
- **Página:** Asociado al número de página en el cual se genera el evento
- **Posición:** Describe la posición dentro de la PLP para el evento registrado

[1] Evento se define como la interacción entre un producto y un cliente

- **Interacción con el usuario:** Impresiones de las PLP en pantalla, asociadas al número de click que se realiza en cada producto, cantidad de visitas a cada producto dentro la PLP y cantidad de visitantes (distintos) en cada producto dentro de la PLP
- **Información transaccional:** Cantidad de unidades vendidas, para cada producto dentro de la PLP y cantidad de dinero recaudado por la venta de cada producto dentro de la PLP
- **Flags:** Indicadores de tipo 0 o 1 que se utilizan para especificar el tipo de evento registrado

Dado que el objetivo es realizar una clusterización de comportamiento similar para las PLP, a través de sus posiciones, para posteriormente modificar el modelo de optimización que ordena los productos dentro de ellas, se dejaron fijos los flags:

- Recomendados = 1
- Call Center = 0

Debido a que la Célula de Relevancia PLP solo optimiza el orden de productos que se muestran por defecto o *recomendado* en la página, dejando fuera el orden alfabético de los productos, el orden de mayor a menor precio o cualquier otro orden distinto al recomendado, que un cliente ingrese como filtro en las categorías. Además, solo se optimizan productos que no son intervenidos por call center[2].

Finalmente, se obtuvo una base de datos compuesta por 10 columnas, con información para las mas de 4000PLP activas en el sitio, a un nivel de granularidad de fecha, categoría, página y posición, con datos para el mes de agosto del año 2020. Un extracto ilustrativo de la base se puede ver en la tabla 6.1

Tabla 6.1: Extracto de base de datos creada. Fuente: Elaboración propia

Categoría	Fecha	Página	Posición	Vistas	Visitas	Visitantes	Unidades Vendidas	Ventas Totales	Flag Orden	Flag Call Center
CAT620161	1/08/2020	1	1	383431	323122	322244	434	$234.324.232	Recomendado	0
CAT620161	1/08/2020	2	1	1234	1003	987	123	$53.234.123	Recomendado	0
CAT23	5/08/2020	1	1	34	21	20	2	$49.990	Recomendado	0
CAT23	5/08/2020	1	2	0	0	0	0	0	Recomendado	0

[2] Productos intervenidos por call center son aquellos productos que son comprados a través del call center, donde no se tiene interacción con la página de Falabella.com

6.1.1. Análisis exploratorio

La motivación de realizar una clusterización para el comportamiento por posición de las listas de productos de Falabella.com, nace a partir del análisis exploratorio de los datos que se presenta a continuación.

La base de datos creada, ilustrada en la tabla 6.1, y descrita en la sección anterior, está compuesta con información para 4077 categorías activas durante el mes de agosto del año 2020. Cuenta con 10 columnas, aproximadamente 4 millones de filas y no presenta valores faltantes o missing values en su composición.

Primero, se realizó un análisis numérico de los datos, encontrando que:

- El número de páginas que posee cada PLP varía entre 1 y 50
- El 70 % de las PLP poseen 10 o menos páginas
- Las PLP contienen entre 1 y 48 elementos, es decir, contienen 48 posiciones

Luego, se analizó cuales eran las categorías que daban origen a las PLP con mayor volumen de ventas y unidades vendidas, encontrando que:

- Las categorías con mayor volumen de ventas acumuladas en el periodo estudiado, fueron las categorias Notebook, con 3460 millones, Smartphone con 2181 millones y Televisores con 1842 millones
- Las categorías con mayor volumen de unidades vendidas acumuladas para el periodo estudiado, fueron Moda Mujer con 15289 unidades, Moda Hombre con 8755 unidades y Smartphone con 8166 unudades.

Sin embargo, el análisis que motivó la realización de la clusterización, se basó en estudiar como se comportan los KPI's de las PLP a través de sus posiciones, al ser separados en primera instancia por categoría y luego por número de página. Para realizar este análisis se utilizó la herramienta de visualización Tableau Desktop gracias a la licencia otorgada por la Universidad de Chile.

6.1.2. Análisis de PLP a través de sus posiciones, por categoría

A través de la herramienta de visualización de datos Tableau, se graficó el comportamiento del total de PLP, para los KPI's Unidades Vendidas, Venta Total e Impresiones PLP (o vistas), separadas por categoría, con el fin de distinguir si existe algún patrón de comportamiento similar entre ellas.

Los resultados obtenidos se muestran en las figuras 6.1, 6.2 y 6.3. Estos resultados permiten concluir dos análisis relevantes, el primero es que existe una correlación positiva entre las primeras posiciones de una PLP y un mayor número de unidades vendidas, y también el mismo tipo de correlación entre las primeras posiciones y la venta total. Además se observa que las impresiones PLP o vistas, asociadas a los click en cada una de ellas, se mantienen constantes, variando en magnitud acorde a cada categoría.

En segundo lugar, se concluye que el comportamiento por posición para los KPI relacionados con cantidad de unidades vendidas y venta total **no presenta un comportamiento homogéneo para todas las categorías**, es más, para las figuras 6.1 y 6.2 es imposible agrupar de manera manual categorías que presenten un comportamiento similar a lo largo del eje x, es decir a través de sus posiciones.

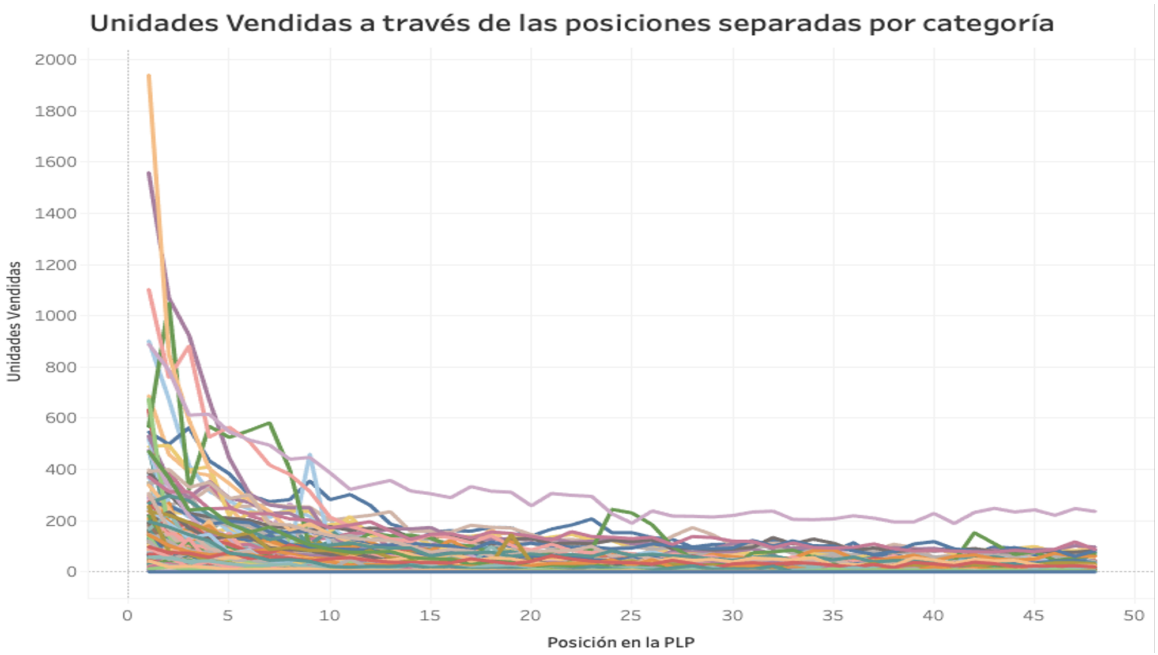


Figura 6.1: Unidades vendidas a través de las posiciones separadas por categoría (cada categoría es un color diferente). Fuente: Elaboración propia

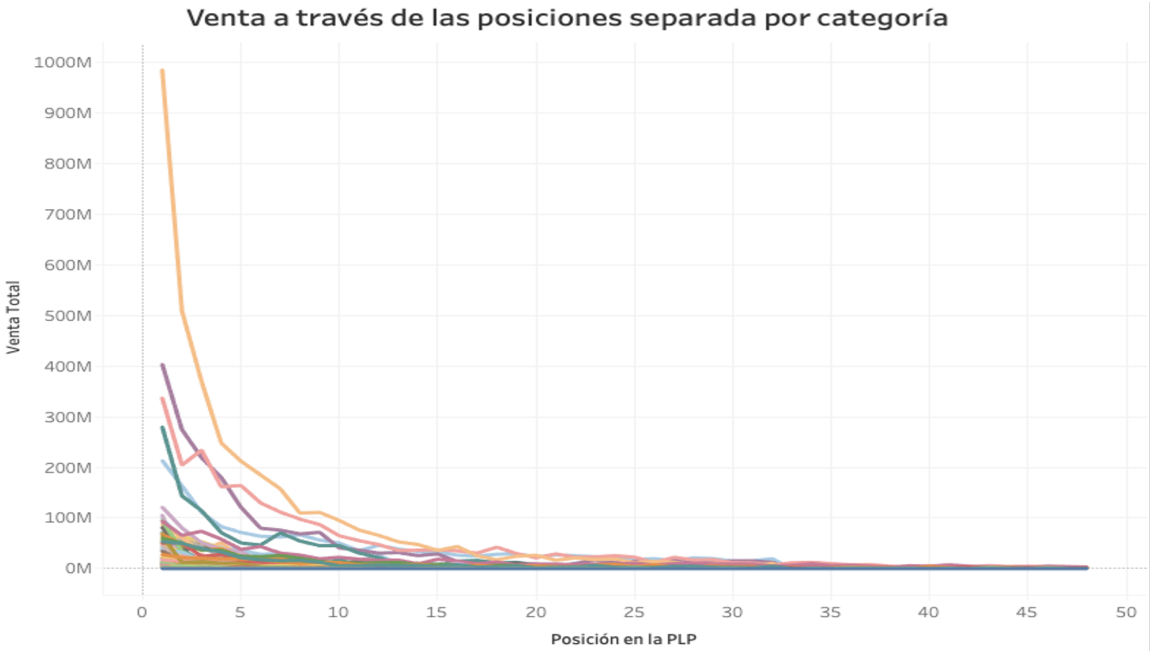


Figura 6.2: venta total a través de las posiciones separadas por categoría (cada categoría es un color diferente). Fuente: Elaboración propia

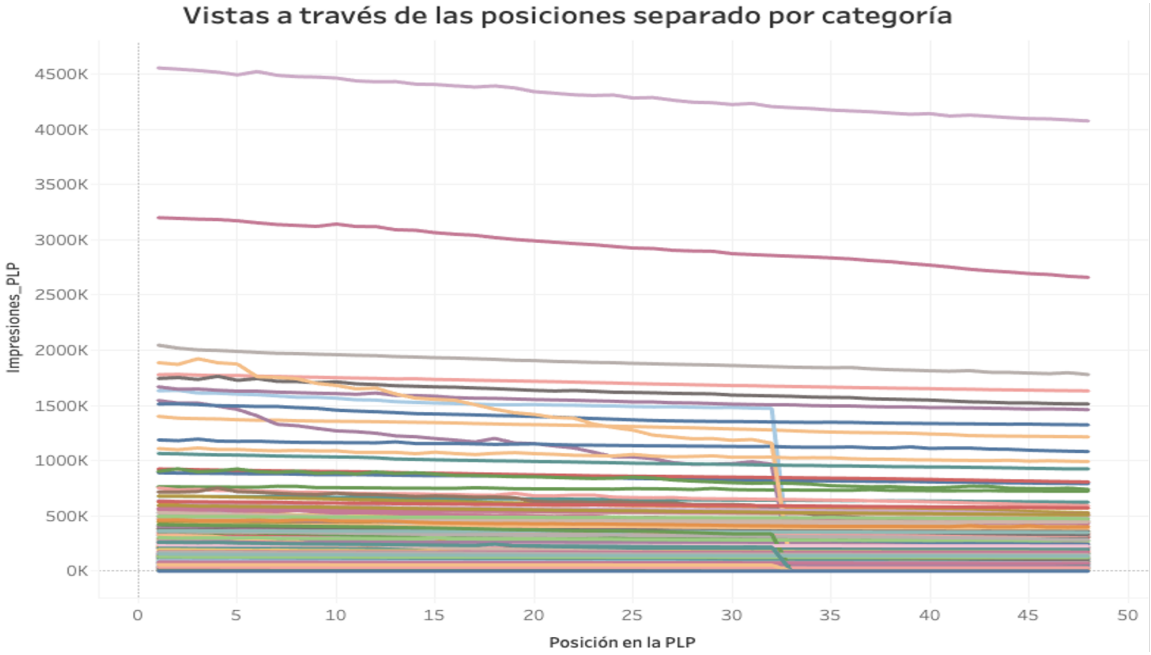


Figura 6.3: Vistas a través de las posiciones separadas por categoría (cada categoría es un color diferente). Fuente: Elaboración propia

Por último, para ilustrar de mejor manera el comportamiento de las PLP, en la figura 6.4 se graficaron las 3 categorías con más venta generada para el periodo estudiado. En naranjo se encuentra Nootebooks con 3460 millones, en rojo Smartphone con 2181 millones y en morado Televisores con 1842 millones

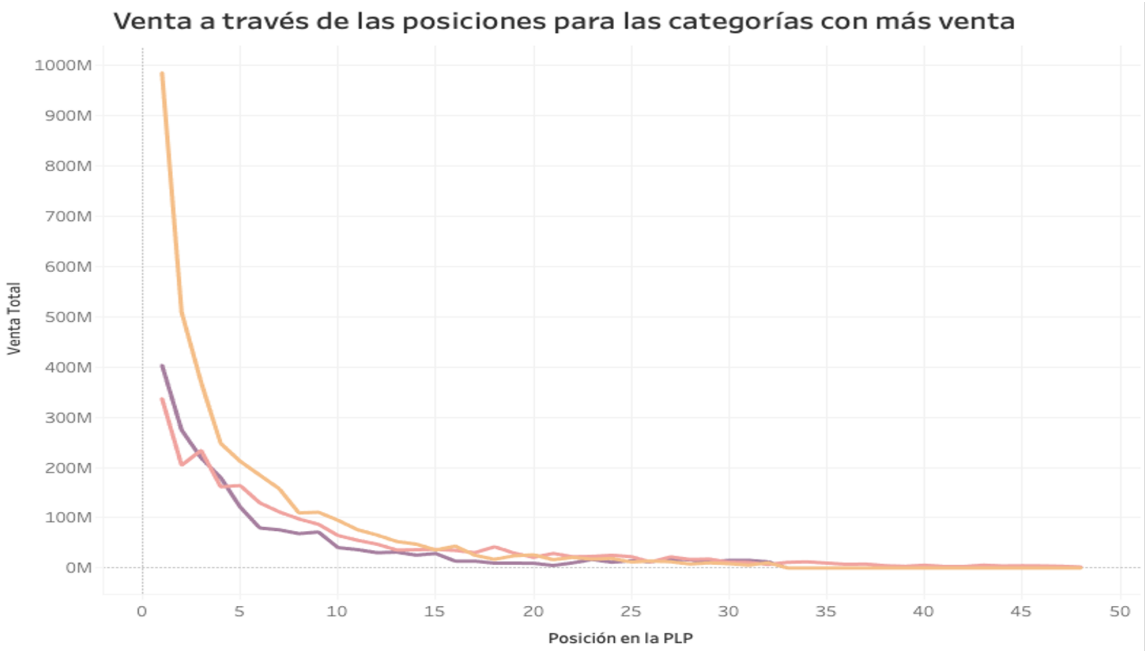


Figura 6.4: Comportamiento de las 3 categorías con mayor venta. Fuente: Elaboración propia

6.1.3. Análisis de PLP a través de sus posiciones, por número de página

Luego, se realizó un análisis exploratorio separando las PLP por número de página, es decir, en lo sucesivo cada línea graficada representa una página distinta, en el eje x se mantienen las posiciones dentro la PLP y en el y se grafica cada KPI respectivamente. Los resultados obtenidos se muestran en la figuras 6.5, 6.6 y 6.7.[3]

De los gráficos se obtuvieron 3 principales conclusiones:

1. El 95 % de las ventas y unidades vendidas se concentra en la primera página de elementos
2. La visualización en pantalla de la página número uno supera 6 veces al resto de las páginas
3. Al diferenciar por página el comportamiento por posición, no se distingue una distribución favorable o atractiva para una clusterización, como en el caso de las categorías

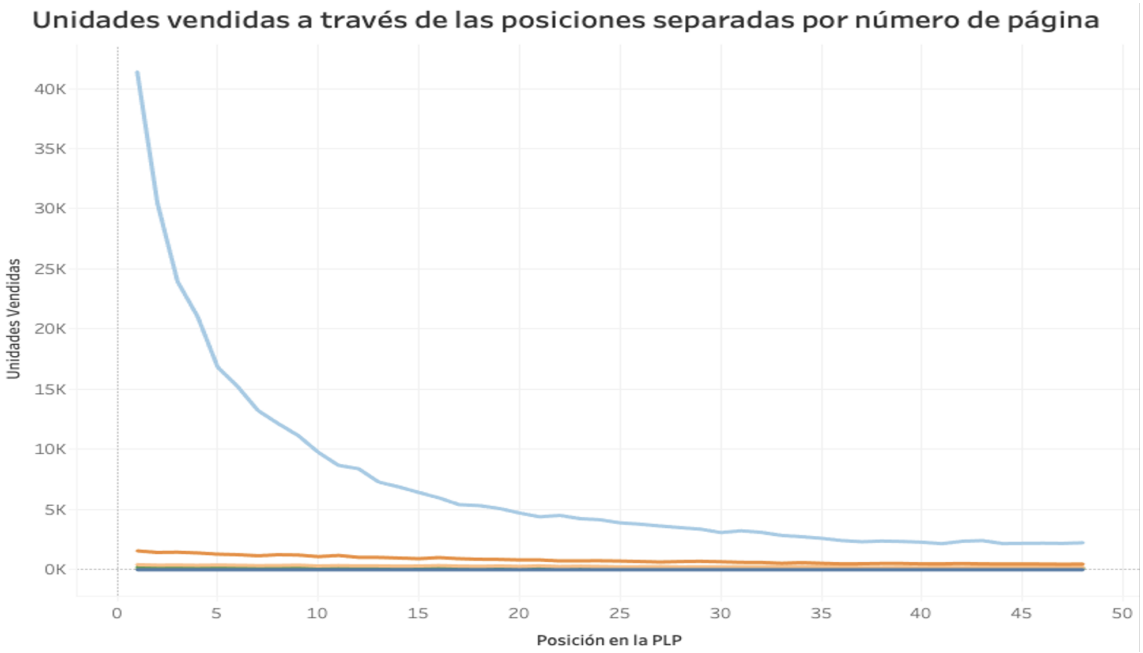


Figura 6.5: Unidades vendidas a través de las posiciones separadas por página (cada página es un color diferente). Fuente: Elaboración propia

[3] La página 1 está representada por el color celeste y la página 2 está representada por el color naranjo.

Figura 6.6: venta total a través de las posiciones separadas por página (cada página es un color diferente). Fuente: Elaboración propia

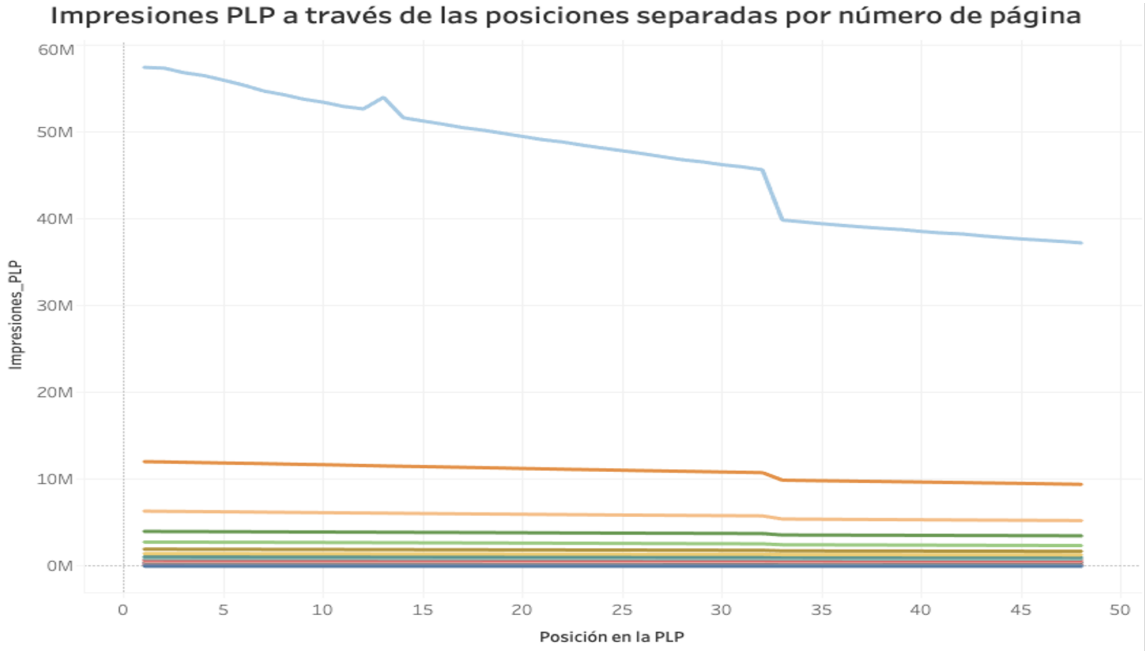


Figura 6.7: Vistas a través de las posiciones separadas por página (cada página es un color diferente). Fuente: Elaboración propia

Posterior al análisis exploratorio, queda comprobada la motivación para realizar una clusterización de comportamiento por posición, utilizando una separación por categoría, para entregar de forma explicita este input al modelo de optimización.

6.2. Preparación de los datos

Con los datos ya almacenados y estudiados, se procedió a prepararlos para poder utilizar un algoritmo de machine learning no supervisado. La preparación de los datos se dividió en dos etapas, primero se modificó la tabla previamente creada (ver tabla 6.1), para crear una segunda tabla, estructurada solo en base al análisis del comportamiento por posición, separado por categoría. Luego, sobre esa tabla se trabajó para adaptar los datos a un formato que sea interpretable por un algoritmo de clusterización.

6.2.1. Reestructuración de los datos

Antes de comenzar con la reestructuración, se definió cuales KPI's iban a ser utilizados en la clusterización. Para esto, dado que se utilizará como medida de similitud la Correlación de Pearson, tener KPI's altamente correlacionados, no aportará información al modelo. Se procedió a calcular la matriz de correlación para cada uno de los KPI's presentes en la tabla 6.1, cuyos resultados se muestran en la tabla 6.2, a continuación:

Tabla 6.2: Correlaciones calculadas para cada KPI. Fuente: Elaboración propia

	Impresiones	**Visitas**	**Visitantes**	**Unidades Vendidas**	**Venta Total**
Impresiones	1.000	0.998	0.998	0.633	0.257
Visitas	0.998	1.000	0.999	0.621	0.253
Visitantes	0.998	0.993	1.000	0.620	0.251
Unidades Vendidas	0.633	0.621	0.620	1.000	0.612
Venta Total	0.257	0.253	0.251	0.612	1.000

Con las correlaciones calculadas, se procedió a eliminar las columnas de visitas y visitantes, manteniendo solo las columnas de Impresiones o vistas, asociada a los clicks en cada PLP y las veces que esta se muestra en pantalla, junto con las variables transaccionales, unidades vendidas y venta total.

Luego para la reestructuración de los datos, teniendo como base la estructura de la tabla 6.1, se procedió a:

1. Disminuir la granularidad a categoría, página y posición, es decir, la magnitud de cada KPI ahora representa la suma de todo el periodo estudiado, eliminando la columna fecha

2. Volver a disminuir la granularidad, a categoría y posición, donde ahora la magnitud de cada KPI, representa la suma de todas las páginas para todo el periodo estudiado, eliminando también la columna número de página

Para ilustrar de mejor forma el proceso realizado, en la tabla 6.3 se muestra el resultado parcial de transformación hasta este punto:

Tabla 6.3: Etapa intermedia de transformación de los datos. Fuente: Elaboración propia

Categoría	**Posición**	**Vistas**	**Unidades Vendidas**	**Venta Total**
CAT720161	1	7442346	2563	$745.213.454
CAT720161	2	5653552	2105	$650.453.434
CAT720161	3	3540968	1200	$315.655.345
CAT983	1	3434	230	$15.434.577
CAT983	2	1459	60	$5.435.455
CAT983	3	4560	567	$40.456.234
CAT664	1	21	0	0

Luego de esto, para estructurar los datos en función las posiciones, separados por categoría, se utilizaron operaciones matriciales, transformando cada KPI en un feature (o columna) que alimente el modelo no supervisado implementado posteriormente.

Lo anterior, transformó la base ilustrada en la tabla 6.3, en la base final, ilustrada en la tabla 6.4

Tabla 6.4: Base de datos transformada. Fuente: Elaboración propia

Categoría	Vistas_1	Vistas_2	Vistas_3	Unidades_1	Unidades_2	Unidades_3	Ventas_1	Ventas_2	Venta_3
CAT720161	546745	454985	325985	123	90	76	$4.466.346	$3.367.346	$2.934.345
CAT345	23	0	0	2	0	0	$23.990	0	0
CAT9302	1235	4566	2045	230	600	340	$23.345.566	$69.346.345	$42.345.772
CAT023	0	0	0	0	0	0	0	0	0

La tabla 6.4 está compuesta por 4077 filas, donde cada fila representa una categoría que da origen a una PLP, y 145 columnas, donde cada columna es un KPI, o de ahora en adelante *feature*, medido para una posición en específico, que alimentará el modelo.

6.3. Modelado

Según lo expuesto en la sección 4.4.1, los pasos para realizar una clusterización son:

1. Elección de variables
2. Elección de medida de similitud
3. Elección de técnica de clúster
4. Validación de resultados

El desarrollo metodológico planteado hasta ahora, abarca el punto número 1.

Para elegir una correcta medida de similitud, según lo expuesto en la sección 4.2, es necesario considerar los resultados esperados para el proceso de clusterización. En este caso, dado que se busca clusterizar PLP que tengan un comportamiento similar a través de sus posiciones, independiente de sus magnitudes, se utilizó como medida de similitud una distancia basada en correlación. En específico se utilizó la distancia basada en Correlación de Pearson.

Para la elección de la técnica de clusterización, según lo expuesto en la sección 4.1.1, existen 2 métodos principales, los algoritmos de partición y los algoritmos jerárquicos. Puesto que se decidió utilizar la Correlación de Person como medida de similitud, se utilizó como técnica de clusterización a los algoritmos jerárquicos. En específico se utilizó un algoritmo jerárquico acumulativo, dada sus bondades de interpretabilidad a través de un dendrograma, su buen ajuste a la medida de similitud escogida y las facultades que otorga para clusterizar sin entregar a priori un número definido de conglomerados.

Antes de proceder a la implementación del algoritmo, fue necesario realizar una modificación extra a la base de datos planteada en la Tabla 6.4. Se procedió a normalizar los features, utilizando una librería de Python llamada StandardScaler, cuya finalidad es normalizar a través de la media y desviación estandar de los datos, para evitar resultados sesgados por la diferencia en magnitud de unidades.

Si bien, ya están definidos los features que alimentarán el modelo, la técnica de clústerización y también la medida de similitud, dado que se utilizó una clusterización jerárquica acumulativa, es necesario definir que método de unión de clúster o linkage se utilizará en dicha clusterización. Según lo expuesto en la sección 4.1.3 esta elección depende de los datos utilizados en cada caso.

Dado esto, a continuación se exponen los resultados de la implementación del algoritmo con los parametros ya definidos, para cada tipo de linkage.

6.4. Resultados del modelado

Como se expone respectivamente en la sección 4.1.3 y 4.5, los resultados de una clusterización jerárquica acumulativa se visualizan a través de un dendrograma y se evalúan a través del coeficiente de la silueta. Un dendrograma se interpreta considerando que en el eje x se encuentran todos los elementos que serán clusterizados, visualizados a través de líneas verticales. Luego, al unir dos líneas verticales, se forma un nuevo conglomerado de datos, que el algoritmo interpretó como cercanos, según la medida de similitud utilizada. El eje y representa la distancia que existe entre conglomerados, es decir, a mayor longitud de una línea vertical, mayor distancia entre un conglomerado y otro.

De forma paralela, el análisis de la silueta se basa en analizar el coeficiente resultante, buscando lograr un valor lo más cercano a 1, que indicaría un ajuste perfecto de los datos. El gráfico para este coeficiente se interpreta mirando el valor resultante en el eje y, que representa el valor del coeficiente según una cantidad determinada de clúster en el eje x.

A continuación se muestran los resultados obtenidos para la clusterización jerárquica acumulativa del comportamiento por posición de las listas de productos de Falabella.com, para los linkage Complete, Average, Single y Weighted.

6.4.1. Linkage Complete

El resultado de la clusterización se muestra en el dendrograma de la figura 6.8.

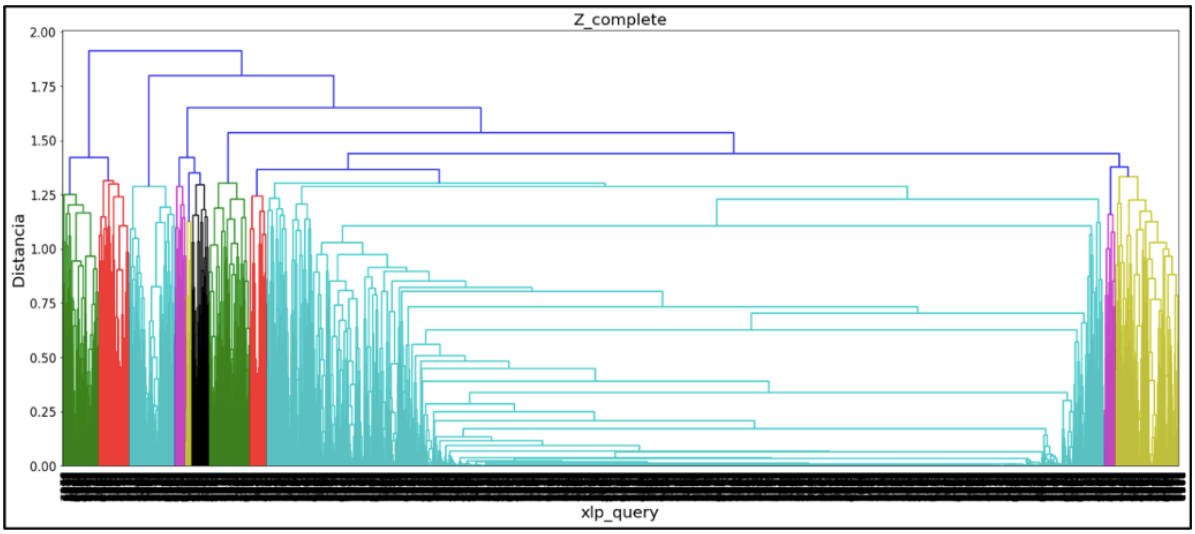


Figura 6.8: Dendrograma resultante para linkage Complete. Fuente: Elaboración propia

Este tipo de linkage, genera un dendrograma con una muy buena diferenciación entre grupos distintos, marcados con diferentes colores, ubicados a distancias en el eje y que permiten la realización de un corte limpio, a la hora de separar en una cantidad K de clúster. Luego, para ver el ajuste matemático de la clusterización se presenta la figura 6.9.

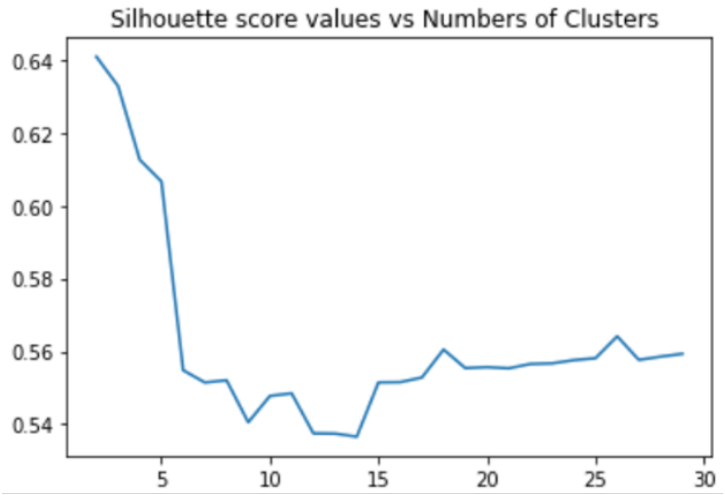


Figura 6.9: Análisis de la silueta para linkage Complete. Fuente: Elaboración propia

De la figura 6.9 se logra corroborar que el método complete no solo presenta una buena visualización a través del dendrograma, si no que muestra un muy buen ajuste para el coeficiente de silueta. En el gráfico se puede observar que en el eje y, el mínimo valor alcanzado se encuentra en torno al 0.54, considerando un rango de 2 a 30 clúster.

6.4.2. Linkage Average

El resultado de la clusterización se muestra en el dendrograma de la figura 6.10

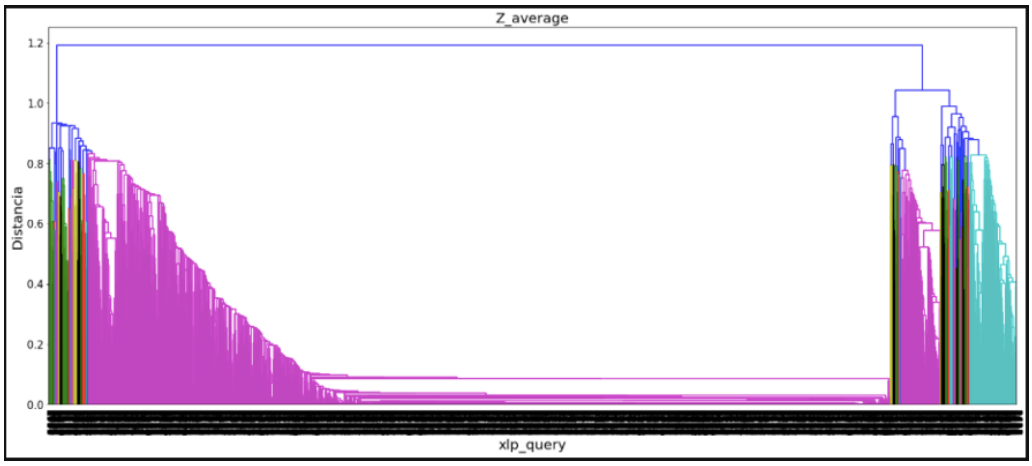


Figura 6.10: Dendrograma resultante para linkage Average. Fuente: Elaboración propia

Para el caso de la figura 6.10 se observa una diferenciación marcada entre diferentes conglomerados de datos (colores diferentes) pero a su vez estos conglomerados están repartidos de forma no uniforme a través del dendrograma, lo que hace difícil la realización de un corte limpio en k diferentes clúster. Para analizar matemáticamente el ajuste de esta clusterización,se calculó el coeficiente de silueta expuesto en la figura 6.11.

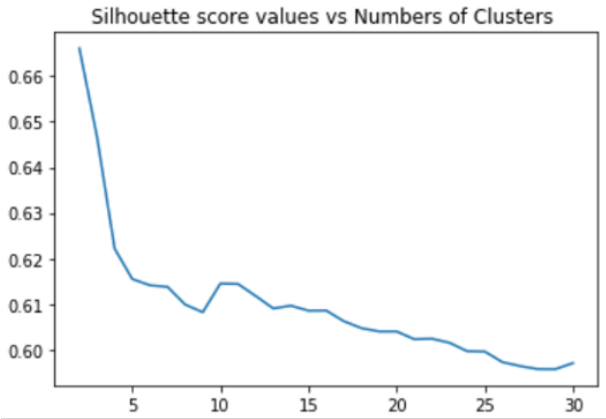


Figura 6.11: Análisis de silueta para linkage Average. Fuente: Elaboración propia

En este caso el análisis de silueta muestra una leve mejoría respecto al ajuste matemático del método anterior, alcanzando un mínimo de aproximadamente 0.58 para el rango propuesto de 2 a 30 clúster. También se observa que la medida es decreciente en la cantidad de clúster propuestos.

6.4.3. Linkage Single

El resultado de la clusterización se muestra en el dendrograma de la figura 6.12

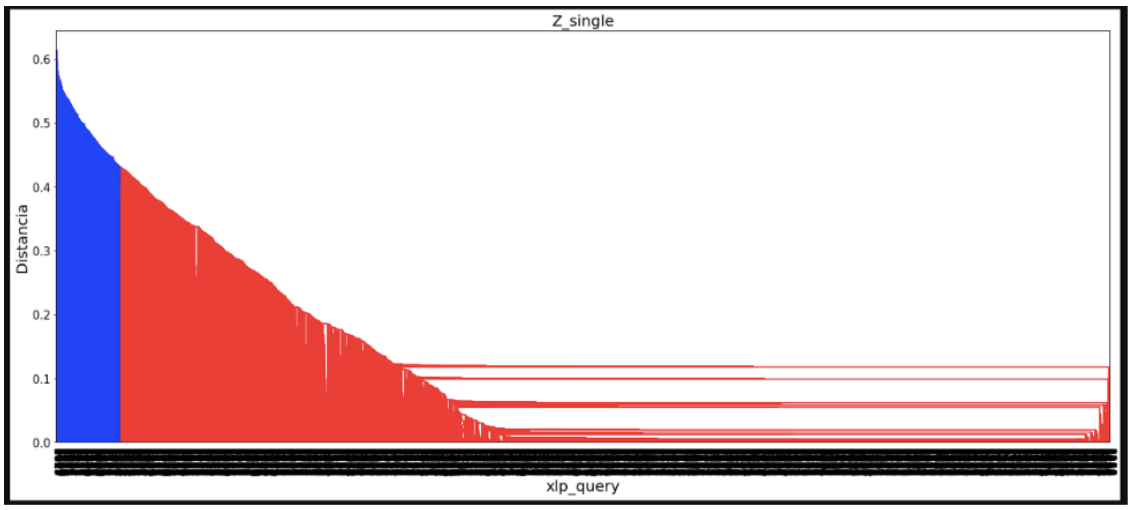


Figura 6.12: Dendrograma resultante para linkage Single. Fuente: Elaboración propia

En este método se observa una diferenciación marcada entre 2 conglomerados de datos (colores rojo y azul), que se encuentran abarcando todo el dendrograma. Este resultado carece de interpretabilidad, ya que no es posible realizar un corte horizontal que determine un numéro k de clúster de forma independiente. De todas formas el análisis para el coeficiente de silueta se presenta en la figura 6.13

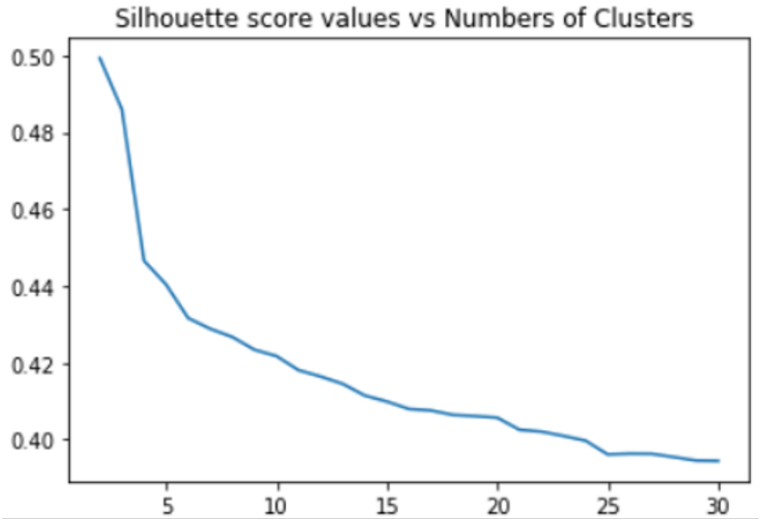


Figura 6.13: Análisis de silueta para linkage Single. Fuente: Elaboración propia

Como era de esperar, los resultados muestran un peor ajuste comparado con los métodos utilizados anteriormente, exhibiendo un mínimo de coeficiente de aproximadamente 0.38.

6.4.4. Linkage Weighted

El resultado de la clusterización se muestra en el dendrograma de la figura 6.14.

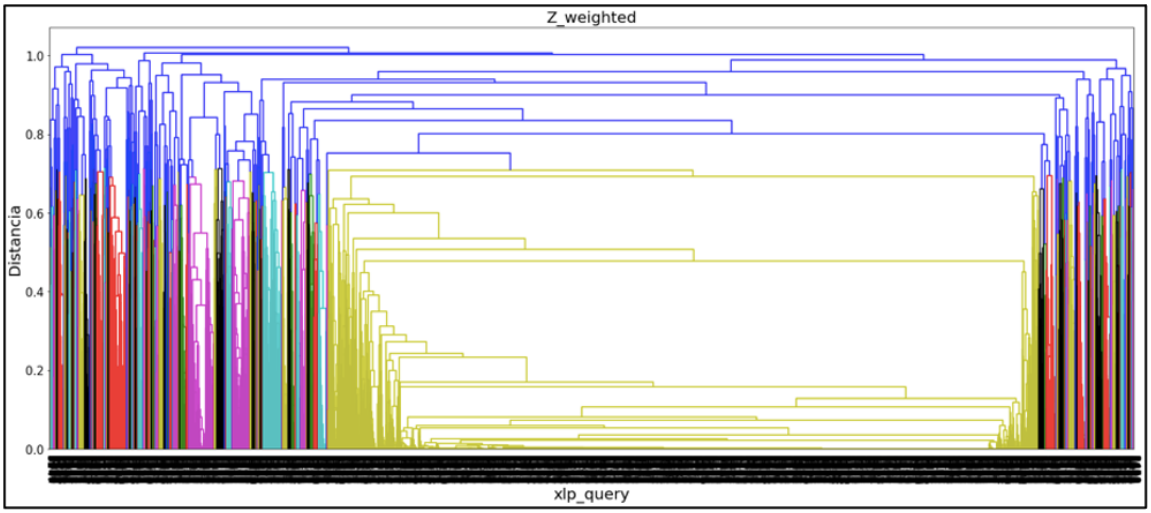


Figura 6.14: Dendrograma resultante para linkage Weighted. Fuente: Elaboración propia

Por último, el método Weighted permite distinguir diferentes conglomerados de datos (colores diferentes), pero a su vez estos conglomerados vuelven a estar repartidos de forma no uniforme a través del dendrograma, lo que hace difícil la realización de un corte limpio en k diferentes clúster. El análisis de la silueta se presenta en la figura 6.15.

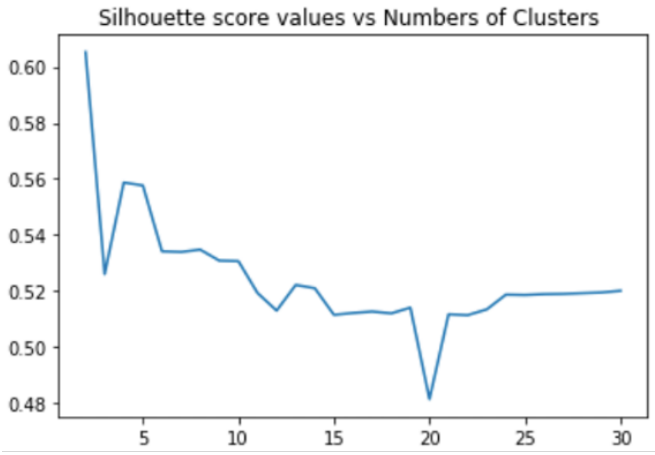


Figura 6.15: Análisis de silueta para linkage Weighted. Fuente: Elaboración propia

El análisis de silueta confirma el buen ajuste matemático de la clusterización, encontrando que el ajuste más bajo se encuentra en aproximadamente 0.48. Si bien es un buen ajuste, dada la dispersión de conglomerados presente en el dendrograma, el coeficiente de silueta es superado por otros linkage vistos anteriormente.

6.4.5. Elección de un número K de clúster

Para la elección de un número K de clúster a utilizar, se siguieron los siguientes pasos:

1. Seleccionar un tipo de linkage
2. Proponer valores candidatos de K clúster en base a juicio experto
3. Analizar valores propuestos a través del coeficiente de silueta

Para el primer paso, se realizó una comparación de los 4 tipos de linkage utilizados, considerando interpretabilidad de su dendrograma y los resultados obtenidos para el coeficiente de la silueta, al dividir la clusterización entre 2 y 30 candidatos.

La tabla 6.5 contiene las valores mínimos alcanzados por cada método, ordenados de mayor a menor.

Tabla 6.5: Coeficientes de la Silueta obtenidos para cada tipo de linkage. Fuente: Elaboración propia

Método Linkage	Coeficiente de Silueta
Average	0.58
Complete	0.54
Weighted	0.48
Single	0.4

De la tabla 6.5 se puede concluir que el método con mejor ajuste para los datos, es el método Average, pero dado que la clusterización es un método de machine learning no supervisado, que no presenta etiquetas en sus datos, la interpretabilidad de sus resultados juega un papel fundamental al momento de tomar decisiones.

Dado lo anterior, y lo mostrado en las figuras 6.8, 6.10, 6.12 y 6.14, se decidió utilizar el método de linkage Complete. El método Complete presenta una diferencia de solo 0.04 unidades respecto al coeficiente de silueta más alto, pero la interpretabilidad de su dendrograma es notablemente mayor.

Con la clusterización realizada y el método de linkage definido, según lo expuesto en la sección 4.1.3 una de las ventajas de la clusterización jerárquica acumulativa, es la definición a posteriori de la cantidad K de cluster a utilizar. En este caso, se propusieron como candidatos los k 5, 11, 15 y 18. (Ver anexo D.2.1)

El k final, se escogió en base a dos reglas aportadas por el juicio de un experto en el negocio, junto al posterior análisis del coeficiente de la silueta para cada candidato. Las reglas fueron las siguientes:

- No puede existir un conglomerado que acumule más del 40 % de la venta
- No puede existir más de un 33 % de conglomerados con menos del 1 % de la venta

En el anexo D.2.1 se encuentras las tablas D.1, D.2, D.3 y D.4 con la distribución del % de venta para cada candidato. Utilizando las reglas ya expuestas, y el análisis para la silueta mostrado en la figura 6.9, se concluyó que la cantidad de clúster k que representa de mejor forma a la clusterización jerárquica acumulativa realizada, equivale a 15 conglomerados.

Esta división cumple con las reglas del negocio y presenta un coeficiente de silueta de 0.55.

6.5. Evaluación de la clusterización

Con la clusterización realizada y el número k de clúster definido, se procedió a analizar el comportamiento de los conglomerados resultantes. Primero se estudió como se distribuyen las categorías en cada uno de los clúster, información disponible en la tabla 6.6.

Tabla 6.6: Distribución de categorías por clúster. Fuente: Elaboración propia

Clúster	**Nº de categorías**
1	134
2	32
3	81
4	167
5	38
6	22
7	60
8	65
9	88
10	61
11	88
12	2974
13	39
14	40
15	188

De la tabla 6.6 se concluye que a pesar de existir conglomerados que acumulan menos del 1 % de las ventas, no existen clúster sin categorías. También es importante notar que existe un conglomerado que acumula una gran cantidad de categorías, lo cual fue consultado y analizado con un experto del negocio, concluyendo que es un comportamiento totalmente esperable, ya que dicho clúster acumula las PLP originadas por categorías que presentan poca o nula interacción a través de sus posiciones, en algunos casos poca cantidad de productos o productos de temporadas anteriores, descontinuados etc.

Posteriormente se analizó como distribuyen los KPI de venta total y cantidad de unidades vendidas, ordenando sus porcentajes de mayor a menor participación en cada clúster, en las tablas 6.7 y 6.8 respectivamente. Para esto se consideró la sumatoria de la venta y unidades vendidas en cada posición de las PLP, para cada conglomerado. Se concluyó que existe una distribución similar de porcentaje para ambos KPI.

Tabla 6.7: Distribución de venta para K=15. Fuente: Elaboración propia

Clúster	% de Venta
9	36,75 %
13	21,16 %
1	9,48 %
8	7,67 %
12	6,44 %
3	5,68 %
4	5,40 %
5	1,81 %
11	1,74 %
15	1,16 %
7	0,96 %
10	0,93 %
2	0,46 %
6	0,27 %
14	0,07 %

Tabla 6.8: Distribución de unidades vendidas para K=15. Fuente: Elaboración propia

Clúster	% de Unidades Vendidas
1	41,30 %
3	17,34 %
4	11,08 %
9	9,75 %
13	6,33 %
8	3,43 %
5	2,70 %
12	2,66 %
2	1,43 %
7	1,33 %
11	0,87 %
15	0,83 %
10	0,52 %
6	0,46 %
14	0,26 %

Luego, se realizó un análisis a través de las posiciones para cada conglomerado, disponible en la figura 6.16. Este análisis consistió en realizar una sumatoria de todas las PLP pertenecientes a cluster, en función de un KPI en particular, para luego graficar todos los conglomerados a la vez, es decir, en el eje x se encuentran las posiciones dentro de la PLP, en el eje y se encuentran los KPI's respectivos, y cada línea representa un clúster diferente.

En la imagen se puede corroborar lo expuesto anteriormente en las tablas 6.7 y 6.8, mostrando que los conglomerados 9 y 1 acaparan los mayores porcentajes de venta total y unidades vendidas respectivamente, y también el clúster número 1 acapara el mayor porcentaje de impresiones PLP o views, asociado al número de click y la cantidad de veces que se imprime en pantalla una PLP. Por último, en la figura 6.16 se logra identificar la distribución que presentan estos y otros conglomerados, de forma agregada, a través de sus posiciones en la PLP.

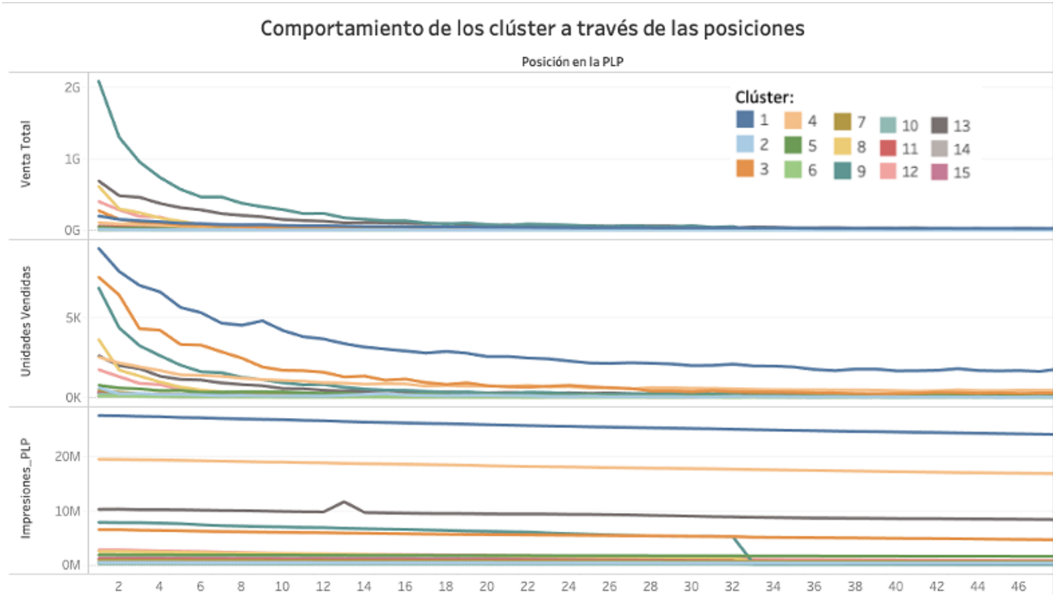


Figura 6.16: Comportamiento de los clúster a través de las posiciones. Fuente: Elaboración propia

Por último, se realizó un análisis de cada conglomerado por separado, para estudiar el comportamiento de cada categoría dentro de su respectivo clúster. A continuación se presenta el análisis para:

- **Clúster 9:** Debido a que es el conglomerado con mayor porcentaje de venta entre todos los clúster
- **Clúster 1:** Debido a que es el conglomerado con mayor porcentaje de unidades vendidas entre todos los clúster
- **Clúster 14:** Debido a que es el conglomerado con menor porcentaje de venta, y a la vez el más ruidoso de todos los clúster

El análisis para todos los conglomerados restantes, se encuentra en el anexo D.3.1.

6.5.1. Análisis clúster 9

El comportamiento a través de las posiciones para el clúster número 9, separado por categoría, se encuentra en la figura 6.17.

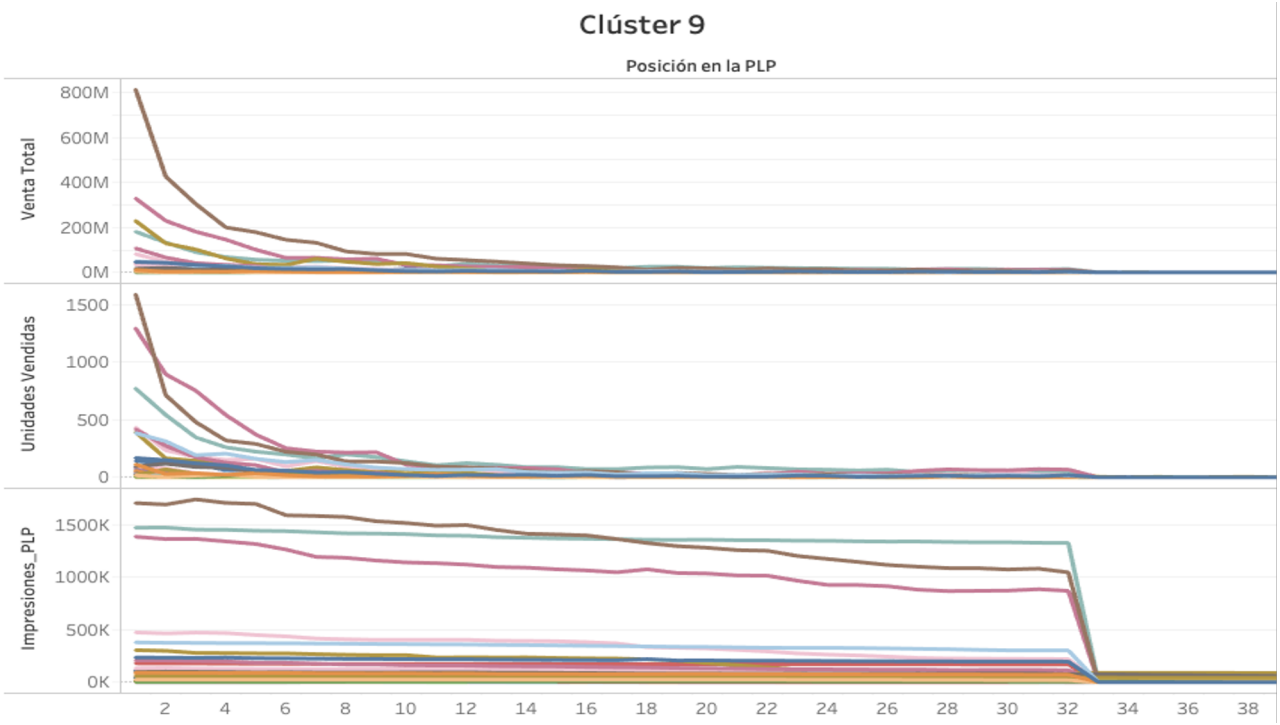


Figura 6.17: Comportamiento a través de las posiciones para las categorías del clúster 9. Fuente: Elaboración propia

De la figura 6.17, se concluye que las categorías que acumulan un mayor porcentaje de la venta, presentan un distribución pareto a través de sus posiciones, para los KPI de unidades vendidas y venta total. Además, estas categorías contienen productos hasta las posición número 33. En específico, este clúster acapara un 36,75 % de la venta total, lo que equivale a 9805 millones de pesos, un 9,75 % de las unidades vendidas, lo que equivale a 33613 unidades y un 6,18 % de las impresiones PLP, que equivalen a 214 millones.

Además, se identifica que el clúster número 9 está compuesto mayoritariamente por categorías de bebé, tecnología, dormitorio y electrodomésticos, como se puede observar en la figura 6.18, correspondiente a un histograma de las categorías presentes en el clúster. Es esperable que dentro de un conglomerado de PLP existan categorías que no se relacionan entre si, ya que el objetivo de la clusterización es clusterizar comportamiento similar a través de las posiciones, independiente de sus magnitudes.

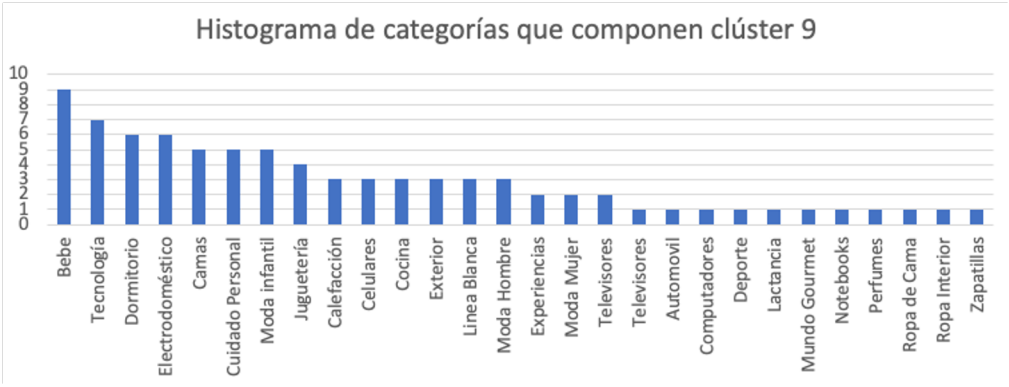


Figura 6.18: Histograma de las categorías presentes en el clúster 9. Fuente: Elaboración propia

Por último, para conocer como se distribuyen en porcentaje las categorías más relevantes de cada KPI dentro del clúster, se presentan las tablas 6.9, 6.10 y 6.11, concluyendo que la distribución de KPI dentro del clúster, es más homogénea que fuera de el.

Tabla 6.9: Distribución de venta total dentro del clúster 9. Fuente: Elaboración propia

Top 5 en Ventas		
Categoría	**Venta Total**	**% del Clúster**
Notebooks	2.919M	29,77 %
Televisores	1.556M	15,97 %
Camas	1.223M	12,47 %
Apple	908M	9,26 %
Tablets	440M	4,49 %

Tabla 6.10: Distribución de unidades vendidas dentro del clúster 9. Fuente: Elaboración propia

Top 5 en Unidades Vendidas		
Categoría	**Unidades Vendidas**	**% del Clúster**
Televisores	6.010	17,88 %
Notebooks	4.977	14,81 %
Camas	4.650	13,83 %
Colchones	2.461	7,29 %
Tablets	2.197	6,54 %

Tabla 6.11: Distribución de impresiones PLP dentro del clúster 9. Fuente: Elaboración propia

Top 5 en Vistas		
Categoría	**Vistas totales**	**% del Clúster**
Camas	44M	20,52 %
Notebooks	43M	20,38 %
Televisores	11M	16,02 %
Tablets	10M	5,15 %
Colchones	7M	5,11 %

6.5.2. Análisis clúster 1

El comportamiento a través de las posiciones para el clúster número 1, separado por categoría, se encuentra en la figura 6.19.

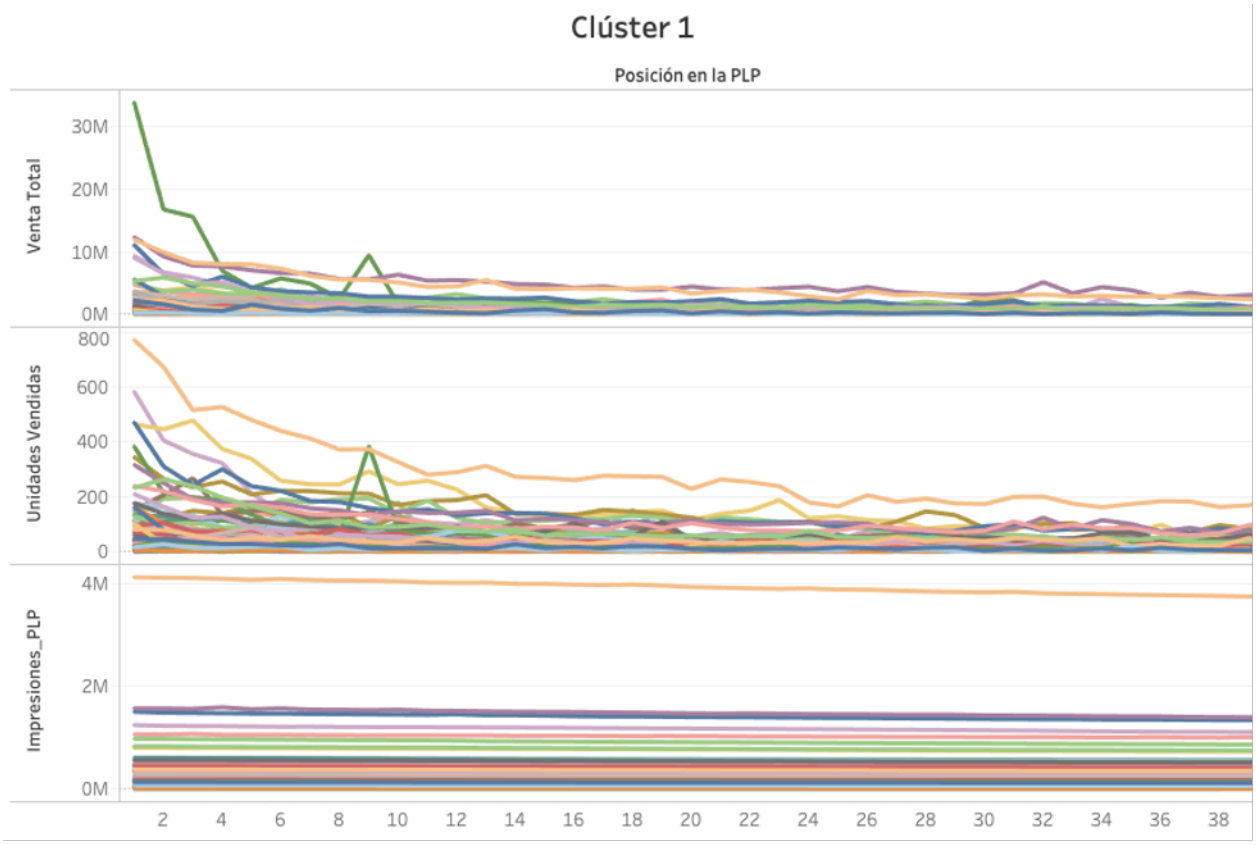


Figura 6.19: Comportamiento a través de las posiciones para las categorías del clúster 1. Fuente: Elaboración propia

De la figura 6.19 se concluye que para el conglomerado con mayor porcentaje de unidades vendidas, la distribución a través de las posiciones para los KPI venta total y unidades vendidas es más uniforme. Además, las PLP presentan productos en todas sus posiciones y las impresiones PLP en pantalla se mantienen constantes. En específico, este clúster acapara un 9,48 % de la venta total, lo que equivale a 2528 millones de pesos, un 41,3 % de las unidades vendidas, lo que equivale a 145.073 unidades y un 35,29 % de las impresiones PLP, que equivalen a 1227 millones.

El conglomerado número 1 está compuesto en su mayoría por categorías de Moda Hombre, Moda Mujer y Moda infantil como se puede observar en la figura 6.20, correspondiente a un histograma de las categorías presentes en el clúster. Desde el punto de vista del negocio, hace sentido que las categorías que más unidades vendan, y que también posean una gran cantidad de impresiones PLP, sean categorías de ropa, debido a su precio y que son productos comprados durante todo el año.

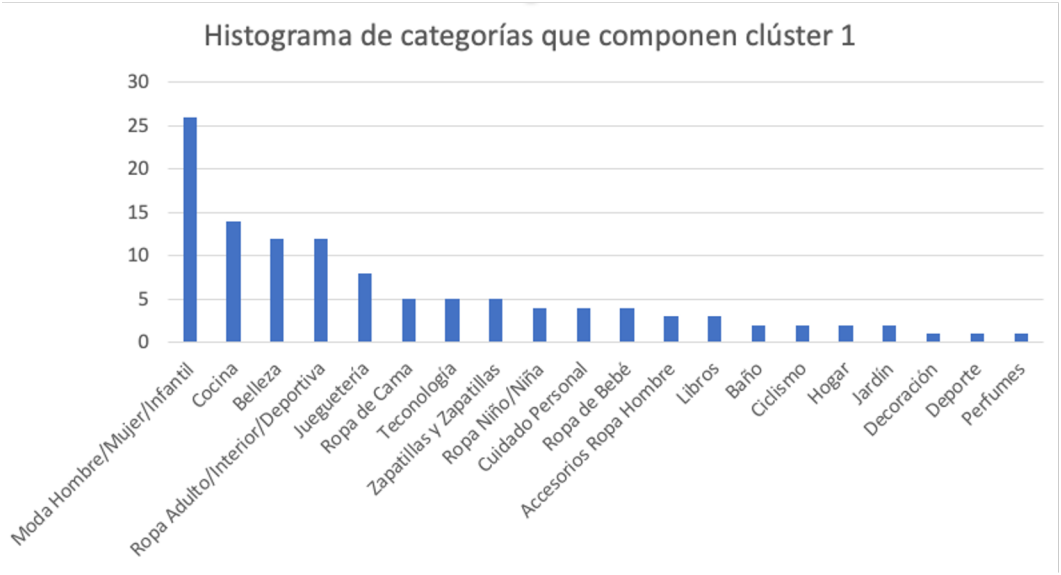


Figura 6.20: Histograma de las categorías presentes en el clúster 1. Fuente: Elaboración propia

Por último, para conocer como se distribuyen en porcentaje las categorías más relevante de cada KPI dentro del clúster, se presentan las tablas 6.12, 6.13 y 6.14, encontrando nuevamente una distribución homoénega dentro de el.

Tabla 6.12: Distribución de venta total dentro del clúster 1. Fuente: Elaboración propia

Top 5 en Venta Total		
Categoría	**Venta Total**	**% del Clúster**
Zapatilla Hombre	216M	8,54 %
Moda Mujer	199M	7,91 %
Audífonos	145M	5,77 %
Moda Hombre	112M	4,46 %
Zapatos Hombre	95M	3,76 %

Tabla 6.13: Distribución de unidades vendidas dentro del clúster 1. Fuente: Elaboración propia

Top 5 en Unidades Vendidas		
Categoría	**Unidades Vendidas**	**% del Clúster**
Moda Mujer	13.036	8,99 %
Moda Niña	7.532	5,19 %
Moda Niño	6.335	4,73 %
Moda Hombre	5.998	4,13 %
Zapatillas Hombre	5.435	3,76 %

Tabla 6.14: Distribución de impresiones PLP dentro del clúster 1. Fuente: Elaboración propia

Top 5 en Vistas		
Categoría	**Vistas totales**	**% del Clúster**
Moda Mujer	186M	215,22 %
Zapatillas Hombre	70M	5,73 %
Moda Hombre	67M	5,45 %
Juguetes	55M	4,53 %
Blusas y Poleras	49M	4 %

6.5.3. Análisis 14

El comportamiento a través de las posiciones para el clúster número 14, separado por categoría, se encuentra en la figura 6.21.

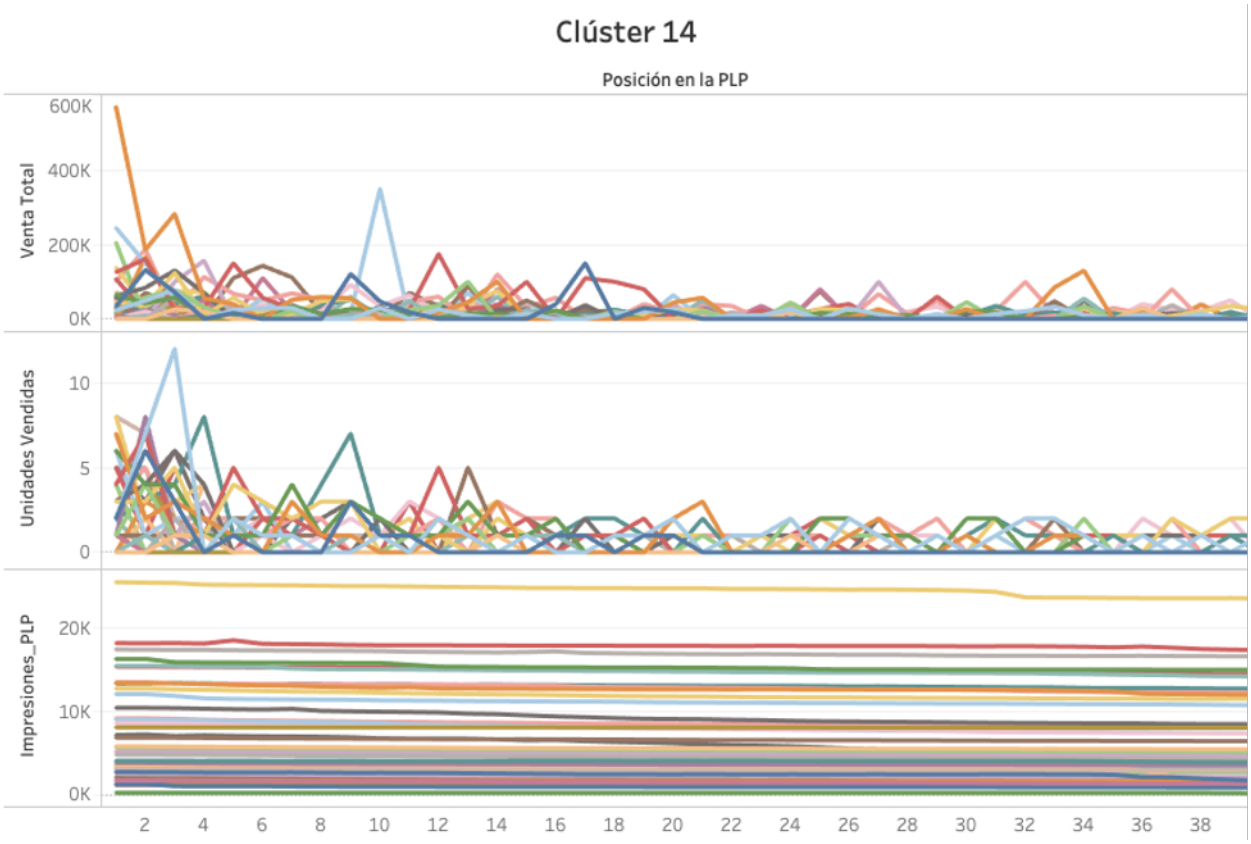


Figura 6.21: Comportamiento a través de las posiciones para las categorías del clúster 14. Fuente: Elaboración propia

Por último, para la figura 6.21 se concluye que el conglomerado con menor porcentaje de venta total, presenta la distribución más ruidosa través de las posiciones, para los KPI venta total y unidades vendidas. Además, las impresiones PLP en pantalla, asociadas a la cantidad de click que se realiza en cada una, se mantienen constantes, pero varían en magnitud, acorde al resto de los KPI. En específico, este clúster acapara un 0,07 % de la venta total, lo que equivale a 19 millones de pesos, un 0,26 % de las unidades vendidas, lo que equivale a 903 unidades y un 0,38 % de las impresiones PLP, que equivalen a 13 millones.

Las categorías mayoritariamente presentes en el clúster más ruidoso, corresponden a belleza, deportes y zapatos, como se puede observar en la figura 6.22, correspondiente a un histograma de las categorías presentes en el clúster.

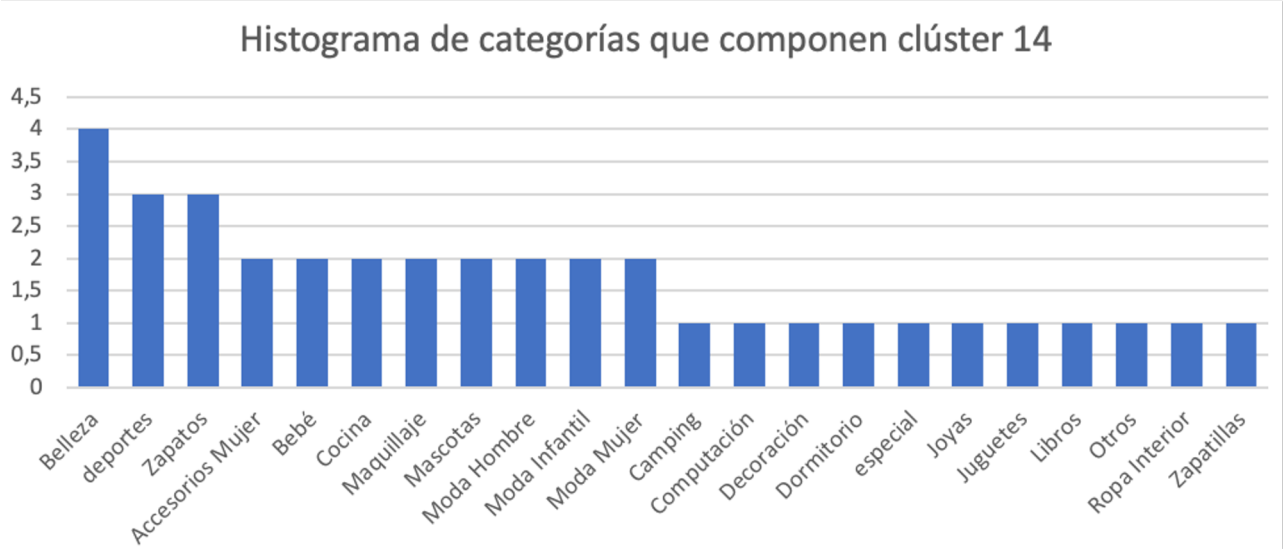


Figura 6.22: Histograma de las categorías presentes en el clúster 14. Fuente: Elaboración propia

6.6. Evaluación

Según lo expuesto en la sección 5.1.5, la evaluación del trabajo se divide en dos ejes. Primero se analizaron los resultados del modelado y se analizó cada clusterización, en las secciones 6.4 y 6.5 respectivamente.

Luego, con la clusterización realizada se procedió a modificar y evaluar el modelo de optimización actualmente desarrollado, para ver su comportamiento en cada clúster.

6.6.1. Construcción de features y target

Para la construcción de los features y targets que alimentarán el modelo, se utilizó la misma lógica de desarrollo presente en el modelo actual, basada en el análisis exploratorio de los datos, la información aportada por el área comercial y la lectura de papers que dan sustento teórico a lo desarrollado. [15], [16], [17]

Las features utilizadas se dividieron en 3 segmentos: navegación, atributos de productos y derivados de los dos segmentos anteriores.

- **Navegación:** interacciones de usuario como clicks, órdenes, venta, agregaciones al carro
- **Atributos de productos:** características propias de los productos como margen, stock, precios, marca
- **Derivadas:** se obtienen manipulando las features de los segmentos anteriores, por ejemplo, la venta acumulada 7 días, porcentaje de descuento, ratio stock (skus con stock > 0 sobre la cantidad total de skus), entre otras

Para el target se definieron 5 niveles de relevancia mediante feedback implícito del usurio (métricas del producto asociada a la navegación):

- **Nivel 0:** producto sin interacción - sin relevancia
- **Nivel 1:** producto sólo clickeado - poco relevante
- **Nivel 2:** producto sólo agregado al carro - medianamente relevante
- **Nivel 3:** producto que se vende con contribución baja - relevante
- **Nivel 4:** producto que se vende con contribución alta - muy relevante

Para dividir la contribución en dos niveles se utilizó la siguiente fórmula: [16]

$$rel_{contrib}(cat, prod) = max(2\frac{contrib_{(cat,prod)}}{(max_{prod \in cat})(contrib_{cat,prod})}) \tag{6.1}$$

Cálculo de niveles de relevancia para contribución. Fuente:[16]

6.6.2. Construcción del modelo por clúster

En lo sucesivo, se utilizó como base el modelo actual desarrollado y proporcionado por la célula de relevancia PLP, modificando los inputs de entrenamiento y predicción, adaptándolos a cada clúster previamente encontrado.

En específico, se desarrollaron 15 modelos distintos, para cada uno de los 15 clúster de comportamiento, generando 15 procesos de entrenamiento y predicción por separado.

El proceso de entrenamiento se llevó a cabo con datos del periodo agosto, septiembre del año 2020, y luego se realizaron predicciones para el mes de octubre del mismo año. Las predicciones utilizan los modelos generados para obtener las probabilidades de pertenencia a cada nivel de relevancia por producto, para luego construir el score en base a la fórmula de la sección 4.6.

6.6.3. Entrenamiento para cada modelo

Para cada uno de los 15 clúster encontrados, se realizó un entrenamiento de forma paralela. El número de productos disponibles para el entrenamiento de cada clúster se encuentra en la tabla 6.18.

Tabla 6.15: Cantidad de productos disponibles para entrenar cada clúster.
Fuente: Elaboración propia

Clúster	Cantidad de productos
1	1.881.045
2	80.034
3	342.088
4	1.542.247
5	171.293
6	72.218
7	114.552
8	127.922
9	213.552
10	58.649
11	117.565
12	394.303
13	429.652
14	58.183
15	182264

Antes de proceder con la utilización del algoritmo de clasificación multiclase, se realizó un preprocesamiento de los datos, donde se destaca la eliminación de las dos únicas variables categóricas en las features de entrenamiento (marca y categoría) ya que al incorporarlas utilizando la técnica de Leave one out encoding se veía disminuído el rendimiento final del modelo.[4]

[4] Según información entregada por la Célula

Además se eliminan de la base de datos los productos cuyo nivel de relevancia es 0, dado que son productos cuya interacción fue nula o muy baja y al igual que con las variables categóricas presentes, disminuyen el rendimiento final del modelo.[5]

Por último, dadas las carácterísticas de la función objetivo y sus niveles de relevancia se realizó un llenado de la base de datos, para evitar el desbalanceo de clases, utilizando la técnica de oversampling[6]. En la tabla 6.19 se pueden observar las clases desbalancedas y posteriormente, en la tabla 6.20 la cantidad de datos para las clases balanceadas.

Tabla 6.16: Cantidad de productos para cada clase sin balancear. Fuente: Elaboración propia

Clúster	**Clase 1**	**Clase 2**	**Clase 3**	**Clase 4**
1	1226904	506320	125765	22056
2	57344	17225	3062	2403
3	212014	86513	34922	8639
4	1201664	286622	40810	13151
5	122339	38131	7570	3253
6	58968	11042	1183	1025
7	86748	22354	2891	2559
8	95686	24360	4434	3442
9	140280	54298	13672	5302
10	46613	10043	1582	411
11	97126	17074	2190	1175
12	329796	54586	7881	2040
13	319762	88584	16479	4827
14	47269	9495	1015	404
15	154567	23686	2822	1189

[5] Según la literatura expuesta en [16]

[6] La ténica de oversampling utilziada fue RandomOversampler, y funciona igualando la clase minoritaria con la mayoritaria.

Tabla 6.17: Cantidad de productos para cada clase balanceada. Fuente: Elaboración propia

Clúster	Clase 1	Clase 2	Clase 3	Clase 4
1	1226904	1226904	**1226904**	**1226904**
2	57344	57344	57344	57344
3	212014	212014	212014	212014
4	1201664	1201664	1201664	1201664
5	122339	122339	122339	122339
6	58968	58968	58968	58968
7	86748	86748	86748	86748
8	95686	95686	95686	95686
9	140280	140280	140280	140280
10	46613	46613	46613	46613
11	97126	97126	97126	97126
12	329796	329796	329796	329796
13	319762	319762	319762	319762
14	47269	47269	47269	47269
15	154567	154567	154567	154567

Luego se procedió a entrenar cada clúster, utilzando la siguiente combinación de hiperparámetros:

- **Max depth:** 6, relacionado a la profunidad en la ramas del algoritmo
- **alpha:** 1, relacionado a la regularización L1 (evitar sobre ajuste)
- **Max delta step:**1, relacionado al trabajo con clases inbalanceadas
- **Nthread:** 32, relacionado a la capacidad de procesamiento en paralelo del algoritmo

Capítulo 7

Resultados

7.1. Resultados por clúster

En la tabla 7.1 se encuentran los resultados en términos de AUC para cada uno de los modelos entrenados. Se observa que para todos los clúster existe un ajuste superior a 0.5, lo cual indica al menos, un mejor resultado comparado con predicciones completamente azarosas.

Tabla 7.1: Métricas de evaluación para modelo por clúster. Fuente: Elaboración propia

Clúster	AUC	F1-Score	Accuracy
1	0.591	0.53	0.53
2	0.593	0.54	0.51
3	0.604	0.51	0.50
4	0.607	0.55	0.59
5	0.606	0.58	0.53
6	0.607	0.66	0.59
7	0.598	0.57	0.51
8	0.641	0.65	0.60
9	0.638	0.61	0.58
10	0.593	0.63	0.57
11	0.618	0.69	0.63
12	0.596	0.68	0.61
13	0.637	0.67	0.63
14	0.583	0.59	0.53
15	0.816	0.73	0.68

Luego, al analizar la métrica F1-score ponderada por la cantidad de elementos en cada clase para cada modelo (ver tabla 7.1) , los resultados muestran una combinación de precisión y recall ponderada en el rango de 0.5 a 0.7.

Similar a lo anterior, el accuracy para cada uno de los modelos representativos de cada clúster, se encuentra en torno al 0.6 (ver tabla 7.1)

Una etapa importante del análisis, desde el punto de vista del negocio, es estudiar como inciden las variables o features que alimentan el modelo de clasificación, en su predicción final. Para lo anterior se desarrolló un estudio del feature importance, en cada uno de los modelos desarrollados (ver anexo E.1 con información de cada clúster).

Es necesario notar como varía la importancia de los features en cada clúster, según su composición en términos del negocio. El clúster 9, que acumula la mayor cantidad de venta en dinero y presenta una distribución eminentemente Pareto, tiene como atributos más importantes para clasificar sus respectivos productos, a los features:

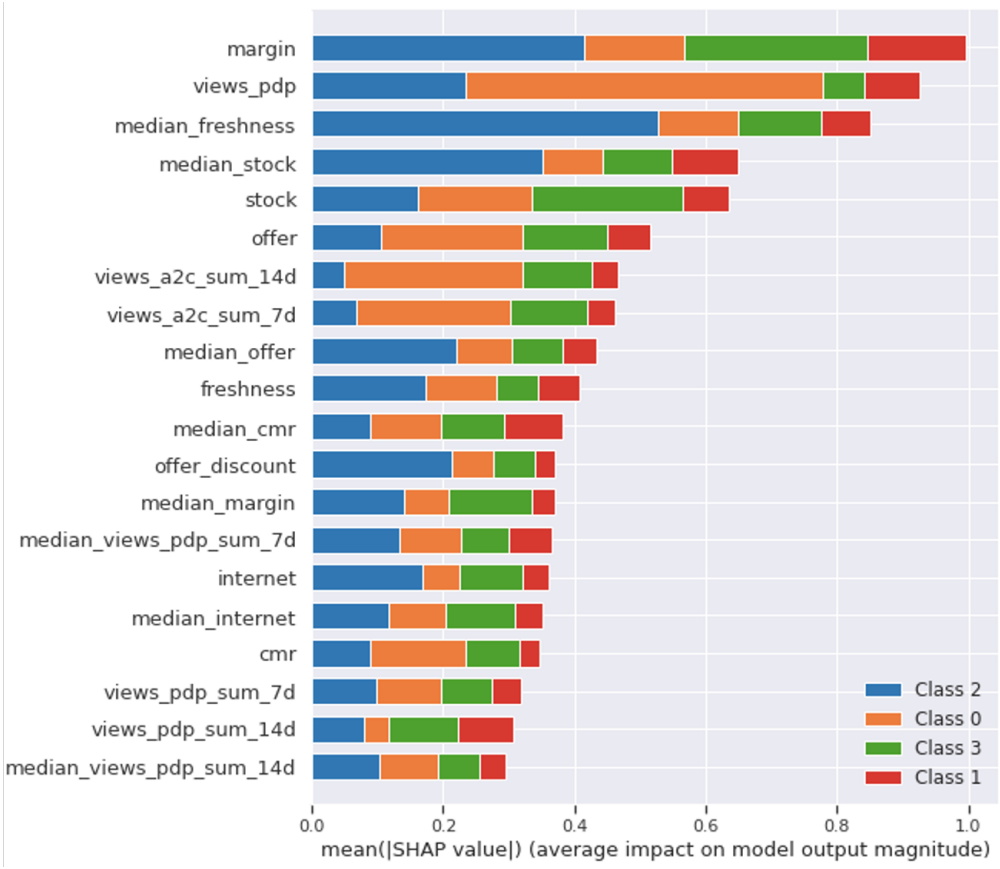


Figura 7.1: Feature Importance para clúster 9. Fuente: Elaboración propia

- Margen (margen atribuible al producto)
- Views pdp (visitas a cada producto dentro de la PLP)
- Median freshness (promedio de las veces que se ha publicado el producto en los últimos 14 días)
- Median stock (promedio del stock para los últimos 14 días)
- Stock (stock en unidades del producto)

Donde los atributos más relevantes en la clasificación del nivel de relevancia 4 (productos que son vendidos, y además dejan una contribución alta a Falabella) son los atributos Margen

y Stock.

Por otro lado el modelo número 1, representante del clúster 1, conglomerado que acumula la mayor cantidad de unidades vendidas y presenta una distribución más uniforme, muestra como atributos más importantes para la clasificación de sus productos a:

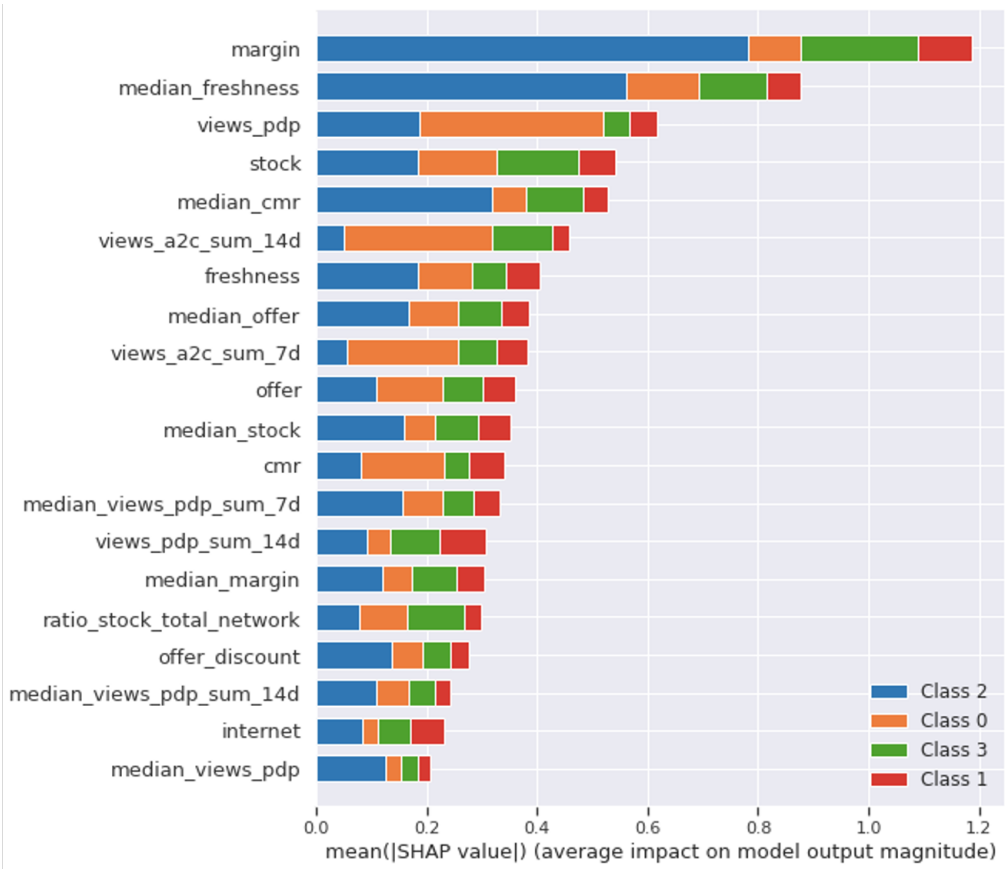


Figura 7.2: Feature Importance para clúster 9. Fuente: Elaboración propia

- Margen (margen atibuible al producto)
- Median freshness (promedio de las veces que se ha publicado el producto en los últimos 14 días)
- Views pdp (visitas a cada producto dentro de la PLP)
- Stock (stock en unidades del producto)
- Median CMR (precio promedio de los últimos 14 días al comprar con tarjeta CMR)

Donde los atributos más relevantes en la clasificación del nivel de relevancia 4, son nuevamente el margen y el stock, pero se diferencia con el anterior en presentar una menor importancia asociada a las visitas de los productos, lo cual está relacionado al comportamiento uniforme del clúster a través de sus posiciones.

Por último, aparece como atributo importante el precio de los productos, al ser comprados con tarjeta CRM, lo cual es lógico desde el punto de vista del negocio, ya que este clúster

acumula la mayor cantidad de unidades vendidas, y la compra con tarjeta CMR está asociada a descuentos en los productos.

7.2. Análisis de NDCG

En el ámbito del Machine Learning, realizar un correcto análisis de las métricas de rendimiento de un modelo, es fundamental para demostrar cuan correctas son nuestras predicciones. Sin embargo, desde el punto de vista del negocio, no basta con mostrar buenos rendimientos de accuracy o recall, si no que es necesario mostrar como se desenvuelve el modelo estudiado en base a los KPI's que se consideren más importantes.

Es por esto, que para los nuevos modelos desarrollados e implementados en la célula de relevancia PLP, se utiliza una evaluación offline, en base a la métrica NDCG. Esta evaluación permite conocer y predecir como afectará la implementación de un nuevo modelo a los KPI's que se obtienen de un modelo previamente desarrollado.

Para realizar esta comparación, primero se definió que KPI's online están correlacionados al valor del NDCG de los KPI's offline, a través de los siguienes pasos:

1. Calcular distintas métricas de relevancia para el NDCG
2. Determinar la correlación entre KPI y NDCG
3. Evaluar las correlaciones y determinar las mejores métricas de relevancia
4. Evaluar las series temporales de los KPI y NDCG

Posterior a este trabajo, se determinó a través de la correlacion y comparacion de series temporales entre KPI vs NDCG, que para las variables Conversion Rate (CR)[1] y Click Through Rate (CTR)[2] Online, la métrica que predice su comportamiento de manera offline es el NDCG de Conversión Rate (CR).

Lo anterior quiere decir, que al aumentar el valor del NDCG para el Conversion Rate, el modelo logrará aumentar el CR y CTR online de las PLP.

De manera similar, se comprobó que para las variables Contribution y Revenue online, la métrica que predice su comportamiento de manera offline es el NDCG de contribución local. (ver anexo E.2)

Por lo tanto, los pasos que se siguieron para probar un nuevo modelo y evaluar su rendimiento fueron:

1. Calcular métricas de NDCG para la contribución local y conversion rate
2. Realizar una comparación temporal para el modelo actual y el modelo nuevo

[1] El CR o tasa de conversión, es el porcentaje de usuarios que realiza una acción específica, ya sea realizar una compra, una descarga, un registro o una reserva

[2] Número de clicks que se obtiene de un elemento, respecto a su número de impresiones que se muestra en pantalla

7.2.1. Análisis de NDCG para modelo por clúster

Posterior al entrenamiento de cada modelo por clúster, se procedió a verificar si existe una potencial ganancia del negocio, al comparar el rendimiento offline de cada clúster versus el modelo actual. La comparación se realizó para el periodo de octubre 2020, y se dividió en tres etapas.

En primera instancia, se procedió a comparar el rendimiento general del modelo actual, versus el nuevo modelo, en términos del NDCG de contribución local, para predecir como se compartaría el nuevo modelo en los KPI Revenue y Contribution Online. Lo anterior siguiendo los siguientes pasos:

1. Se calcularon predicciones por clúster de PLP
2. Se unieron todas las predicciones en una tabla
3. Se realizaron predicciones para todas las PLP con el modelo actual
4. Se compararon los valores de NDCG de contribución local para ambos resultados

Se obtuvieron los valores presentes en la figura 7.3:

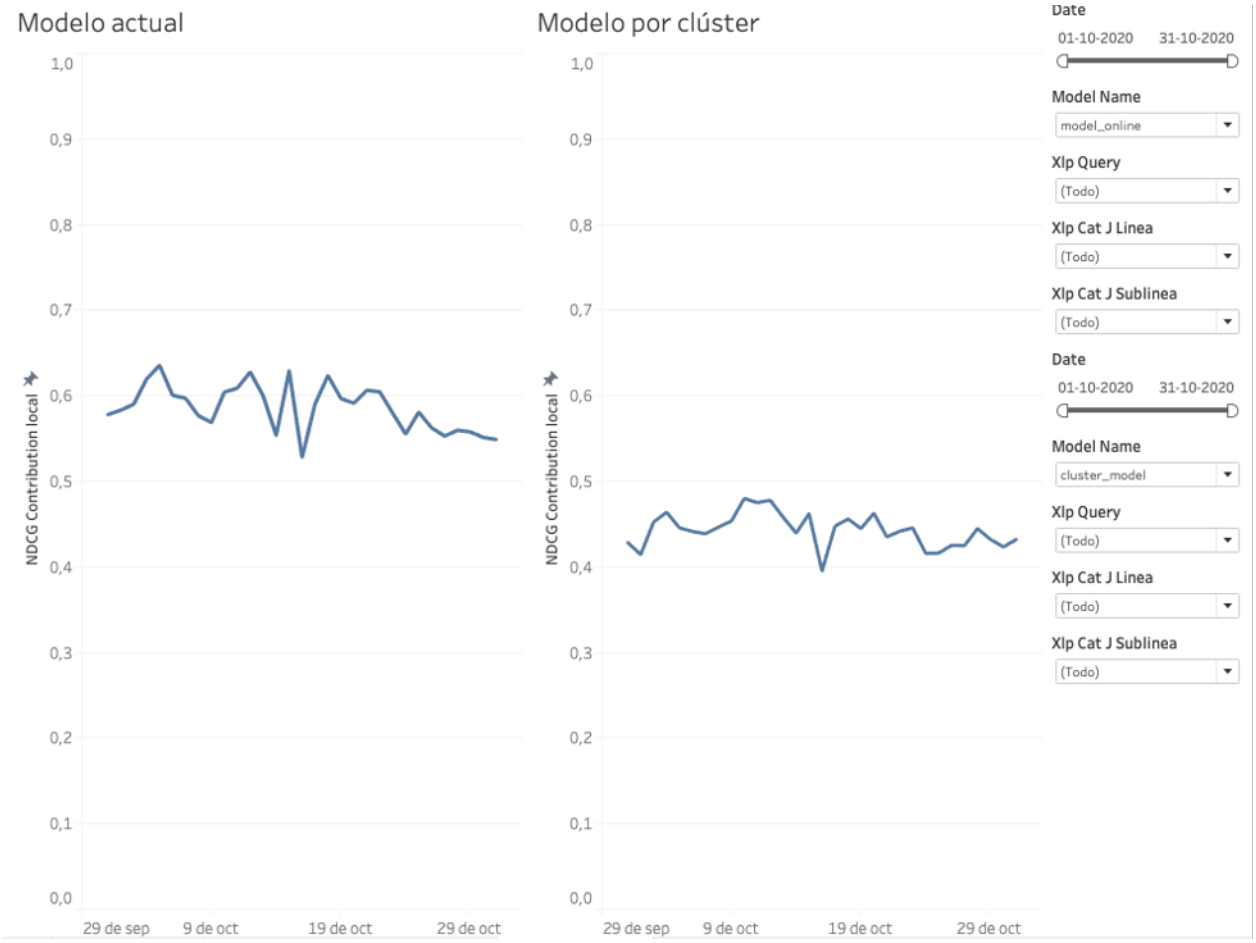


Figura 7.3: Comparación general entre valores de NDCG de contribución local para modelo actual versus modelo por clúster . Fuente: Elaboración propia

De la figura 7.3 se puede observar que el modelo por clúster presenta un menor rendimiento para el NDCG de contribución local, lo cual indica que no es un buen candidato para reemplazar el modelo actual, dado que este índice offline es un predictor del rendimiento en términos del negocio, que presentaría el modelo por clúster en caso de ser llevado a producción.

Específicamente, se obtiene que el valor promedio del NDCG de contribución local para el modelo actual es 0,58, versus los 0,46 obtenidos para el modelo por clúster, mostrando una disminución del 20,6 % entre ambos valores.

En segunda instancia, se procedió a comparar el rendimiento general del modelo actual, versus el nuevo modelo, en términos del NDCG de conversión local, para predecir como se comportaría el nuevo modelo en los KPI Conversion Rate (CR) y Click Through Rate (CTR) Online. Lo anterior a través de los siguientes pasos:

1. Se calcularon predicciones por clúster de PLP
2. Se unieron todas las predicciones en una tabla
3. Se realizaron predicciones para todas las PLP con el modelo actual
4. Se compararon los valores de NDCG de conversión local para ambos resultados

Se obtuvieron los valores presentes en la imagen 7.4

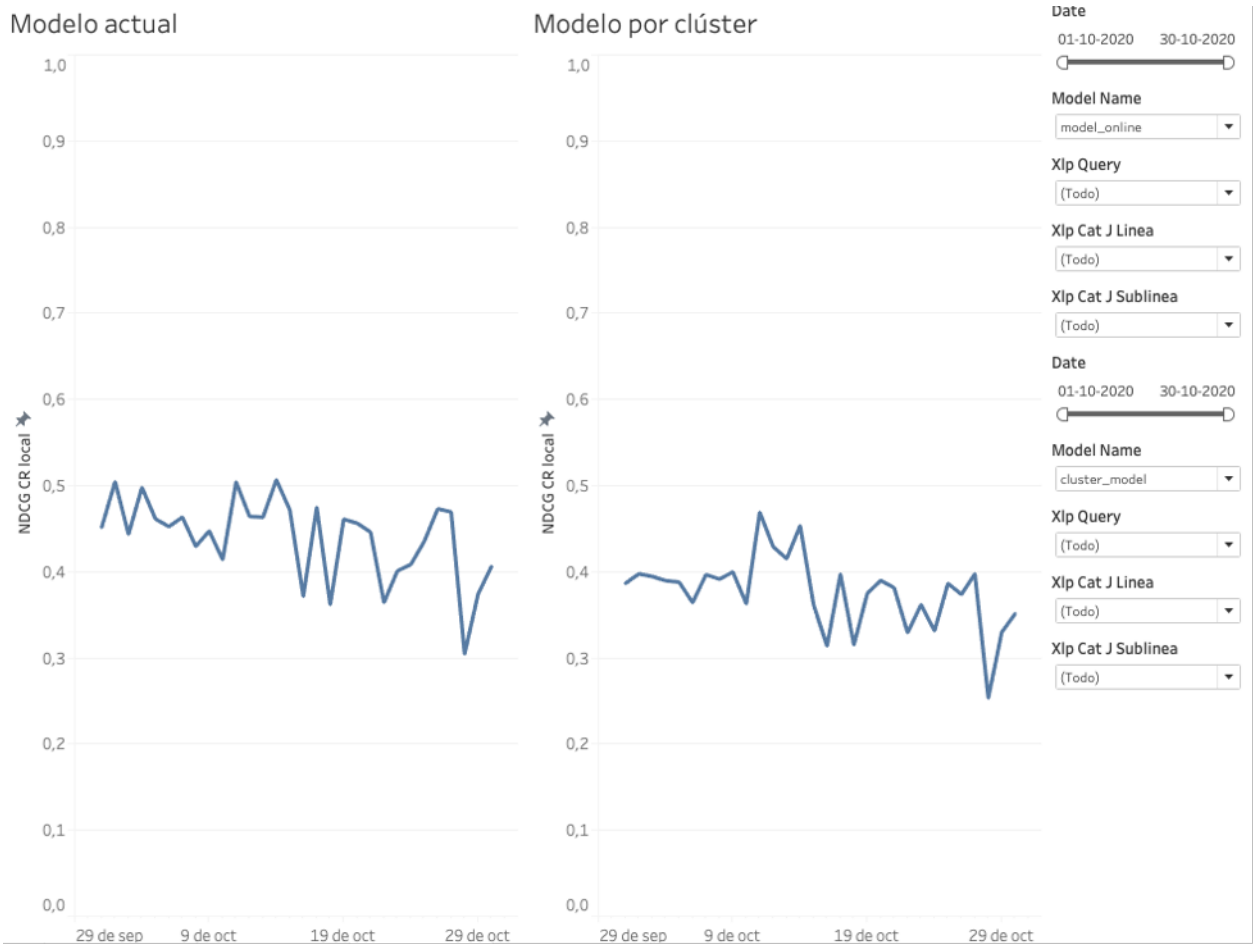


Figura 7.4: Comparación general entre valores de NDCG de conversión local para modelo actual versus modelo por clúster. Fuente: Elaboración propia

A pesar de obtener mejores resultados para el modelo por clúster en el NDCG de conversión local, este no logra igualar o superar al rendimiento del modelo actual, lo que se traduce nuevamente en indicios de un peor desempeño del nuevo modelo para los KPI online, Conversion Rate (CR) y Click Through Rate (CTR)

Específicamente, se obtiene que el valor promedio del NDCG de conversión local para el modelo actual es 0,43, versus los 0,38 obtenidos para el modelo por clúster, mostrando una disminución del 11,6 % entre ambos modelos.

En tercera instancia, para una evaluación más específica, se realizó un análisis de NDCG para cada uno de los clúster por separado, comparándolos con el desempeño del modelo actual en cada uno de las categorías de los clúster, para el periodo de octubre 2020.

A continución se muestran los resultados que se consideran más relevantes desde el punto de vista del negocio, La totalidad de resultados se expone en el anexo E.2.2.

El clúster que acumula una mayor cantidad de ventas en dinero, corresponde al conglomerado número 9, cuya distribución a través de las posiciones de la PLP se distingue del tipo pareto. Para este conglomerado, el rendimiento del NDCG de contribución local es notoriamente menor que el rendimiento del mismo indicador, para las mismas PLP, utilizando el modelo actual, ver figura 7.5.

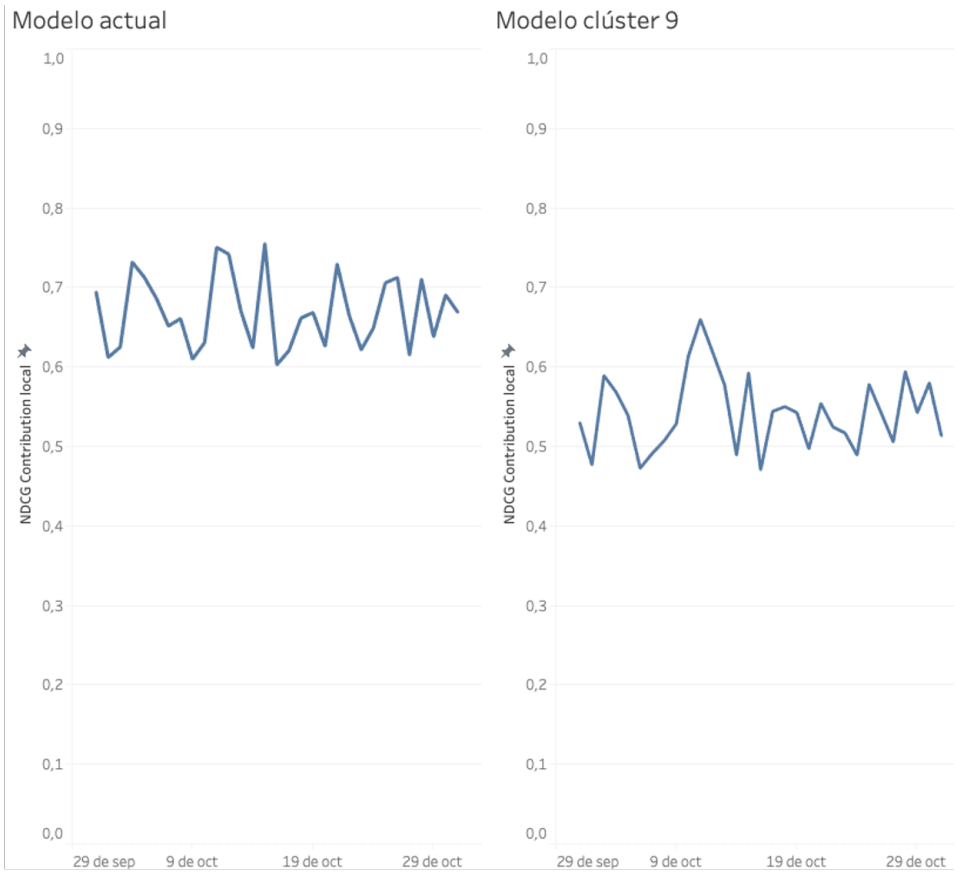


Figura 7.5: Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 9. Fuente: Elaboración propia

Luego, si se analiza el NDCG de conversión local, ocurre que si bien el modelo por clúster supera al modelo actual en el periodo del 9 al 19 de octubre, también ocurre que posterior al 19 de octubre, el modelo actual obtiene valores de NDCG notablemente más elevados, alcanzando un mínino de tan solo 0,3 NDCG, versus el 0,5 presente en el modelo por clúster, ver figura 7.6.

El conglomerado que presenta la mayor cantidad de unidades vendidas, corresponde al clúster 1, cuya distribución a través de las posiciones es más uniforme. En este caso, al analizar el NDCG tanto de contribución como conversión local, se obtienen resultados notablemente peores que los obtenidos con el modelo actual. Específicamente se observa una disminución del 37,9 % en el valor promedio del NDCG de contribución, y una dismunución del 23,9 % para el valor promedio del NDCG de conversión entre ambos modelos, ver figuras 7.7 y 7.8.

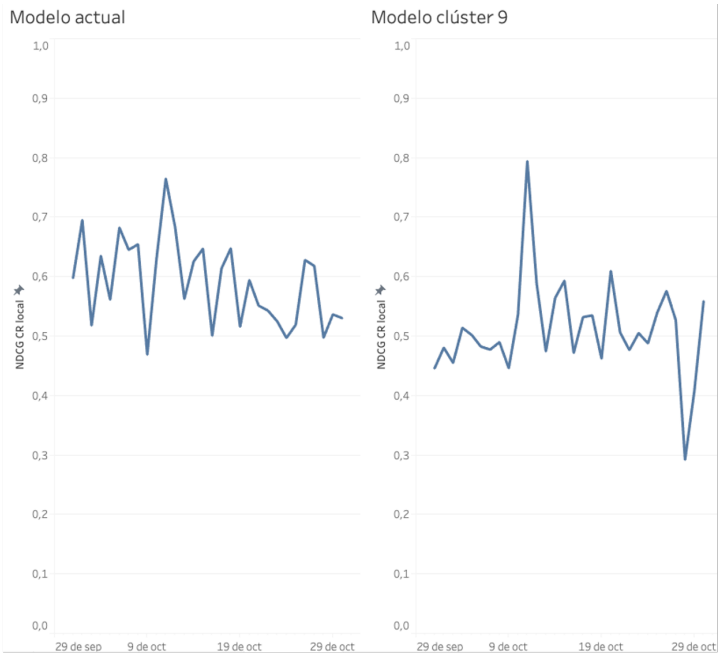


Figura 7.6: Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 9. Fuente: Elaboración propia

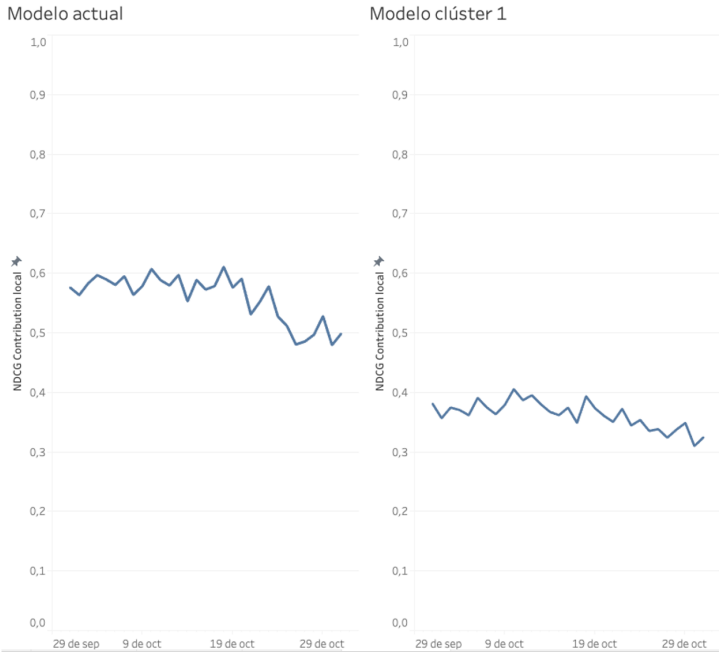


Figura 7.7: Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 1. Fuente: Elaboración propia

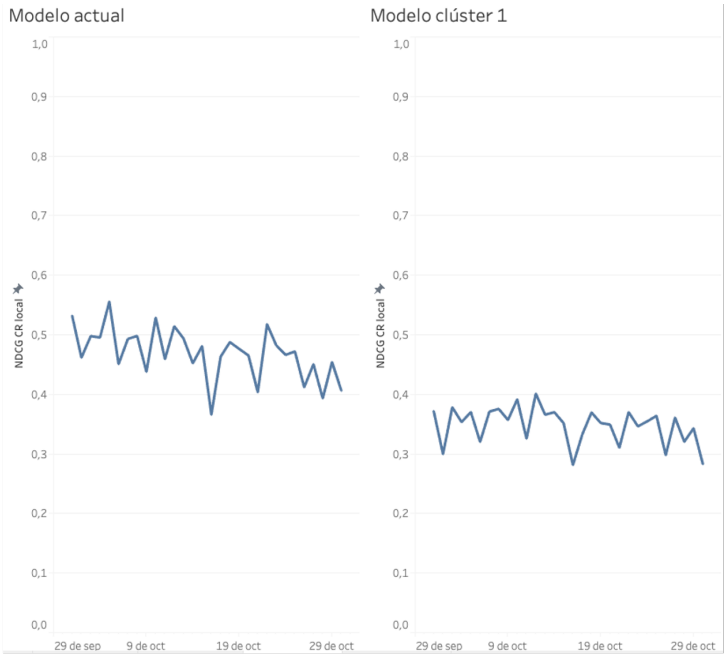


Figura 7.8: Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 1. Fuente: Elaboración propia

Por último, se analizaron dos casos cuyo comportamiento es relevante desde el punto de vista de los datos. El clúster número 15, que presenta el mejor ajuste para el modelo de clasificación multiclase, y el clúster número 14, cuya distribución a través de las posiciones en las PLP es la más ruidosa de todas las opciones.

Para el clúster número 15, se obtuvieron rendimientos muy similares tanto para el NDCG de contribución como de conversión, mostrando líneas de tendencia similar a través del tiempo y movimientos acotados dentro dentro de un mismo rango de valores, ver figuras 7.9, 7.10.

En el caso del clúster 14, a pesar de presentar la distribución más ruidosas de los conglomerados, los rendimientos obtenidos no son los peores observados entre los modelos de cada clúster. Específicamente, como se puede observar en la imagen 7.11, para el NDCG de conversión local existe un patrón de comportamiento similar, pero una disminución de rendimiento desde el día 19 de octubre en adelante.

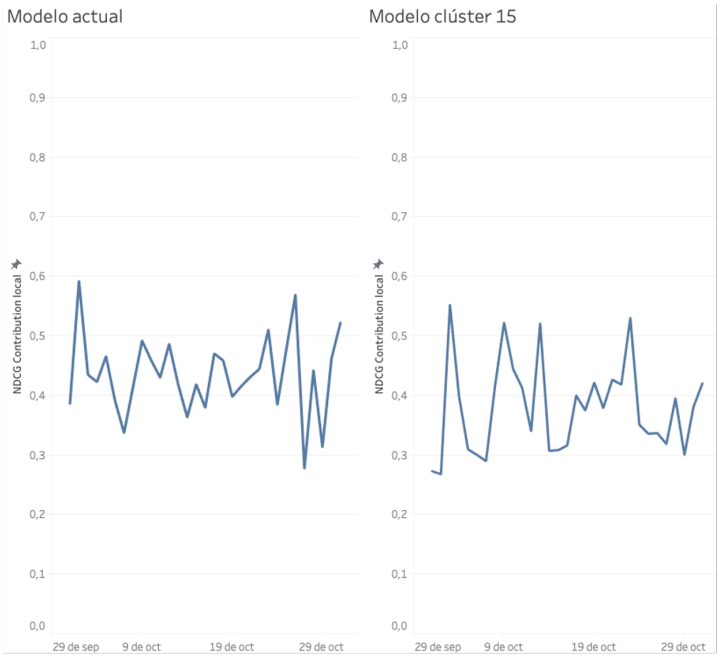


Figura 7.9: Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 15. Fuente: Elaboración propia

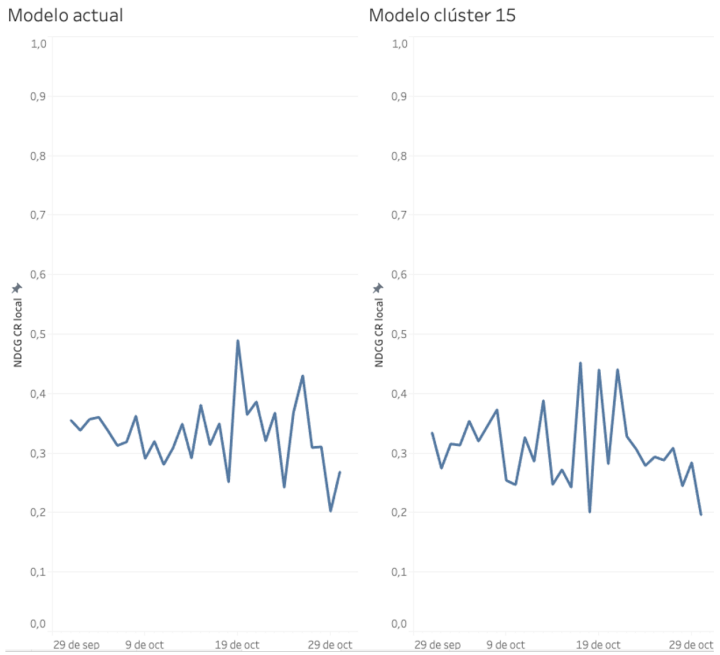


Figura 7.10: Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 15. Fuente: Elaboración propia

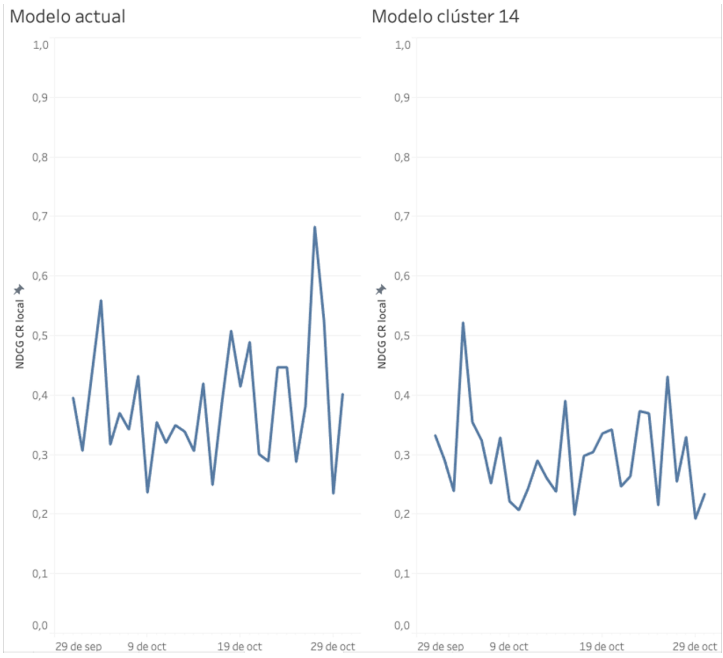


Figura 7.11: Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 14. Fuente: Elaboración propia

Capítulo 8

Conclusiones

La cantidad de transacciones que se realizan diariamente a través de los canales de ecommerce, aumentan la cantidad de datos que se almacenan, procesan y generan en las páginas web. Trabajar en Falabella.com fue una prueba fehaciente de esto, logrando constatar que las Listas de Productos de cada una de las categorías presentes en la página, mueven terabytes datos y representan entre un 20 % y un 30 % de las ventas totales del sitio, para julio 2020.

Dado lo anterior, es que Falabella busca optimizar la forma en que se muestran los productos dentro de estas listas en la página web. Para este objetivo, se realizó un clustering de PLP's en base a su comportamiento por posición, con el fin de entregar este input de forma explícita al modelo que se utiliza en el área, buscando optimizar sus resultados.

La clusterización fue realizada a través un método de clustering jerárquico acumulativo, utilizando como medida de distancia la correlación de pearson y un método de unión de clúster (o linkage) de tipo Complete. Se obtuvieron 15 diferentes conglomerados, con interpretabilidad desde el punto de vista matemático y del negocio.

De la clusterización se puede concluir principalmente, que existen diferencias en el comportamiento de la venta y unidades vendidas para PLP originadas por diferentes categorías, lo cual implica un comportamiento diferente de cara al usuario que navega en la página, cuando visita categorías de diferente tipo. En específico, se puede concluir que:

1. Las categorías que más venden en dinero se encuentran relacionadas a tecnología
2. Las categorías que más venden en unidades están asociadas a Moda Hombre, Moda Mujer y Moda Infantil
3. Mientras mayor es el porcentaje de venta, el clúster presenta una distribución más cercana a una distribución de pareto
4. Mientras menor es el porcentaje de venta, mas ruidoso es el comportamiento a través de las posiciones del clúster
5. Los clúster que abarcan mayor porcentaje de unidades vendidas, se asemejan más a una distribución uniforme

6. Las impresiones PLP, asociadas a las veces que se imprime una PLP en pantalla, que se a su vez se encuentran relacionadas con los clicks de una PLP, se mantienen constantes, variando en magnitud acorde a las demás variables.

7. Las categorías de ropa (Moda Hombre, Moda Mujer y Moda Infantil) están presentes de en más de un clúster, dada sus múltiples variaciones (recordar que una categoría puede dar origen a más de una PLP)

8. Al analizar los clicks dentro de los elementos de una PLP, es decir los clicks a cada producto a través de las posiciones, estos se relacionan con la venta en un 62 %.

Posterior a lo anterior, se realizó una modificación del modelo globlal de optimización, para entrenar por separado cada conglomerado de datos, y comparar el rendimiento de ambos modelos a través de la métrica NDCG, para octubre del año 2020.

De la evaluación de rendimiento para el modelo por clúster, a pesar de obtener mayor NDCG en determinados conglomerados (ver figura 7.6), también se obtuvieron peores resultados en más de una ocasión (ver figura 7.9, 7.5, 7.10), lo cual indica de que entregar este input de forma explicita, no mejora los resultados obtenidos hasta el momento.

Para realizar un análisis general entre ambos modelos, se unieron la totalidad de las PLP, de forma posterior a su entrenamiento por clúster, y luego se compararon con el modelo actual. De esta comparación, se concluyó que el modelo por clúster efectivamente presenta un menor rendimiento a través del tiempo, disminuyendo, en promedio, un 29,6 % el NDCG de contribución local, y un 11,6 % el NDCG de conversión local para el periodo de tiempo estudiado.

Al analizar específicamente cada modelo por clúster, se obtuvieron las siguientes conclusiones:

1. Para los clúster que acumulan un mayor porcentaje de venta, el rendimiento es notoriamente menor, comparado al modelo actual

2. El nivel de ruido que presentan las distribuciones a través de las posiciones de una PLP, no está correlacionado con el resultado del rendimiento del modelo.

Finalmente, dado el trabajo realizado y los resultados obtenidos, se concluye que realizar una clusterización de comportamiento por posición para las PLP originadas por las categorías del sitio Falabella.com, aporta información valiosa con interpretabilidad del negocio acerca del comportamiento de los usuarios en la página, pero utilizar esta información como input directo para el modelo global que busca optimizar la contribución, no genera un cambio positivo ni aumento de los indicadores de interés económico.

Por lo tanto, ya que las conclusiones no fueron las esperadas, se dejan planteadas 3 hipótesis que buscan explicar el origen de los resultados.

En primer lugar, actualmente para la confección del clasificador multiclase utilizado, se construye el target de ranking considerando 5 niveles de relevancia, que van desde el nivel 0, con productos que no presentan interacción con el usuario, al nivel 5, con productos que

son vendidos y además registran una alta contribución para Falabella [16]. La construcción anterior responde a un trade off entre la especificidad de la clasificación de los productos, versus la cantidad de niveles de relevancia que tendrá el modelo, en otras palabras, que tan específico se busca clasificar los productos, versus, cuantas clasificaciones distintas se tendrán.

Dicha distinción se basa en el desarrollo acerca de las aplicaciones del Learning to Rank en las búsquedas de ecommerce, presentado por Shubhra Kanti Karmaker Santu, Parikshit Sondhi, ChengXiang Zhai (2019) y Ping Li, Christopher J.C. Burges, Qiang Wu (2007). Una posible hipótesis de porque se obtuvieron peores resultados a través de los clúster, es debido al alto nivel de exigencia con que el modelo busca clasificar los productos para los 5 niveles de relevancia. Si bien la experiencia teórica muestra resultados correctos, en el desarrollo empírico se pueden buscar otras alternativas de relevancia.

Una segunda hipótesis está relacionada con la cantidad de categorías que alimenta cada modelo. Actualmente, el modelo global optimización toma como set de entrenamiento las 4000 PLP activas en el sitio. Al dividir las PLP según su comportamiento por posición, los conglomerados resultantes son una división de la cantidad total de categorías (ver tabla 6.6) por lo que al tener menor cantidad de datos, se podría traducir en menor poder predictivo.

Por último, una hipótesis que busca una explicación teórica de los resultados, tiene relación con la medida de similitud utilizada en la clusterización y la herramienta de machine learning utilizada en la construcción del clasificador. Al utilizar como herramienta un árbol de decisión para la clasifición multiclase, se tiene que los resultados se obtienen a través de agrupamientos y cortes en los nodos creados por el algoritmo, basados en el cumplimiento de reglas de decisión sobre grupos de características similares. Lo anterior, no toma en cuenta la correlación en el comportamiento de las PLP a través de las posiciones, regla fundamental con la cual fue realizada la clusterización y posterior análisis de los datos.

8.1. Trabajos futuros

En pos de mejorar las métricas del negocio, se dejan propuestos posibles cambios en la metodología utilizada, que escapan al alcance de este trabajo, pero que buscan obtener posibles mejores resultados.

Acorde a las hipótesis sobre los resultados obtenidos, se plantean 3 posibles cambios:

1. Utilizar diferentes niveles de relevancia que permitan tener menos especificidad para clasificar productos según las interacciones con el usuario
2. Construir un tipo de clusterización que entregue como resultado clústers que obtengan mejores medidas de precisión accuracy y recall para el modelo de optimización utilizado actualmente
3. Construir un nuevo modelo de optimización, utilizando herramientas de machine learning que tomen en consideración de forma explicita la correlación en el comportamiento productos a través de las posiciones de las PLP

Con lo anterior, queda propuesta una nueva comparación de modelos, para lo cual se debe seguir la metodología planteada en el presente trabajo de título.

Bibliografía

[1] World Economic Forum. Future of Retail Insight Report. (2017). URL http://www3.weforum.org/docs/IP/2016/CO/WEF_AM17_FutureofRetailInsightReport.pdf

[2] Diario Financiero. Cierre forzado de tiendas golpea a Falabella y reporta pérdidas de US 157 millones. (2020). URL https://www.df.cl/noticias/empresas/retail/cierre-forzado-de-tiendas-golpea-a-falabella-y-reporta-perdidas-por-us/2020-08-25/193024.html

[3] Gareth James. Las 10 tendencias del eCommerce en el mundo. (2020). URL https://www.prochile.gob.cl/wp-content/uploads/2019/01/Las_10_Tendencias_del_ecommerce_en_el_mundo-cap_10_.pdf

[4] América Retail. Venta Online de Falabella crece 124 % durante el segundo trimestre de 2020. (2020). URL https://www.america-retail.com/chile/venta-online-de-falabella-crece-124-durante-el-segundo-trimestre-de-2020/

[5] Diario Financiero. Cierre forzado de tiendas golpea a Falabella y reporta pérdidas de US$157 millones. (2020). URL https://www.df.cl/noticias/empresas/retail/cierre-forzado-de-tiendas-golpea-a-falabella-y-reporta-perdidas-por-us/2020-08-25/193024.html

[6] Jorge Juan (1990) Platón. La República o El Estado. Madrid. Editorial EDAF , 30 - 28001.

[7] Beca Cofre, S. Clustering Difuso con Selección de Atributos. (2007) URL http://repositorio.uchile.cl/handle/2250/104686

[8] Alboukadel Kassambara (2017) Practical Guide to Cluster Analisys in R. STHDA

[9] Santiago De La Fuente Fernandez (2011). Análisis de Conglomerados. Madrid. UAM

[10] Gareth James, D. W. (2018). An Introduction to Statistical Learning. New York, Springer.

[11] Keith Henderson, Brian Gallagher, and Tina Eliassi-Rad. (2015). EP-MEANS: an efficient nonparametric clustering of empirical probability distributions. Association for Computing Machinery, New York, NY, USA, 893–900.

[12] Eric-Joel Blanco-Hermida Sanz (2016) Algoritmos de clustering y aprendizaje automático aplicado en Twitter. Univerisdad Politécnica de Cataluña. Cataluña

[13] Diego Fernando Vallejo Huanga (2015 - 2016) Clustering de documentos con restricciones de tamaño. Universidad Politécnica de Valencia. Valencia

[14] Victor Galan Cortina (2015) Aplicación de la metodología CRISP-DM en un proyecto de minería de datos en el entorno universitario. Universidad Carlos III de Madrid, Madrid.

[15] Falk K. Practical (2019) Recommender Systems, Manning Publications Co. United Sta-

tes of America.

[16] Li P. and Burges C. and Wu Q. (2007) McRank: Learning to Rank Using Multiple Classification and Gradient Boosting, Advances in Neural Information Processing Systems, Canada.

[17] Kanti S. and Sondhi P. and Zhai C. (2017) On Application of Learning to Rank for E-Commerce Search, International ACM SIGIR Conference on Research and Development in Information Retrieval, Japan.

Anexo A

Antecedentes generales

A.1. Anexo 1

Organigrama de la Célula de Relevancia PLP

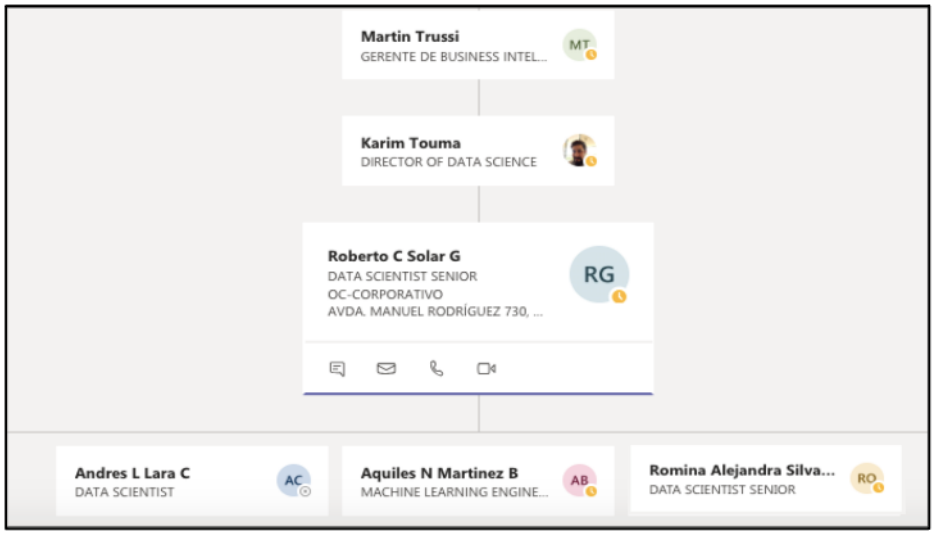


Figura A.1: Organigrama Célula de Relevancia PLP. Fuente: Teams

Anexo B

Oportunidad identificada

B.1. Anexo 2

Historial de interacciones e información transaccional registrada para los clientes de Falabella.com al navegar dentro de una PLP

Tabla B.1: Descripción de interacciones registradas . Fuente: Elaboración propia

Elemento	Descripción
Date	Fecha en la que se registra el evento
Xlp_page_number	Número de página del evento registrado
Xlp_position	Posición del evento registrado
Flags	Representan elecciones binarias (0 o 1) que se utilizan para especificar los eventos registrados
Xlp_order	Indica el orden en que se registró el evento dentro de la página (recomendados, menor a mayor precio etc..)
Xlp_query	Indica la categoría a la cual pertenece el evento
Pdp_cat_j	Indica el id del producto del cual se registra el evento
Views_xlp_pos_local	Cantidad de veces que se imprime en pantalla el evento
Visits_xlp_pos_local	Cantidad de visitas registradas para el evento
Visitors_xlp_pos_local	Cantidad de visitantes registrado para el evento
Gross_purchase	Dinero vendido asociado al evento resgistrado
Units_purchase	Cantidad vendida asociada al evento registado

Anexo C

Marco teórico

C.1. Anexo 3

C.1.1. Algoritmos de partición

Este grupo de métodos se caracteriza por ser en general el más preciso (pero no necesariamente el más válido) para generar un criterio de clustering. En particular, los algoritmos de partición requieren de un parámetro entregado por el usuario, que indique una cantidad fija de clúster a encontrar dentro de los datos en estudio.

Luego, a través de una función objetivo, típicamente WGSS (Within-Group Sum of Square errors) se evaluan las particiones candidatas a través de la varianza de los sub-grupos resultantes. De esta manera, al encontrar un mínimo local de la función objetivo, se considera haber llegado a un óptimo de clustering para el criterio planteado.

Clásicamente estos métodos han estado asociados únicamente a poder encontrar clúster con formas esféricas y de tamaños similares, no obstante, con el cambio de la métrica de distancia utilizada, es posible encontrar estructuras muy diversas. Algunos métodos clásicos en esta área son el algoritmo K-Means y Fuzzy C-Means.[7]

C.1.2. K-Means

El algoritmo de K-means es una conocida aproximación para particionar un conjunto de datos en K distintos grupos no solapados entre si. Un requisito para el uso del algoritmo consiste en especificar a priori el número k de clúster que se desea encontrar, asignando a cada observación uno y solo uno de los cluster encontrados.

El algoritmo funciona en base a una iteración de distancias, minimizando las distancias intra-cluster y maximizando las distancias inter-cluster. En la figura C.1 se observa el resultado de utilizar K-means para 150 observaciones simuladas en dos dimensiones, usando tres distintos valores de K. [8]

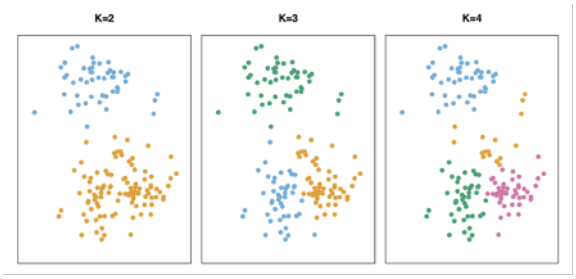


Figura C.1: : Ejemplo de clusterización con K-mean para 150 observaciones en R. Fuente: Practical Guide to Cluster Analysis in R [8]

C.2. Anexo 4

C.2.1. Medidas de distancia geométrica

Distancia Euclidiana: Es la medida de distancia más usada, se basa en el teorema de Pitágoras para calcular la distancia directa entre dos puntos. Por ejemplo, al suponer a y b dos puntos dentro del plano bidimensional, la distancia euclidiana entre a y b es calculada como la distancia física entre ambos puntos según la siguiente figura C.2: [8]

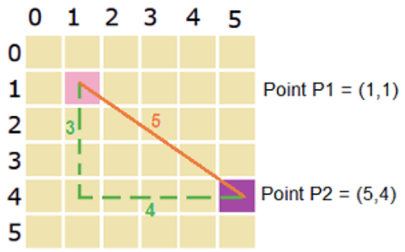


Figura C.2: Definición y descripción de la medida de distancia euclidiana. Fuente: Practical Guide to Cluster Analysis in R [8]

Distancia de Manhattan: Medida de distancia que captura la distancia entre dos puntos agregando la diferencia absoluta en pares de observaciones,C.3 es decir, ya no es la distancia recta en el plano cartesiano, si no que es la distancia asumiendo movimientos solo en el eje X e Y (para espacios bidimensionales).[8]

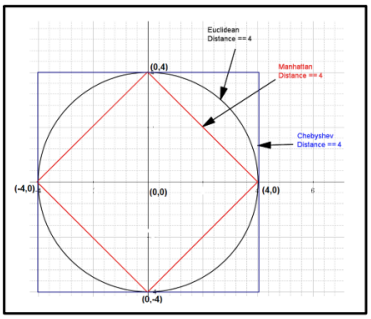


Figura C.3: Definición y descripción de la medida de distancia de Manhattan. Fuente: Practical Guide to Cluster Analysis in R. [8]

C.3. Anexo 5

C.3.1. Métricas de evaluación para modelo global de optimización

A continuación se detallan los pasos requeridos para calcular la métrica NDCG sobre un conjunto de documentos

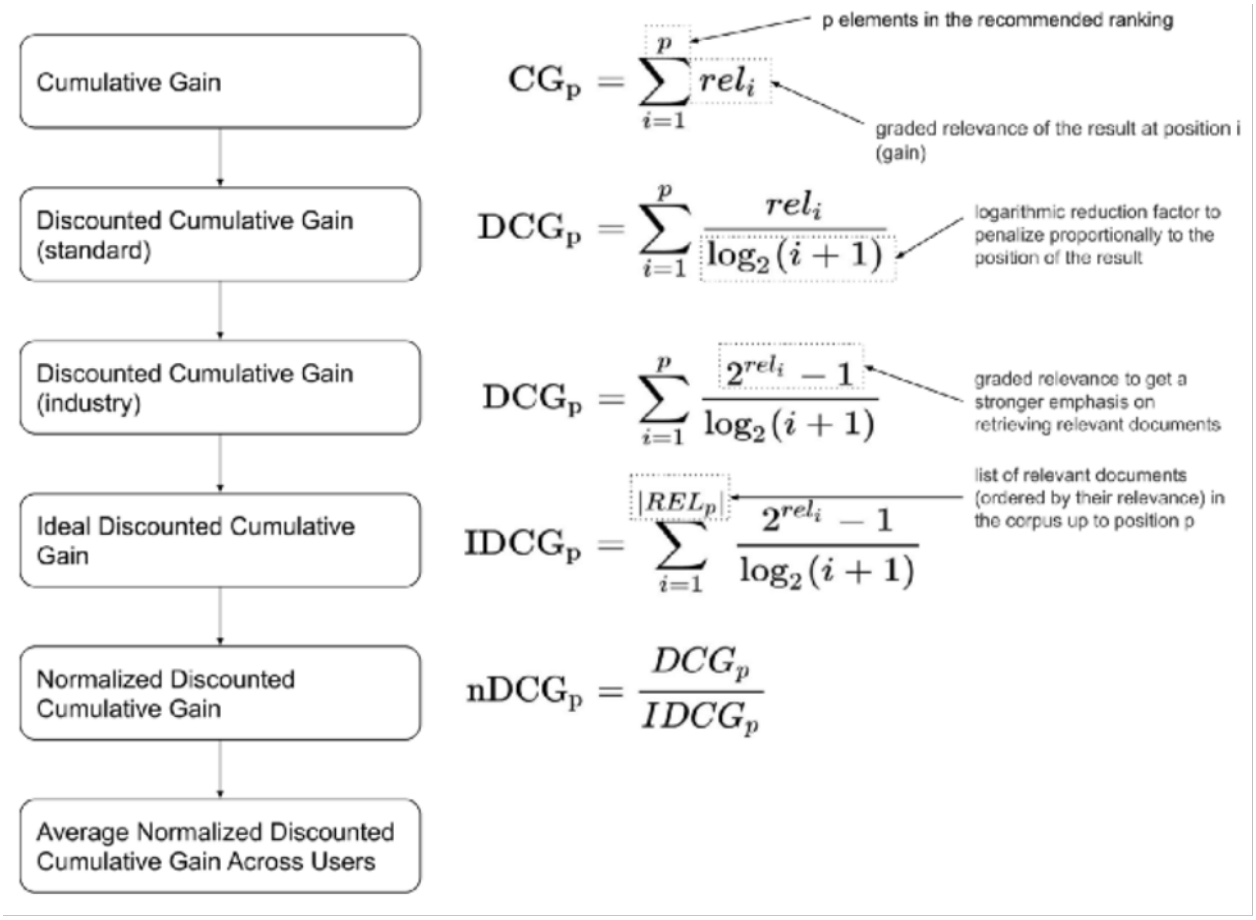


Figura C.4: : Metodología de cálculo de NDCG. Fuente: [8]

C.4. Anexo 6

C.4.1. Estado del arte en clusterización de distribuciones

En abril de 2015, Keith Henderson, Brain gallagher y Tina Eliassi publicaron su investigación titulada "EP-MEANS: An efficient nonparametric clustering of empirical probability distributions" [11], un algoritmo basado en K-means que, dada una colección de distribuciones de probabilidad empírica, permite clusterizar de manera eficiente distribuciones con formas arbitrarias, unidimensionales y continuas.

El algoritmo se sustenta en 3 principios fundamentales, ser eficiente, con un costo de procesamiento O(N), ser no paramétrico, es decir no depender de elecciones manuales de

parámetros y ser empírico, siendo capaz de clusterizar distribuciones sin conocer la familia a la cual pertenecen.

Los autores realizaron como ejemplo de motivación una aplicación de este algoritmo en la clusterización de rutas aéreas para aerolíneas de EE.UU. La base de datos utilizada consistió en cientos de registros para distintas aerolíneas que mostraba 3 columnas de datos Route, Airline y Distance (nombre de la ruta, nombre de la aerolínea y distancia en millas respectivamente).

El objetivo a lograr, era clusterizar las aerolíneas por modelo de negocio, encontrando grupos que presentaran rutas similares. Luego de realizado el experimento, se obtuvieron resultados alentadores, (ver figura figura C.5), encontrando tres clúster para las rutas estudiadas, es decir, había tres distribuciones de viajes predominantes.

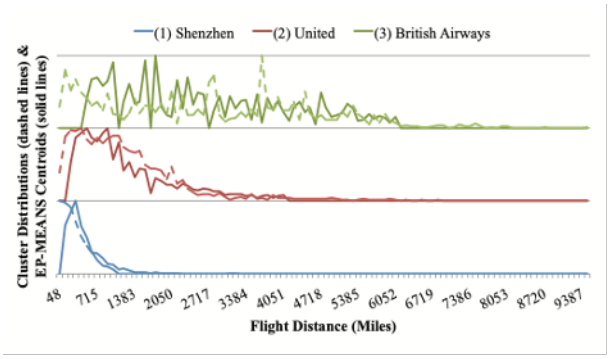


Figura C.5: Resultados obtenidos de aplicar algoritmo EP-MEANS para clusterizar aerlíneas basadas en sus distancias de vuelo. Fuente: [11]

Tabla C.1: Descripción de conglomerados encontrados por modelo de negocio. Fuente:[11]

Clúster 1 Local and Regional	**Clúster 2 Domestic**	**Clúster 3 International**
Shenzen	United	British Airways
Wizz Air	Ryanair	Korean Air
Xiamen	Delta	Emirates
Hellas Jet	American	Qatar Airways
Sichuan	US Airways	Transaero

Anexo D

Desarrollo metodológico

D.1. Anexo 7

D.1.1. Funnel de compra

Por definición, un funnel de compra identifica los pasos que debe seguir un cliente en un comercio online, para la obtención de un objetivo final, que puede variar según el elemento que se este analizando (formulario, subscripción, email marketing o ecommerce).

En el caso de Falabella, y en particular en el área de Data Science, el funnel analizado corresponde al funnel de compra, cuyo objetivo final consiste en medir la conversión de los productos, e identificar cuantas visitas se van perdiendo en cada una de las etapas de compra, desde la PLP hasta la conversión final.

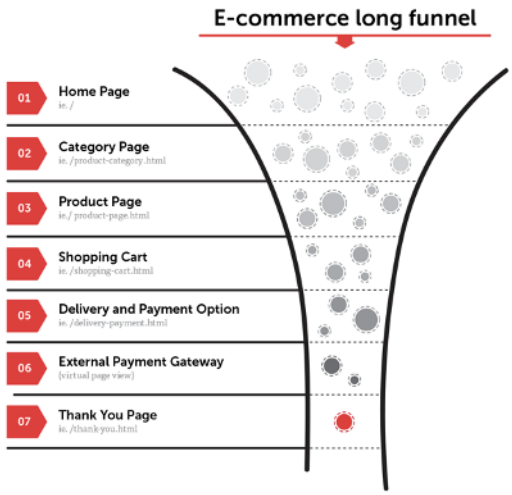


Figura D.1: Funnel de compra para un Ecommerce. Fuente: wwww.ibeschool.cl

Actualmente, el funnel desarrollado por el área de Relevancia PLP, sirve como input para el modelo global, que busca mejorar la contribución de los productos en base al ordenamiento de estos dentro de la PLP, según los datos que se van obteniendo del funnel.

D.2. Anexo 8

D.2.1. Elección de un número k de clúster

Se presentan a continuación las distribuciones del % de venta para los candidatos k= 5, 11, 15, y 18.

Tabla D.1: Distribución de venta para K=5 . Fuente: Elaboración propia

Clúster	% de Venta
2	77,22 %
4	10,64 %
1	7 %
5	4,5 %
3	0,64 %

Tabla D.2: Distribución de venta para K=11 . Fuente: Elaboración propia

Clúster	% de Venta
7	44,42 %
10	21,16 %
1	9,48 %
9	8,19 %
2	6,14 %
3	5,40 %
4	1,81 %
11	1,23 %
6	0,96 %
8	0,93 %
5	0,27 %

Tabla D.3: Distribución de venta para K=15 . Fuente: Elaboración propia

Clúster	% de Venta
9	36,75 %
13	21,16 %
1	9,48 %
8	7,67 %
12	6,44 %
3	5,68 %
4	5,40 %
5	1,81 %
11	1,74 %
15	1,16 %
7	0,96 %
10	0,93 %
2	0,46 %
6	0,27 %
14	0,07 %

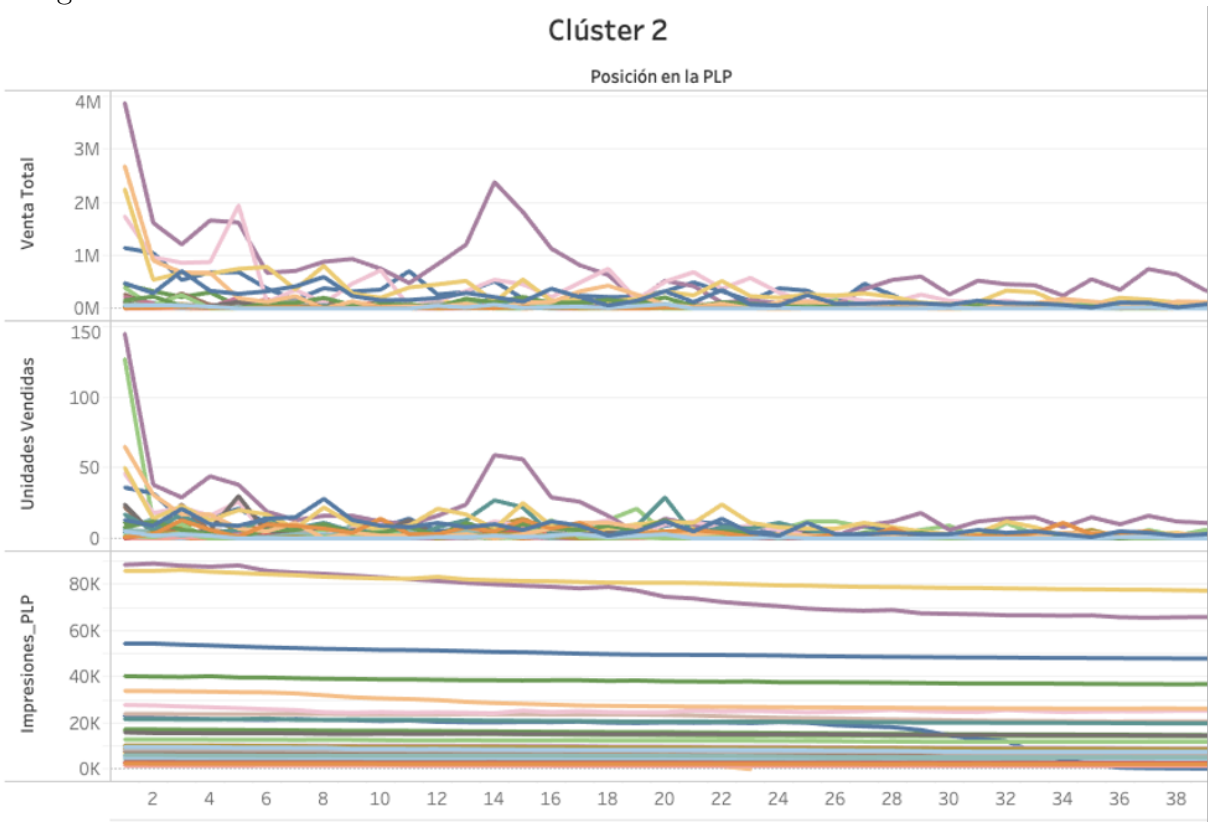
Tabla D.4: Distribución de venta para K=18 . Fuente: Elaboración propia

Clúster	**% de Venta**
11	36,75 %
16	21,16 %
1	9,48
10	7,67 %
15	6,25 %
5	5,40 %
4	5,38 %
6	1,81 %
13	1,74 %
18	1,16 %
12	0,93 %
8	0,75 %
2	0,46 %
3	0,30 %
7	0,27 %
9	0,22 %
14	0,20 %
17	0,07 %

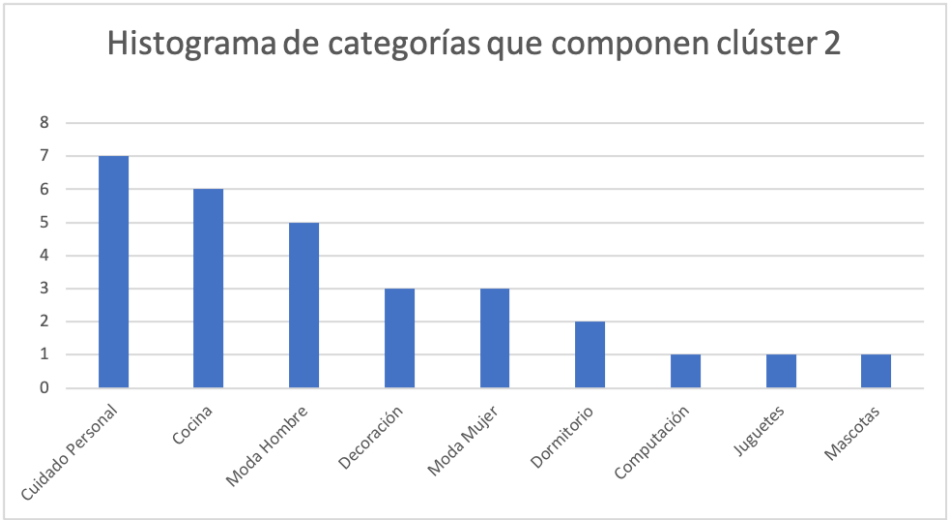
D.3. Anexo 9

D.3.1. Resultados de la clusterización

A continuación se presenta el comportamiento por posición para la totalidad de clúster generados, justo con su histograma de categorías y la distribución interna de KPI en los top de categorías.



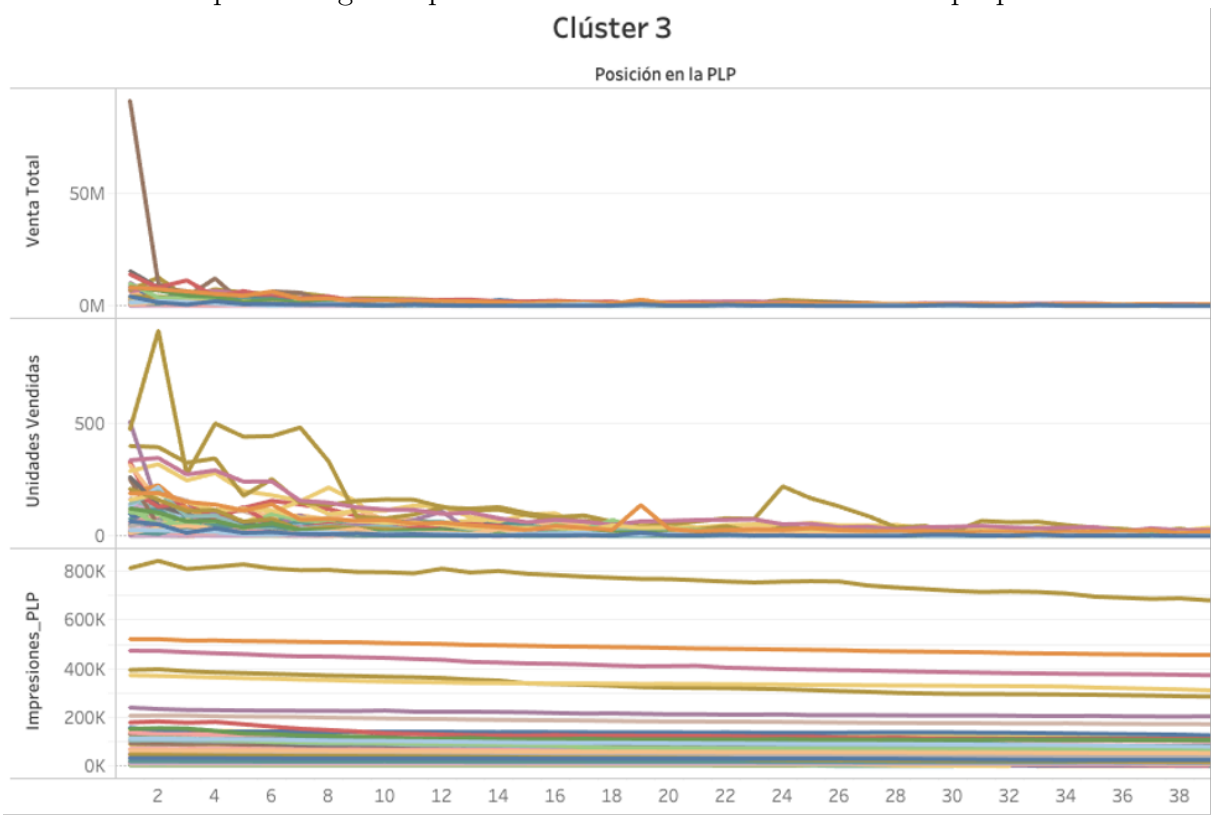
Comportamiento de KPI a través de las posiciones para el clúster 2. Fuente: Elaboración propia



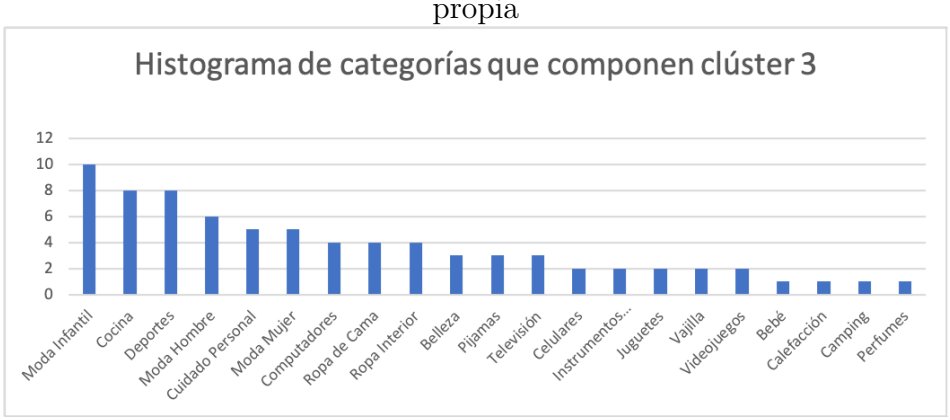
Histograma de categorías que componen el clúster 2. Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$	% CLÚSTER		CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
Afeitadoras	$ 33.005.221	27,15%		Afeitadoras	875	17,41%		Parkas Hombre	3.840.855	15,39%
Aspiradoras	$ 15.726.180	12,94%		Parkas Hombre	471	9,37%		Afeitadoras	3.535.810	14,17%
Parkas Hombre	$ 15.433.290	12,70%		Individuales	358	7,12%		Billeteras	1.817.006	7,28%
Ollas a Presión	$ 11.613.812	9,55%		Ollas a Presión	288	5,73%		Cuidado Del Rostro	1.347.757	5,40%
Billeteras	$ 3.409.890	2,79%		Aspiradoras	282	5,61%		Aspiradoras	1.183.194	4,74%

Top de categorias para el clúster 2. Fuente: Elaboración propia



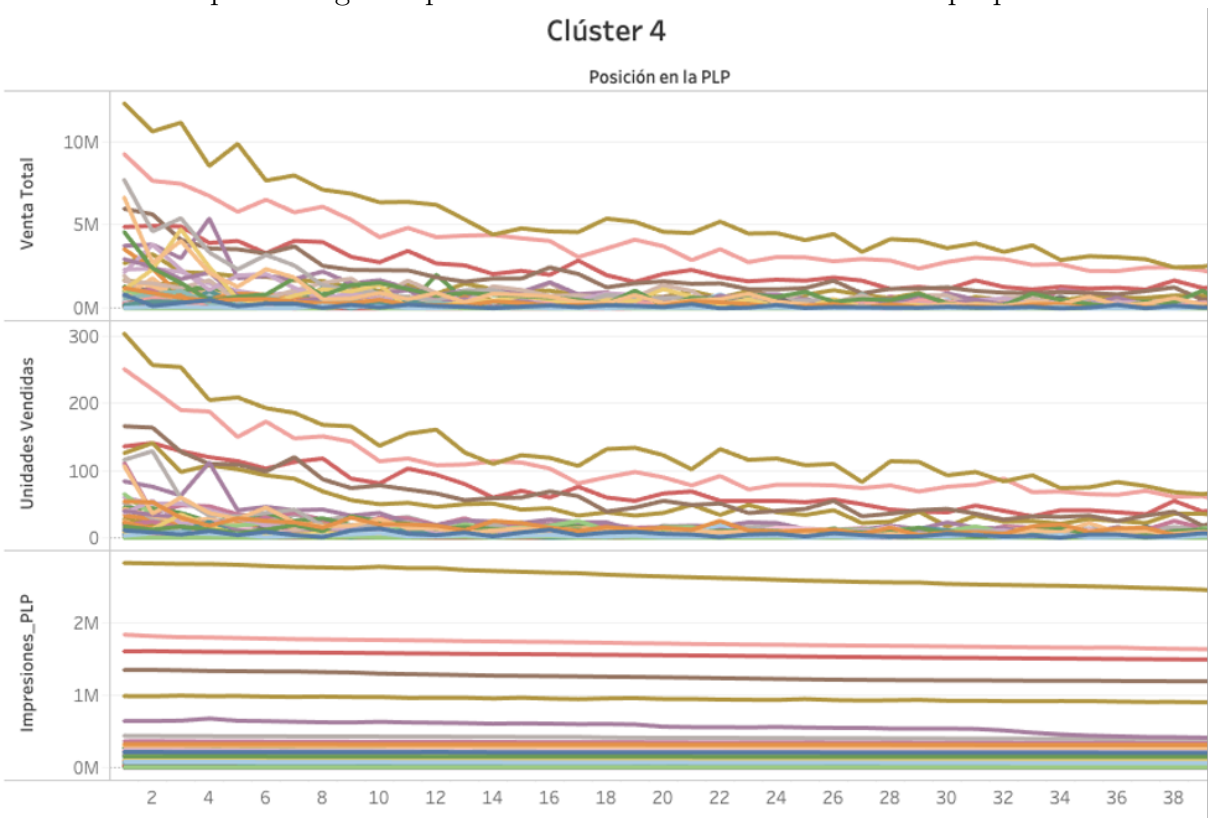
Comportamiento de KPI a través de las posiciones para el clúster 3. Fuente: Elaboración propia



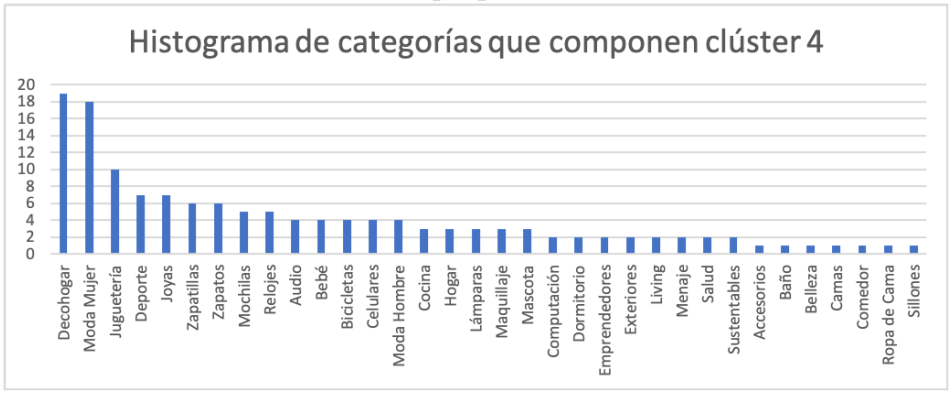
Histograma de categorías que componen el clúster 3. Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$	% CLÚSTER		CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
Nintendo	$ 145.020.780	9,56%		Jeans Mujer	6.744	11,14%		Jeans Mujer	35.877.542	13,43%
Jeans Mujer	$ 99.492.704	6,56%		Ropa de Cama	4.064	6,71%		Perfumes	23.033.534	8,63%
Sábanas	$ 98.417.728	6,49%		Jeans Hombre	3.922	6,48%		Sábanas	19.496.131	7,30%
Perfumes	$ 88.547.129	5,84%		Bóxers	3.730	6,16%		Cubrecamas	15.951.929	5,97%
Hornos Eléctricos	$ 88.168.820	5,81%		Perfumes	2.329	3,85%		Jeans Hombre	15.458.035	5,79%

Top de categorias para el clúster 3. Fuente: Elaboración propia



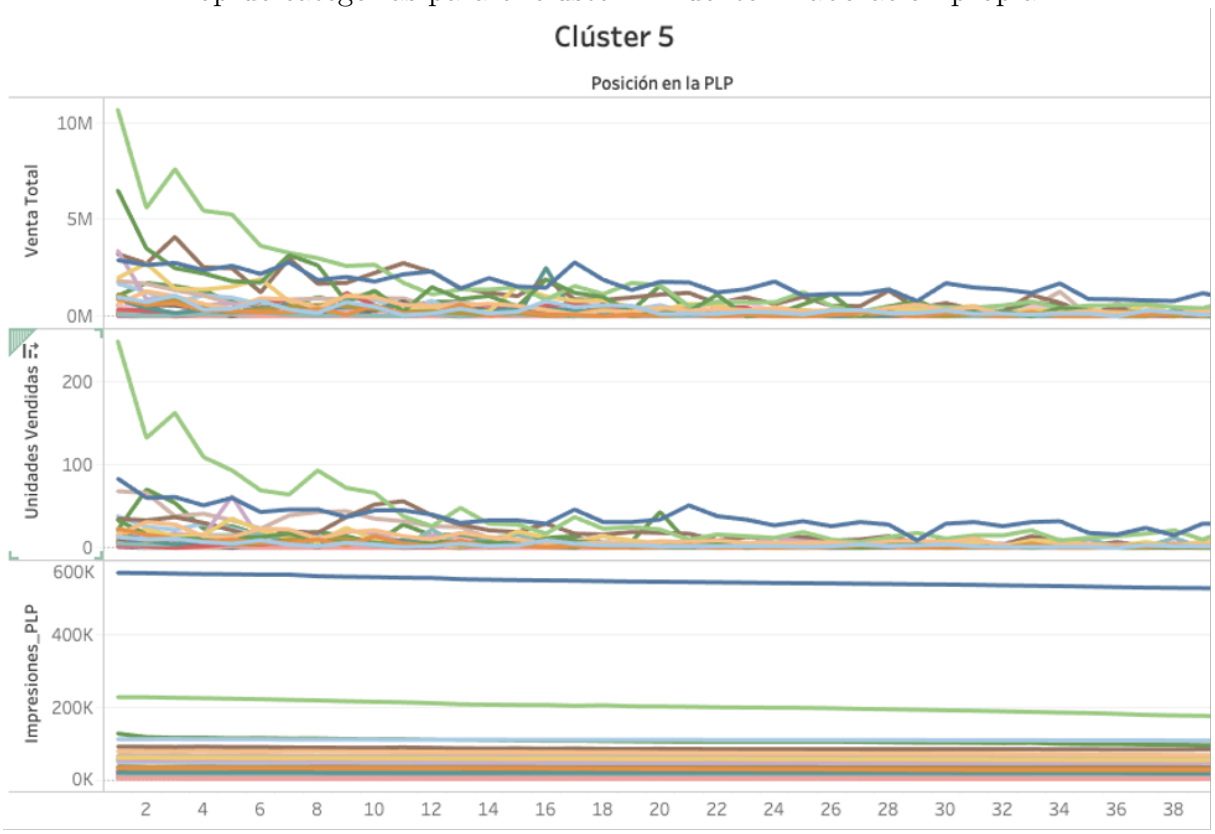
Comportamiento de KPI a través de las posiciones para el clúster 4. Fuente: Elaboración propia



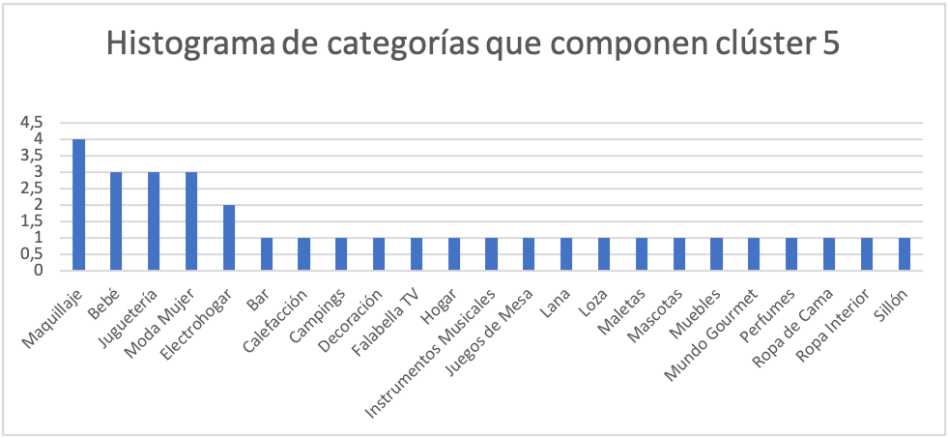
Histograma de categorías que componen el clúster 4. Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$	% CLÚSTER		CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
Zapatillas	$ 233.633.083	16,20%		Zapatillas	5.804	14,90%		Zapatillas	124.380.183	14,33%
Zapatillas Mujer	$ 174.868.694	12,10%		Zapatillas Mujer	4.639	11,90%		Zapatillas Mujer	81.489.046	9,39%
Zapatos Mujer	$ 102.630.258	7,10%		Zapatos Mujer	3.114	8,00%		Zapatos Mujer	73.688.693	8,49%
Botines	$ 83.779.020	5,80%		Botines	2.689	6,90%		Botines	59.583.887	6,87%
Abrigos y Chaquetas	$ 48.635.432	3,40%		Abrigos y Chaquetas	2.189	5,60%		Abrigos y Chaquetas	45.098.857	5,20%

Top de categorias para el clúster 4. Fuente: Elaboración propia



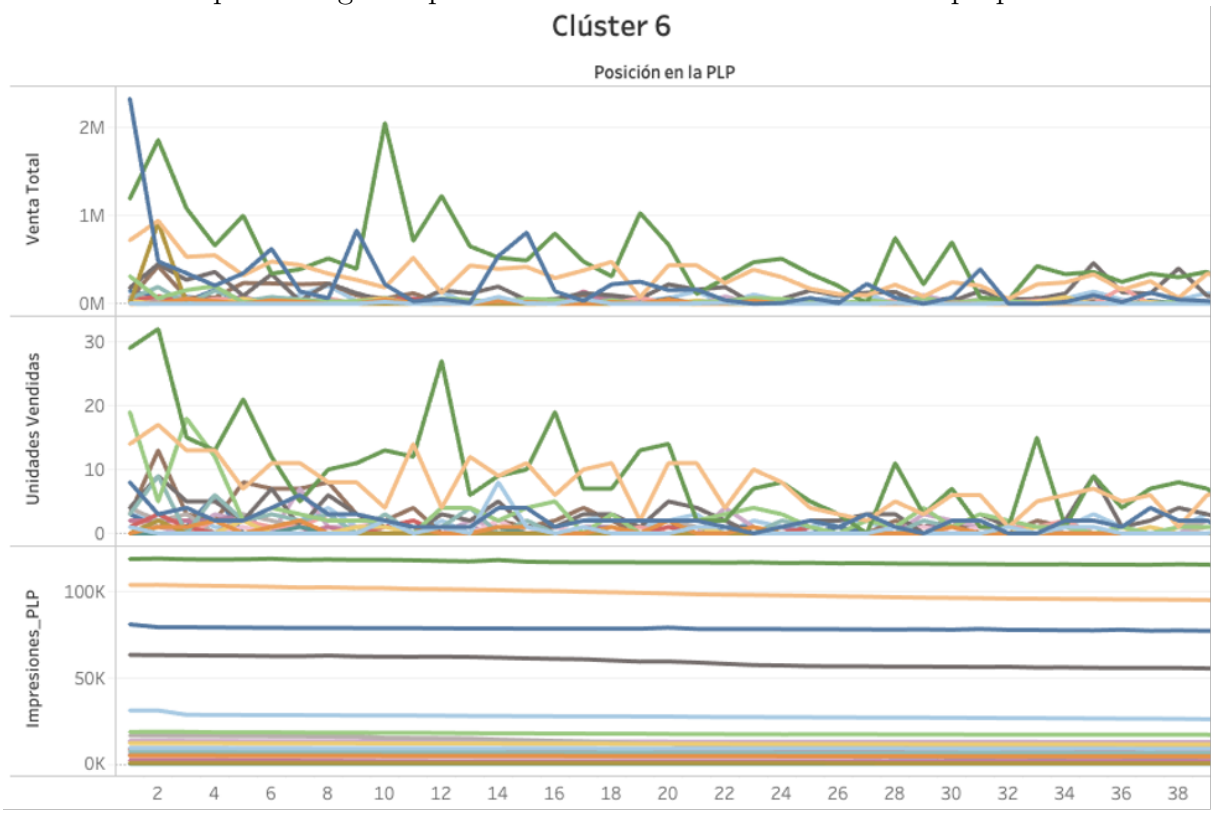
Comportamiento de KPI a través de las posiciones para el clúster 5. Fuente: Elaboración propia



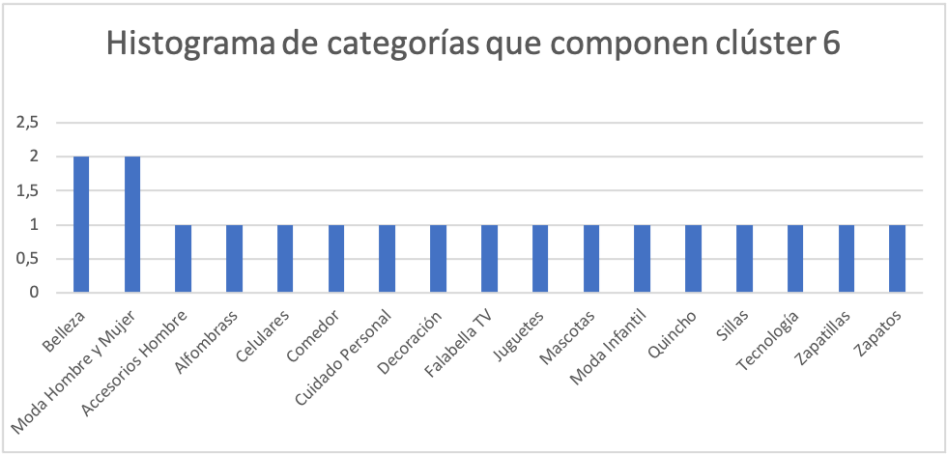
Histograma de categorías que componen el clúster 5. Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$	% CLÚSTER		CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
Perfumes Hombre	$ 69.862.704	16,54%		Perfumes Hombre	1.797	18,96%		Alfombras	259.892.646	30,20%
Alfombras	$ 76.795.880	15,91%		Alfombras	1.613	17,02%		Perfumes Hombre	9.402.429	10,86%
Aspiradoras	$ 55.039.351	11,40%		Aspiradoras	765	8,07%		Escritorio	5.252.837	6,07%
Sofá Cama y Futones	$ 42.388.460	8,78%		Scooter Skate y Patines	646	6,82%		Sofá Cama y Futones	5.014.988	5,79%
Sillas de bebé	$ 29.816.200	6,18%		Estufas Eléctricas	575	6,07%		Aspiradora	4.102.868	4,74%

Top de categorias para el clúster 5. Fuente: Elaboración propia



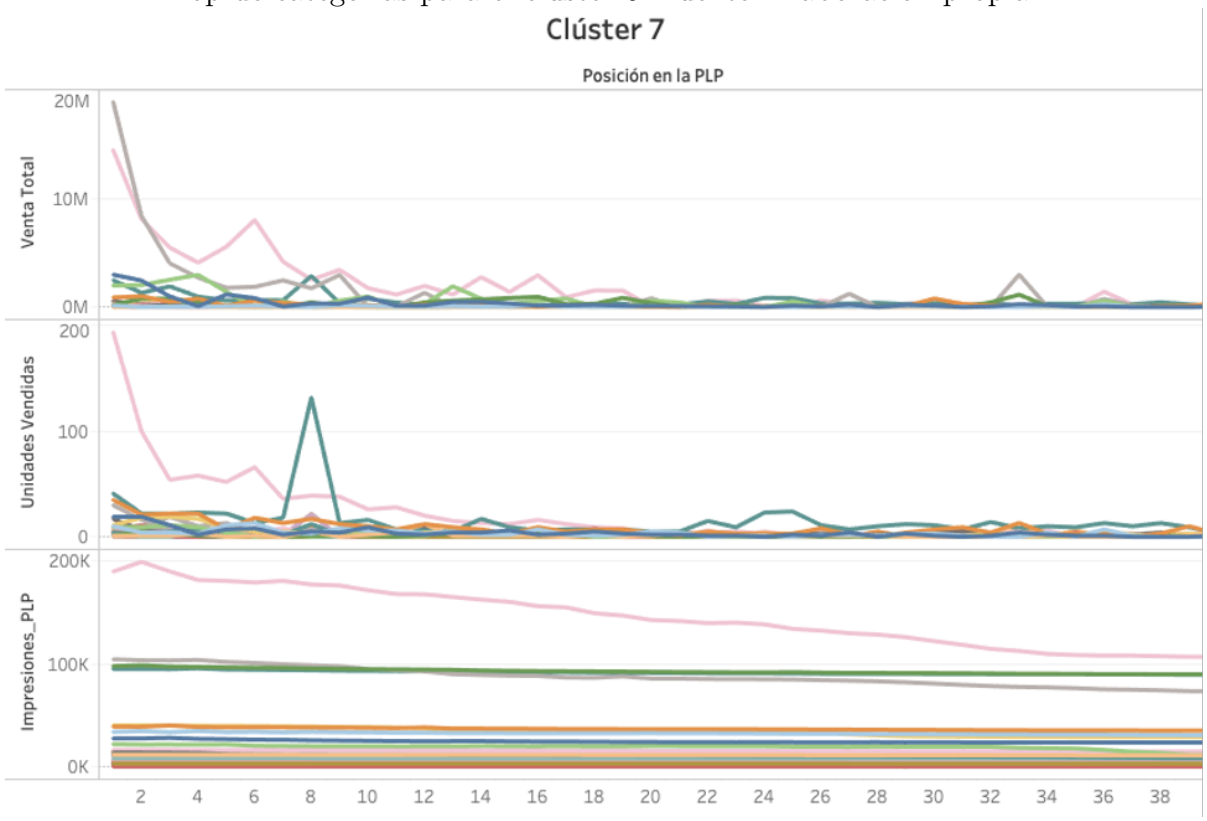
Comportamiento de KPI a través de las posiciones para el clúster 6. Fuente: Elaboración propia



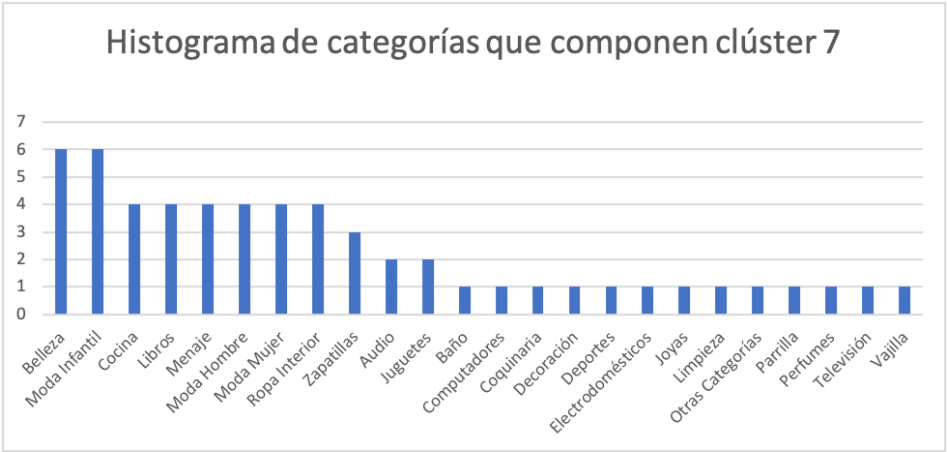
Histograma de categorías que componen el clúster 6. Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$	% CLÚSTER		CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
Sillas y Poltronas	$ 22.898.270	33,53%		Sillas y Poltronas	420	26,12%		Sillas y Poltronas	5.595.272	22,66%
Botines Hombre	$ 14.301.515	20,06%		Botines Hombre	339	21,08%		Botines Hombre	4.715.316	19,10%
Parrillas	$ 10.212.330	14,33%		Cinturones	141	8,77%		Parrillas	3.756.346	15,21%
Moda Hombre y Mujer	$ 6.488.220	9,10%		Moda Hombre y Mujer	128	7,96%		Moda Hombre y Mujer	2.818.537	11,42%
Moda Hombre Gap	$ 2.555.670	3,59%		Parrillas	99	6,16%		Zapatillas	1.324.089	5,36%

Top de categorias para el clúster 6. Fuente: Elaboración propia



Comportamiento de KPI a través de las posiciones para el clúster 7. Fuente: Elaboración propia

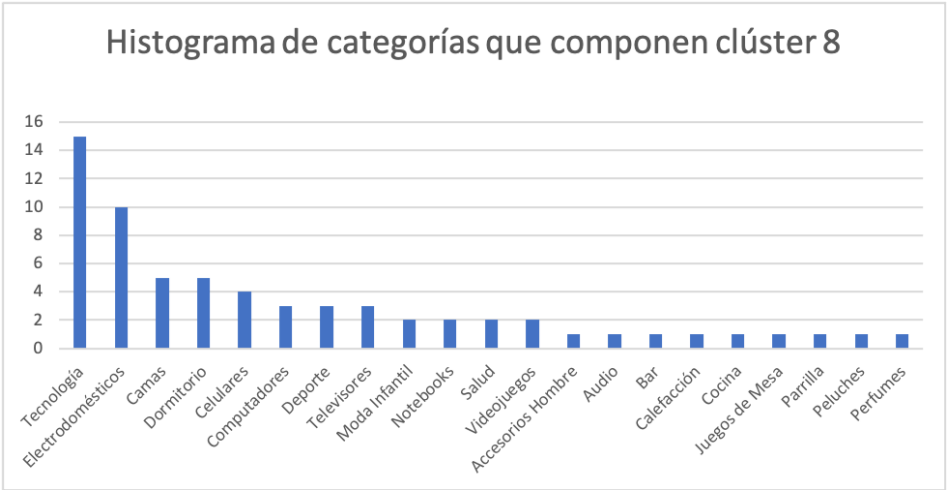


Histograma de categorías que componen el clúster 7 . Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$		% CLÚSTER	CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
Batidoras	$	79.744.447	31,05%	Batidoras	859	18,42%		Batidoras	6.691.193	16,89%
Computadores	$	55.893.730	21,76%	Audífonos Bluetooth	682	14,63%		Hi-Fi Audio	4.438.675	11,20%
Audífonos Bluetooth	$	24.431.080	9,51%	Cuchillería	370	7,93%		Audífonos Bluetooth	4.383.951	11,07%
Hornos Empotrables	$	19.902.850	7,75%	Tratamientos Faciales	178	3,82%		Computación Gamer	4.078.740	10,30%
Aspiradoras	$	14.486.011	5,64%	Aspiradoras	149	3,20%		Cuchillería	1.762.363	4,45%

Top de categorias para el clúster 7. Fuente: Elaboración propia

Comportamiento de KPI a través de las posiciones para el clúster 8. Fuente: Elaboración propia

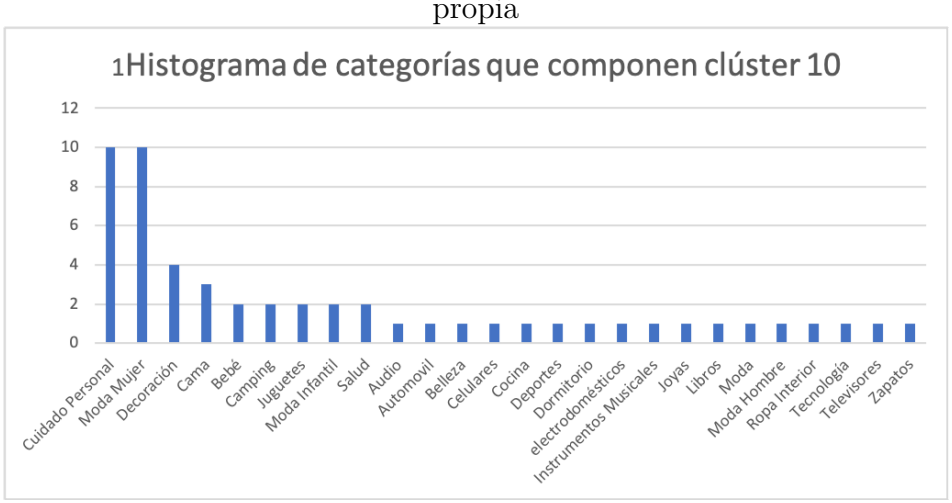


Histograma de categorías que componen el clúster 8. Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$	% CLÚSTER		CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
SmartWatch	$ 299.407.320	14,63%		SmartWatch	1.700	14,12%		SmartWatch	9.326.847	13,17%
PlayStation	$ 183.523.581	8,97%		Secadoras	967	8,03%		PlayStation	7.465.541	10,54%
Secadoras	$ 180.568.741	8,82%		PlayStation	909	7,55%		Impresoras	6.045.086	8,53%
Aspiradoras Robot	$ 139.098.050	6,80%		Impresoras	832	6,91%		Video Juegos	5.404.766	7,63%
Spinning	$ 122.788.310	6,00%		Aspiradoras Robot	750	6,23%		Cómodas y Cajoneras	4.367.340	6,17%

Top de categorias para el clúster 38. Fuente: Elaboración propia

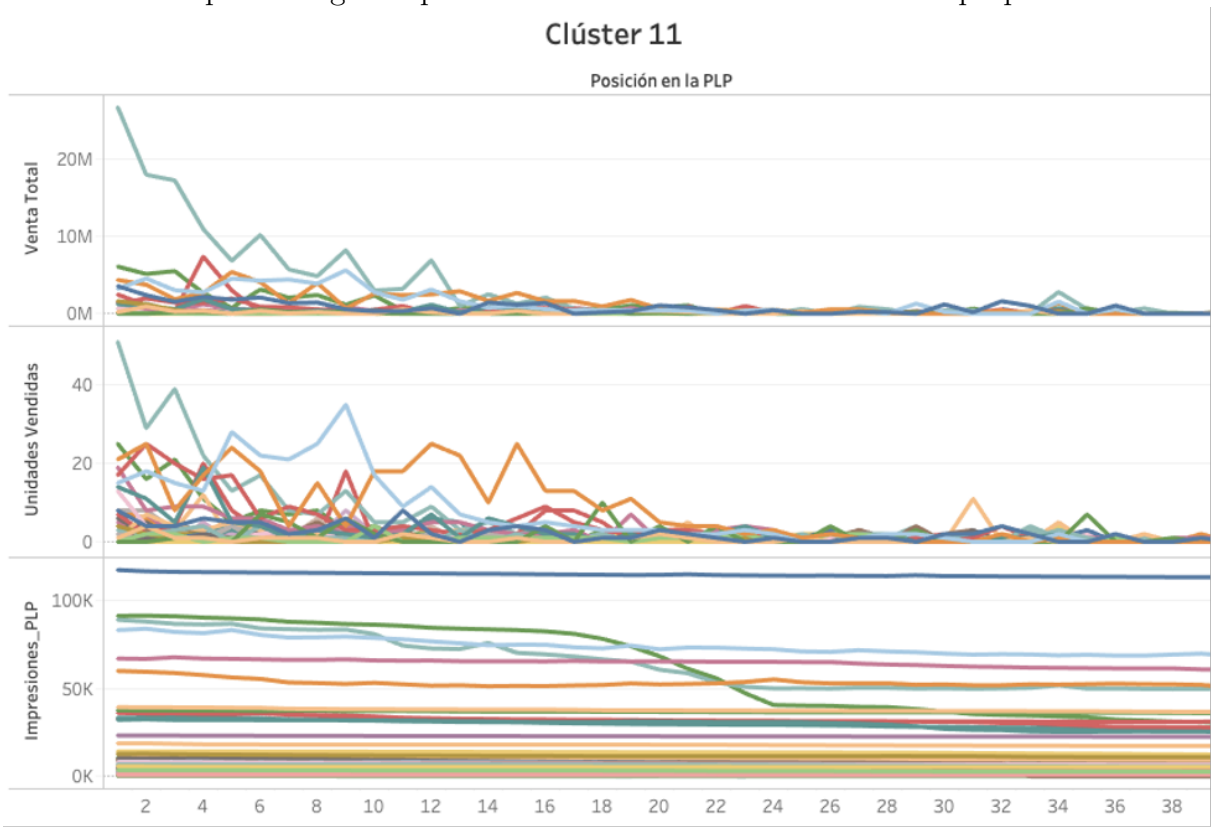
Comportamiento de KPI a través de las posiciones para el clúster 10. Fuente: Elaboración propia



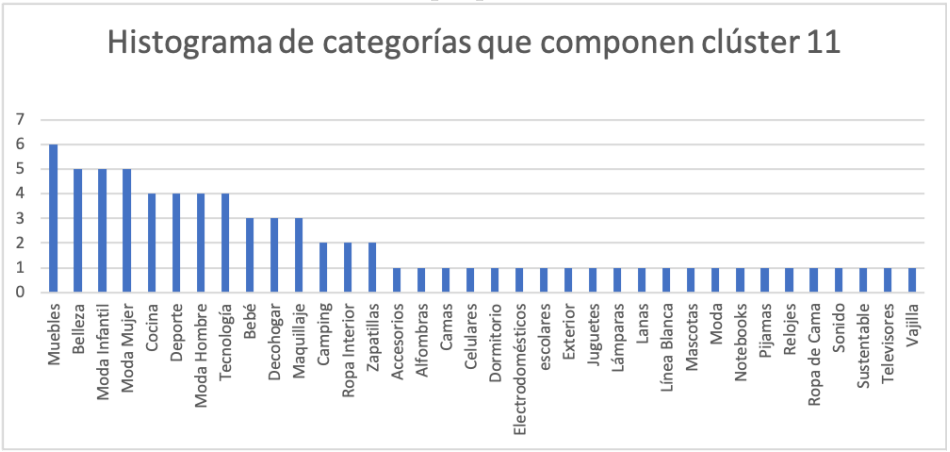
Histograma de categorías que componen el clúster 10. Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$	% CLÚSTER		CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
LED sobre 55 pulgadas	$ 150.367.260	60,38%		LED sobre 55 pulgadas	274	15,00%		LED sobre 55 pulgadas	1.838.523	16,35%
Mundo DrimKip	$ 30.372.880	12,20%		Afeitadoras	168	9,20%		Afeitadoras	567.953	5,05%
Sillas de Auto	$ 8.507.370	3,42%		Mundo DrimKip	112	6,13%		Parlantes y Subwoofer	403.659	3,59%
Camas	$ 7.574.690	3,04%		Triciclos	80	4,38%		Sillas de Auto	391.109	3,48%
Afeitadoras	$ 6.546.320	2,63%		Sillas de auto	77	4,21%		Colchones 1,5 Plaza	359.751	3,20%

Top de categorias para el clúster 10. Fuente: Elaboración propia



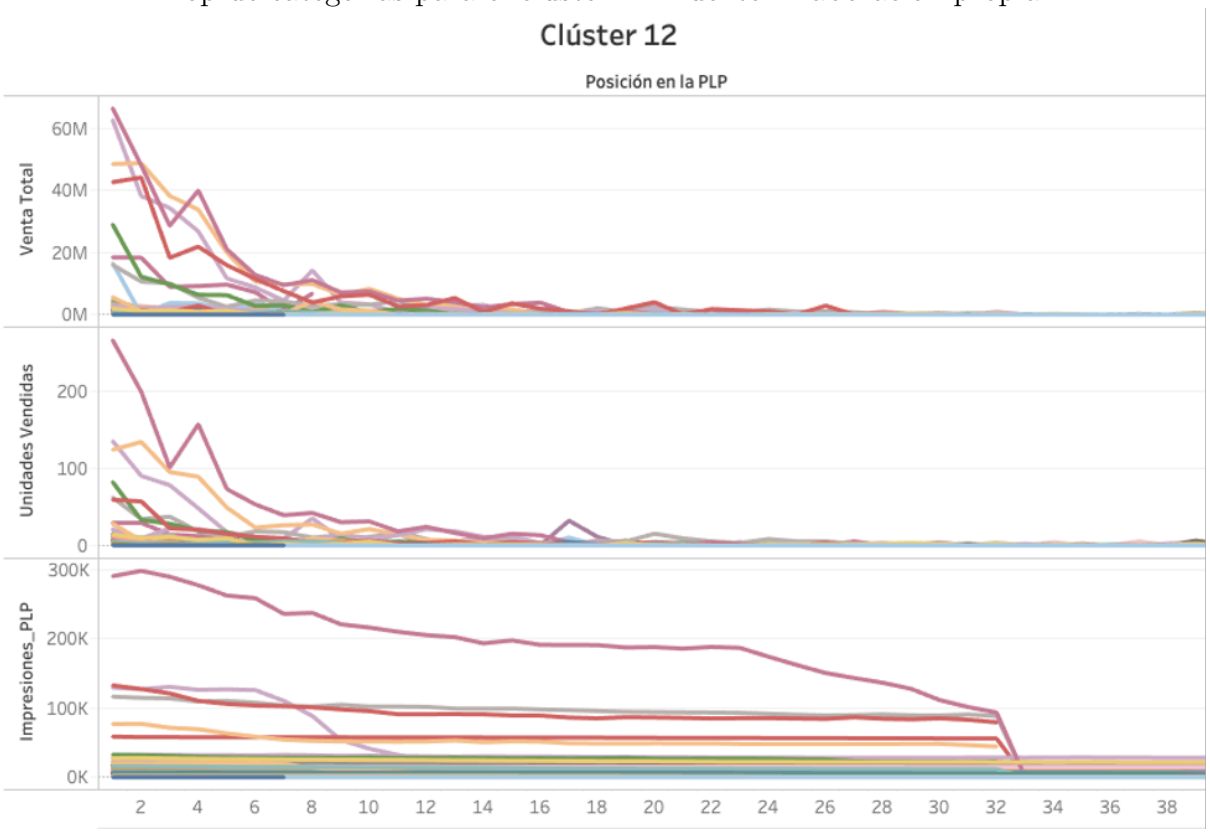
Comportamiento de KPI a través de las posiciones para el clúster 11. Fuente: Elaboración propia



Histograma de categorías que componen el clúster 11. Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$	% CLÚSTER		CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
Notebooks HP	$ 139.341.480	29,94%		SoundBar y Home Theater	325	10,63%		Comedor	1.838.523	12,42%
SoundBar y Home Theater	$ 56.004.399	12,03%		Máquina de Cocer	290	9,49%		Máquina de cocer	567.953	7,94%
Todo Samsung	$ 54.933.002	11,80%		Notebooks HP	252	8,24%		Muebles de cocina	403.659	6,93%
Comedor	$ 41.483.510	8,91%		Todo Samsung	139	6,12%		Computadores HP	391.109	6,69%
Drones	$ 34.158.070	7,34%		Autos a Batería	111	4,55%		SoundBar y Home Theater	359.751	6,09%

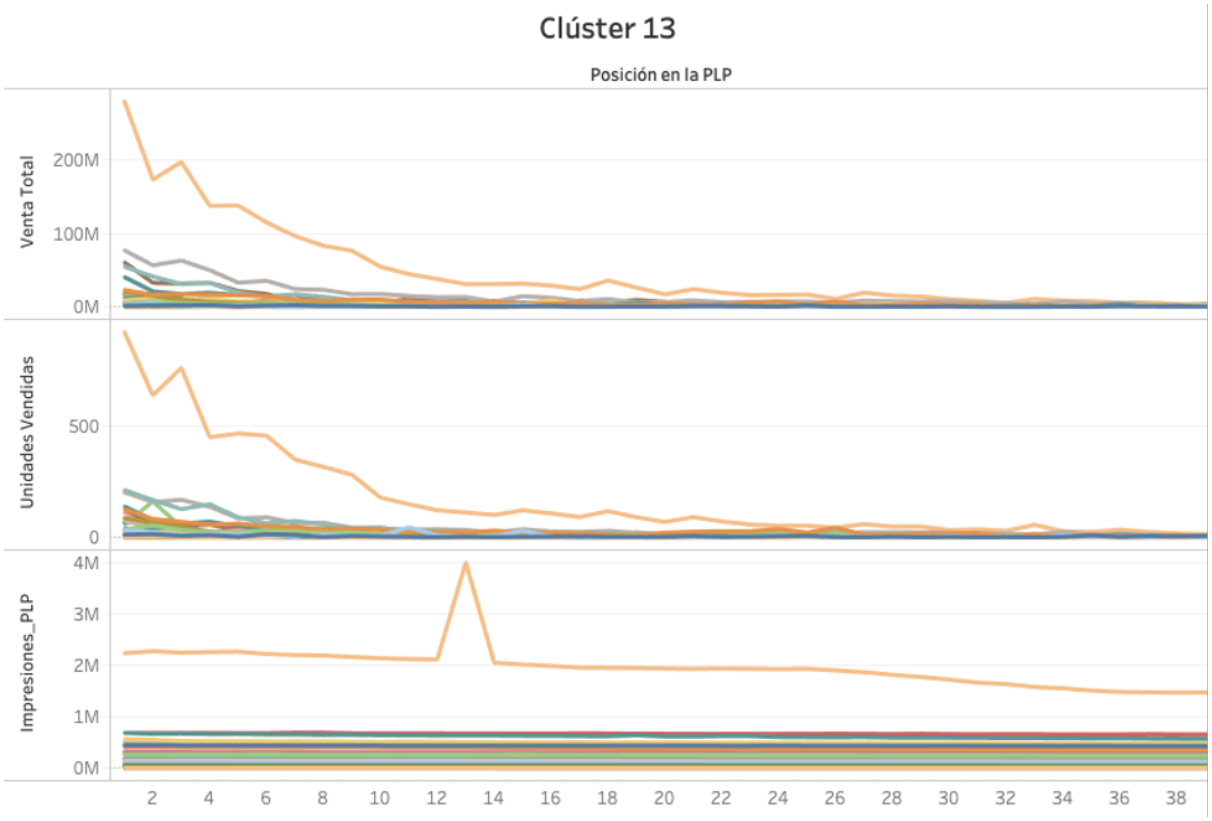
Top de categorias para el clúster 11. Fuente: Elaboración propia



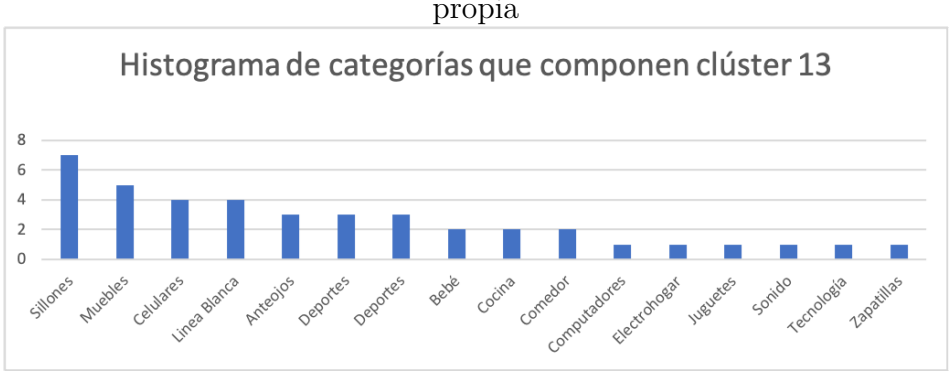
Comportamiento de KPI a través de las posiciones para el clúster 12. Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$	% CLÚSTER		CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
Cocina a Gas	$ 283.183.755	16,50%		Cocina a Gas	1.110	11,90%		Comedor	1.838.523	12,42%
Lavadoras y Secadoras	$ 254.022.550	14,80%		Lavadoras y Secadoras	645	6,90%		Máquina de cocer	567.953	7,94%
Trotadoras	$ 217.943.101	12,70%		Trotadoras	459	4,90%		Muebles de cocina	403.659	6,93%
Notebooks Gamer	$ 210.147.640	12,20%		Smart TV	335	3,60%		Computadores HP	391.109	6,69%
Smart TV	$ 87.796.650	5,10%		Teclados y Pianos	237	2,55%		SoundBar y Home Theater	359.751	6,09%

Top de categorias para el clúster 12. Fuente: Elaboración propia



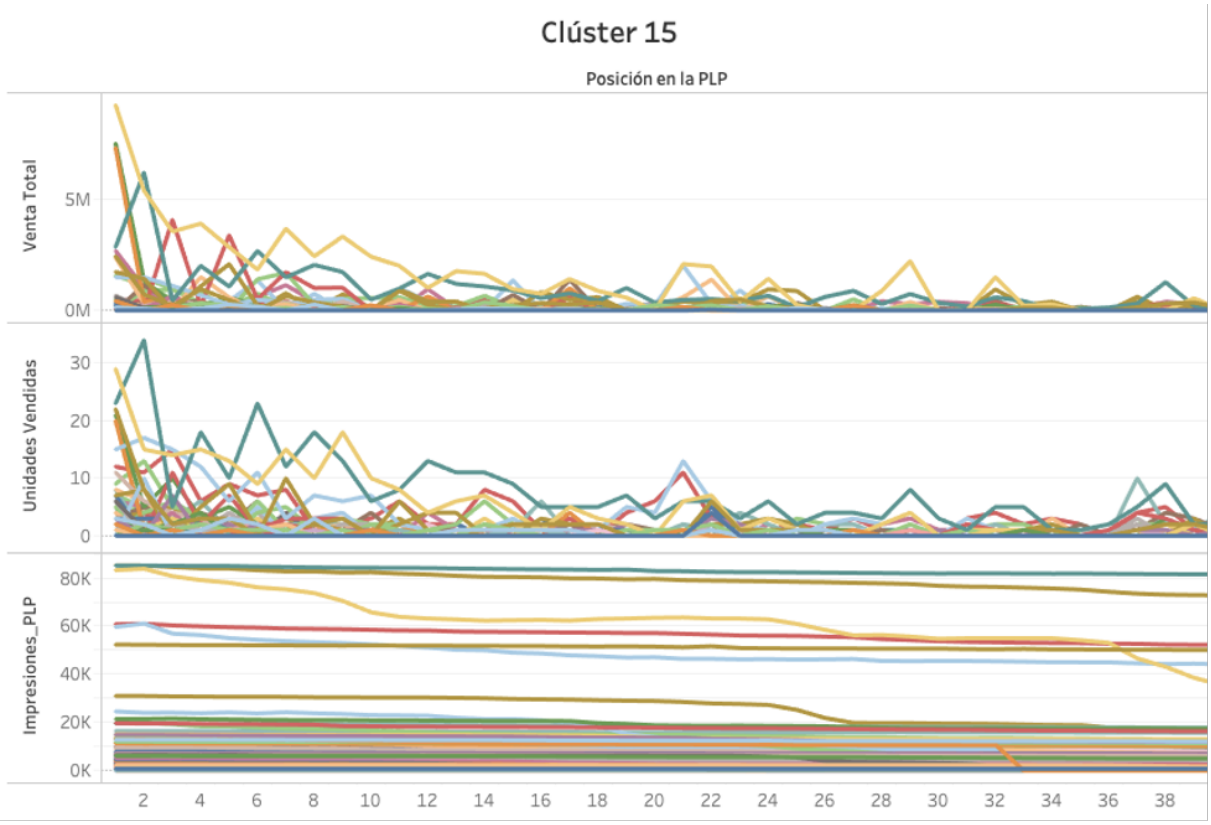
Comportamiento de KPI a través de las posiciones para el clúster 13. Fuente: Elaboración propia



Histograma de categorías que componen el clúster 13. Fuente: Elaboración propia

TOP 5 EN VENTAS				TOP 5 EN UNIDADES VENDIDAS				TOP 5 EN VISTAS		
CATEGORÍA	VENTA TOTAL	% CLÚSTER		CATEGORÍA	UNIDADES VENDIDAS	% CLÚSTER		CATEGORÍA	CANTIDAD	% CLÚSTER
Smartphone	$ 1.918.287.912	33,97%		Smartphone	6.867	30,89%		Smartphone	90.243.211	20,13%
Refrigeradores	$ 653.388.559	11,57%		Refrigeradores	1.778	8,00%		Sofás y sillones	32.229.718	7,19%
Lavadora	$ 366.613.640	6,49%		Lavadora	1.500	6,75%		Línea Blanca	29.355.582	6,55%
Computadores	$ 353.192.700	6,25%		Celulares y teléfonos	1.224	5,51%		Comedor	23.952.252	5,34%
Electrohogar	$ 289.127.749	5,12%		Línea Blanca	948	4,26%		Refrigeradores	21.151.448	4,72%

Top de categorias para el clúster 13. Fuente: Elaboración propia



Comportamiento de KPI a través de las posiciones para el clúster 15. Fuente: Elaboración propia

TOP 5 VENTAS				TOP 5 UNIDADES VENDIDAS				TOP 5 VIEWS		
CAT	$	% CLÚSTER		CAT	UNIDADES	% CLÚSTER		CAT	CANTIDAD	% CLÚSTER
Berges	$ 62.877.790	20,30%		Coches Bebé	339	11,60%		Coches	3.972.469	7,32%
Coches bebé	$ 40.497.060	13,10%		Parkas deportivas	221	7,60%		Terrazas	3.741.875	6,90%
Terrazas	$ 16.417.470	5,30%		Berges	179	6,10%		Parkas Deportivas	2.668.802	4,92%
Drones	$ 12.887.388	4,20%		Xbox	166	5,70%		Berges	2.653.056	4,89%
Xbox	$ 11.177.470	3,60%		Bar	91	3,10%		Bar	2.449.074	4,71%

Top de categorias para el clúster 15. Fuente: Elaboración propia

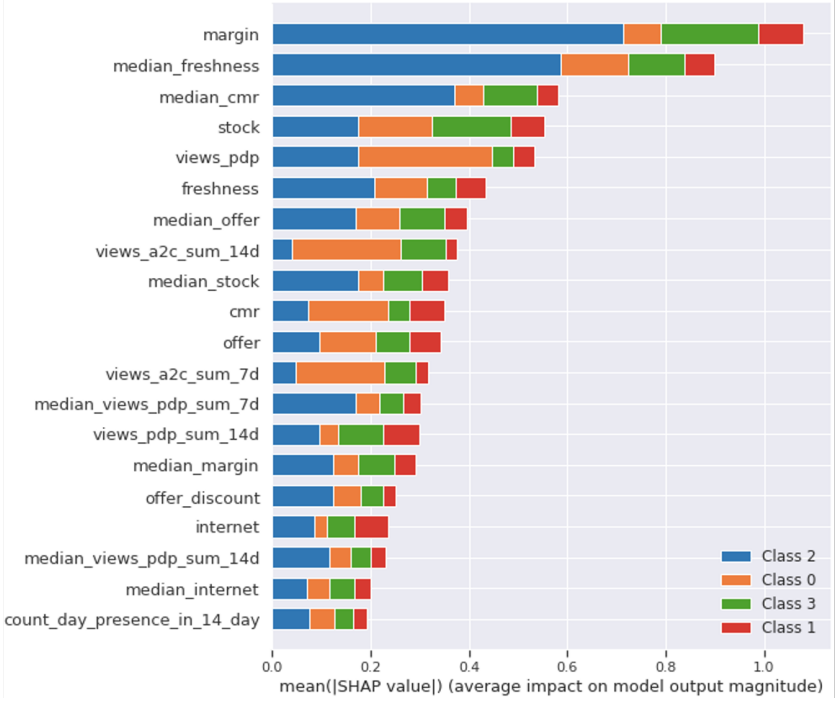
Anexo E

Resultados

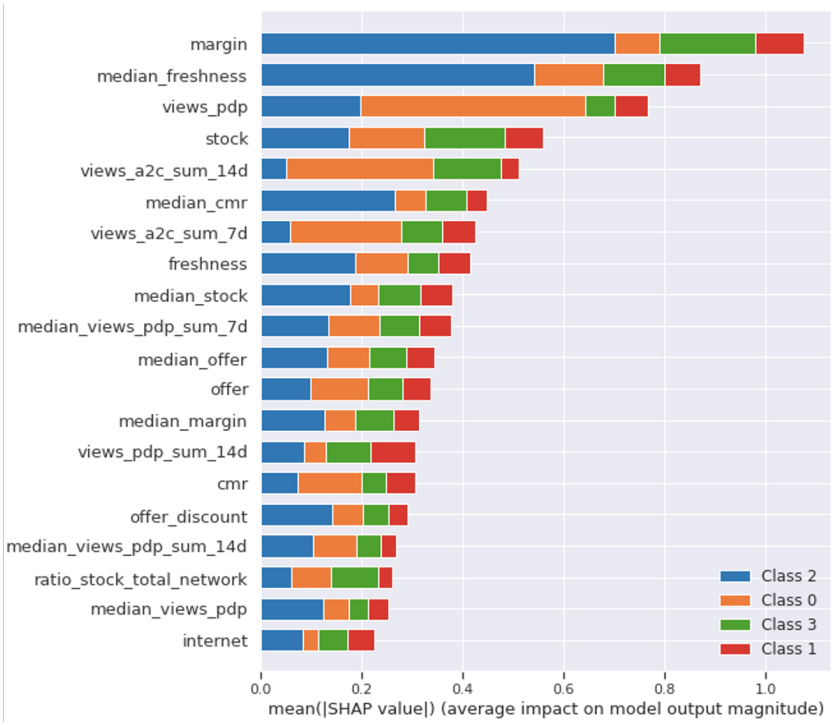
E.1. Anexo 10

E.1.1. Resultados por clúster

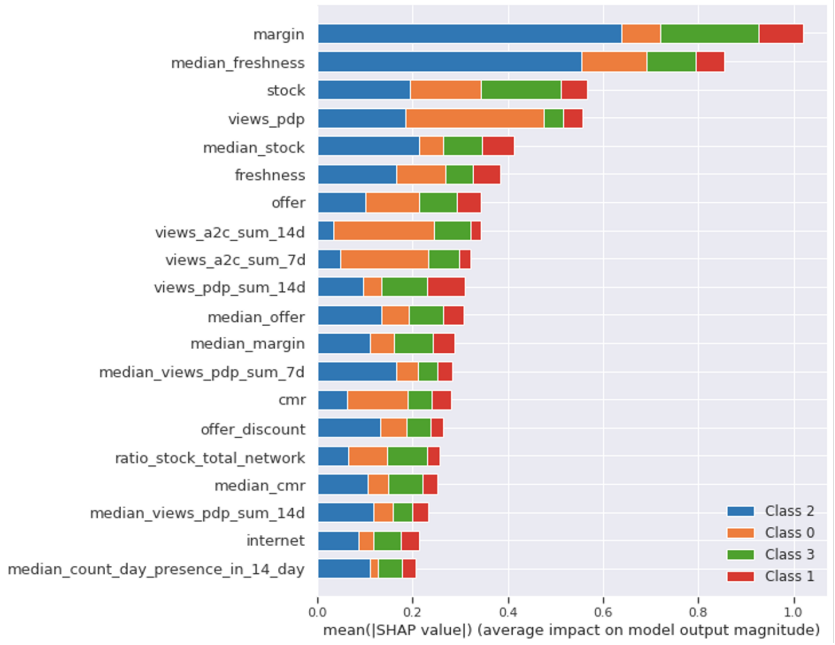
A continuación se muestran el análisis de Feature Importance para cada modelo.



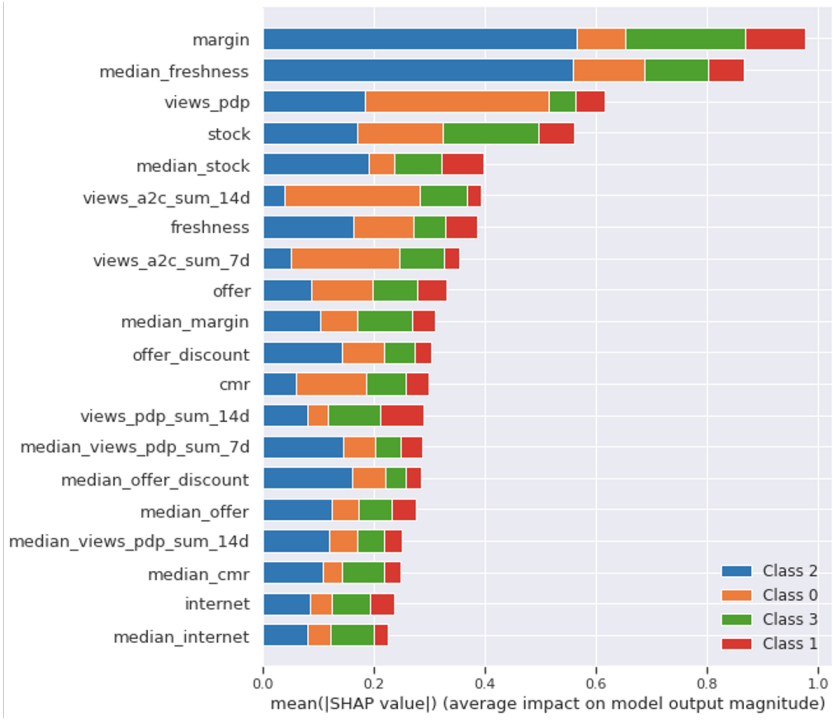
Feature Importance para clúster 2. Fuente: Elaboración propia



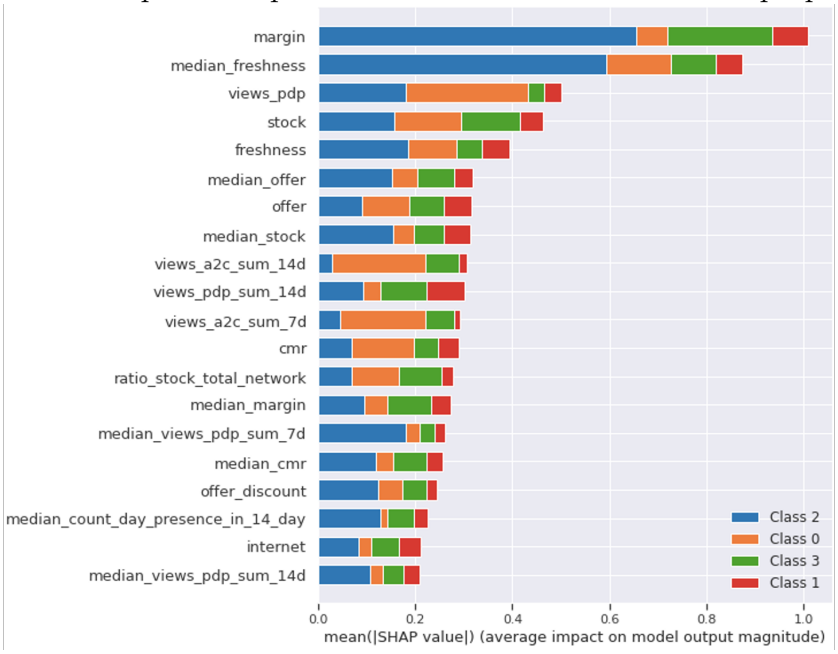
Feature Importance para clúster 3. Fuente: Elaboración propia



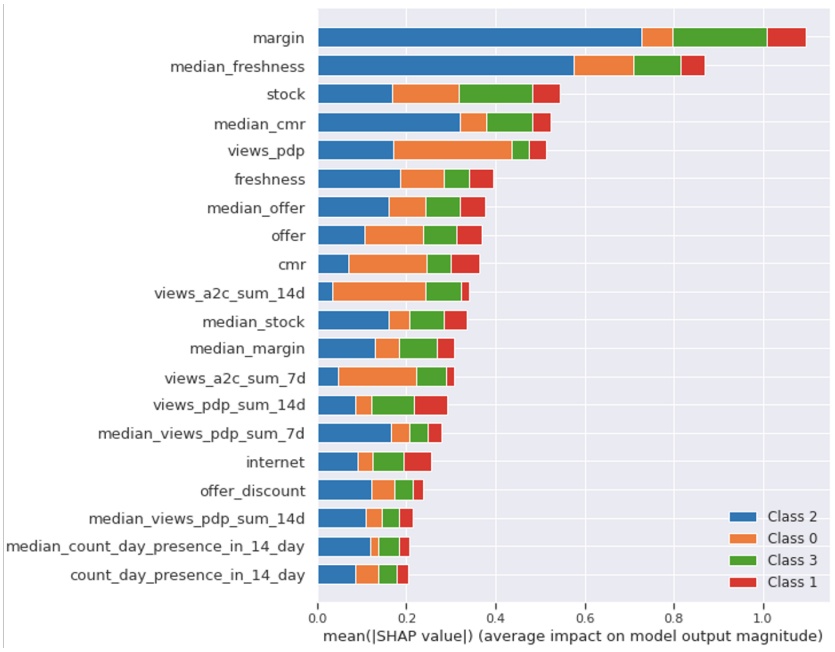
Feature Importance para clúster 4. Fuente: Elaboración propia



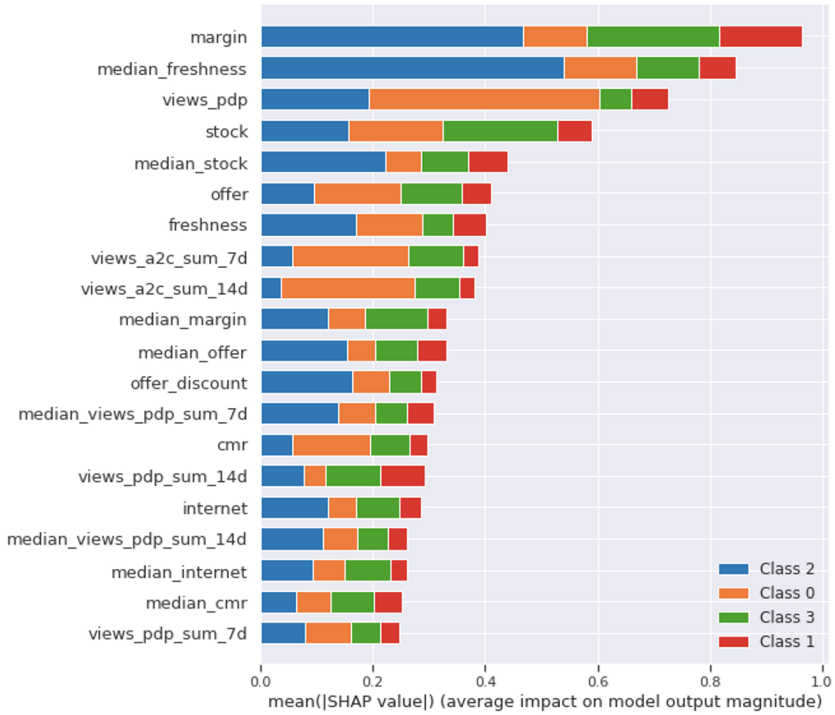
Feature Importance para clúster 5. Fuente: Elaboración propia



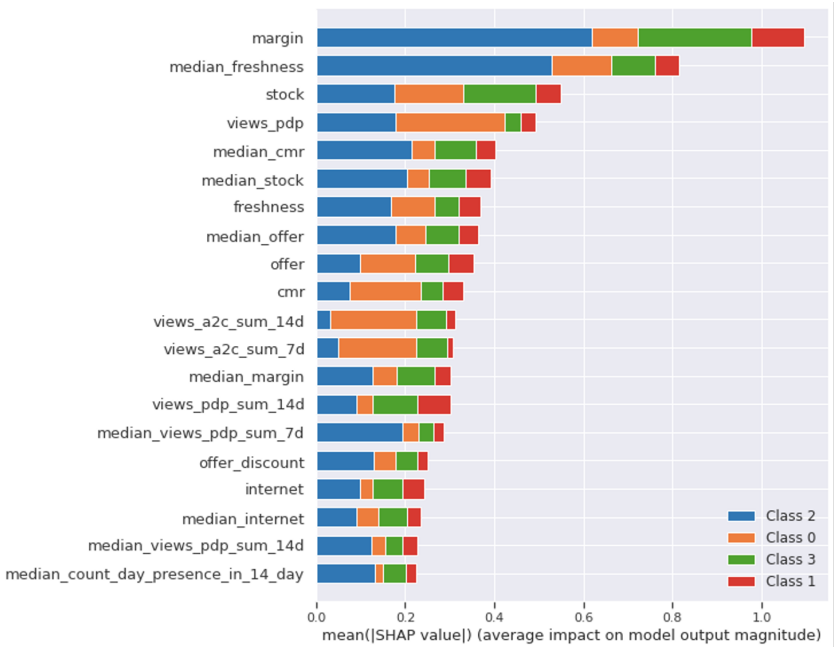
Feature Importance para clúster 6. Fuente: Elaboración propia



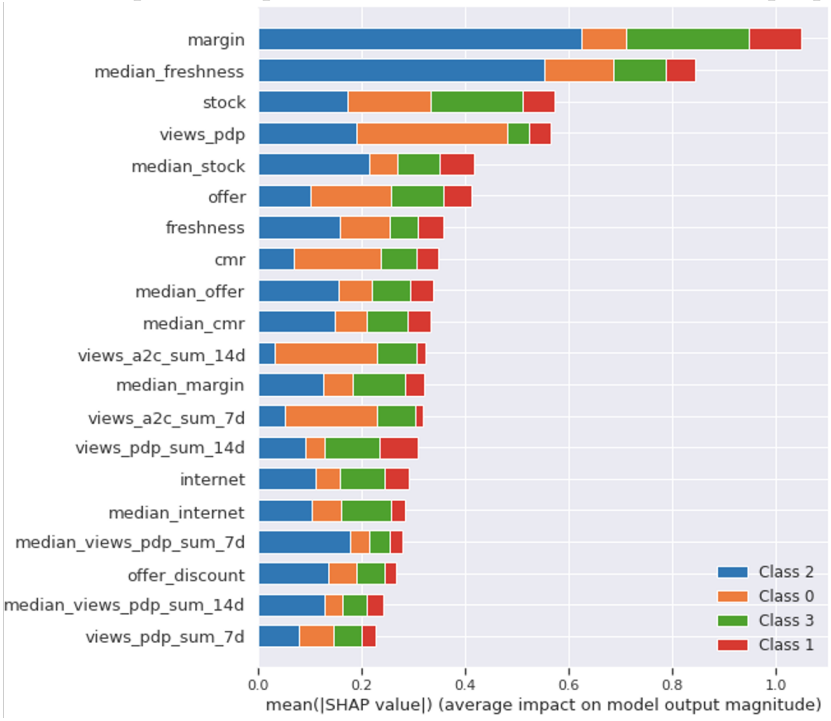
Feature Importance para clúster 7. Fuente: Elaboración propia



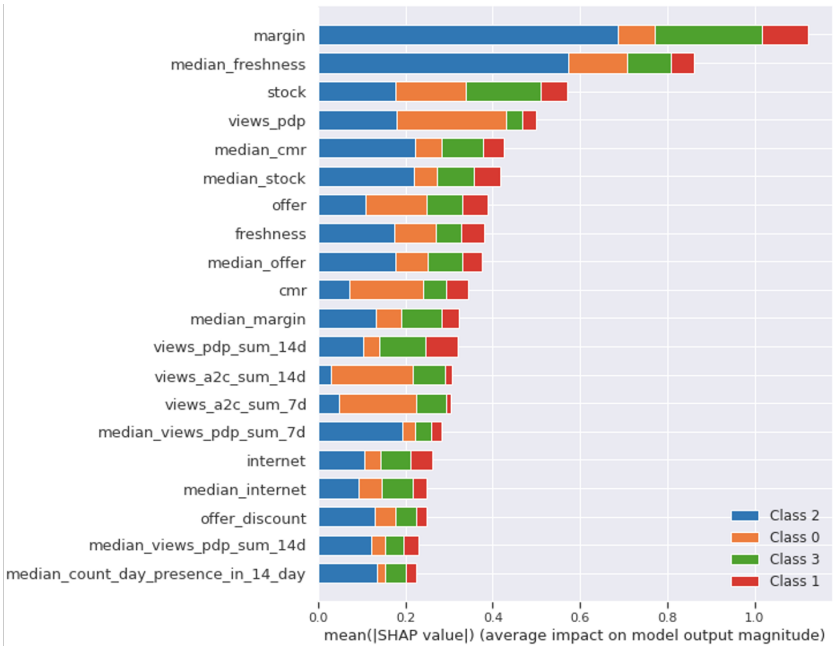
Feature Importance para clúster 8. Fuente: Elaboración propia



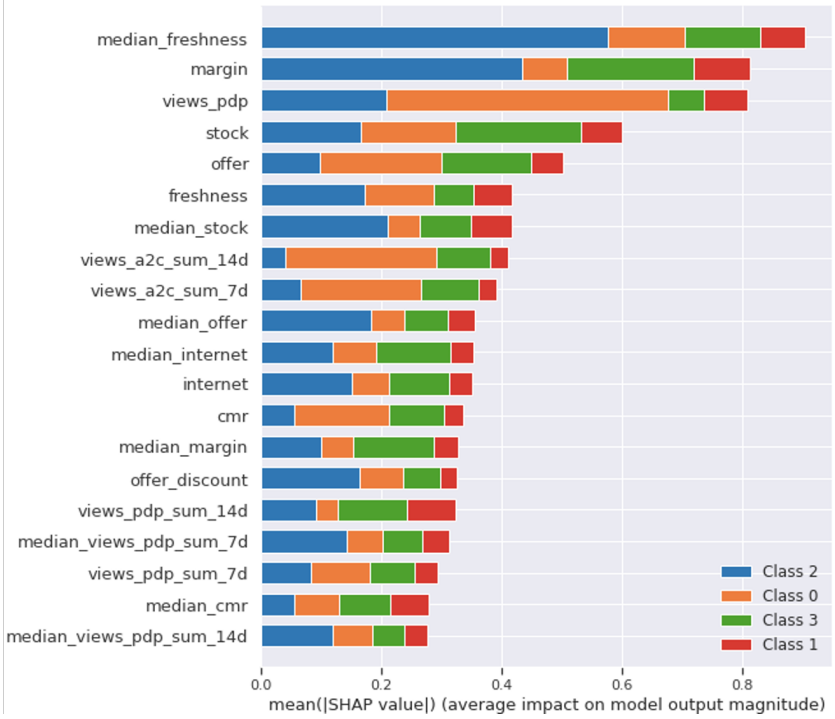
Feature Importance para clúster 10. Fuente: Elaboración propia



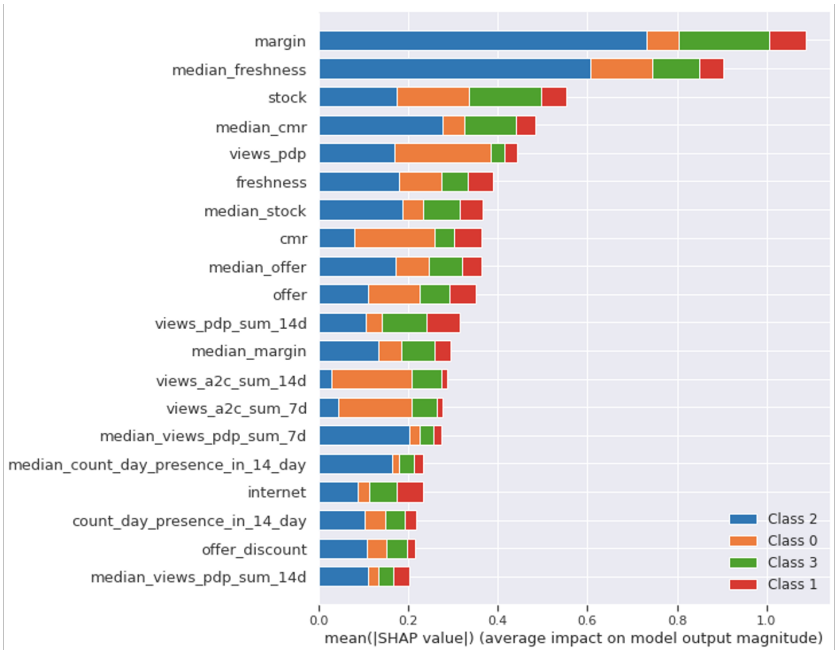
Feature Importance para clúster 11. Fuente: Elaboración propia



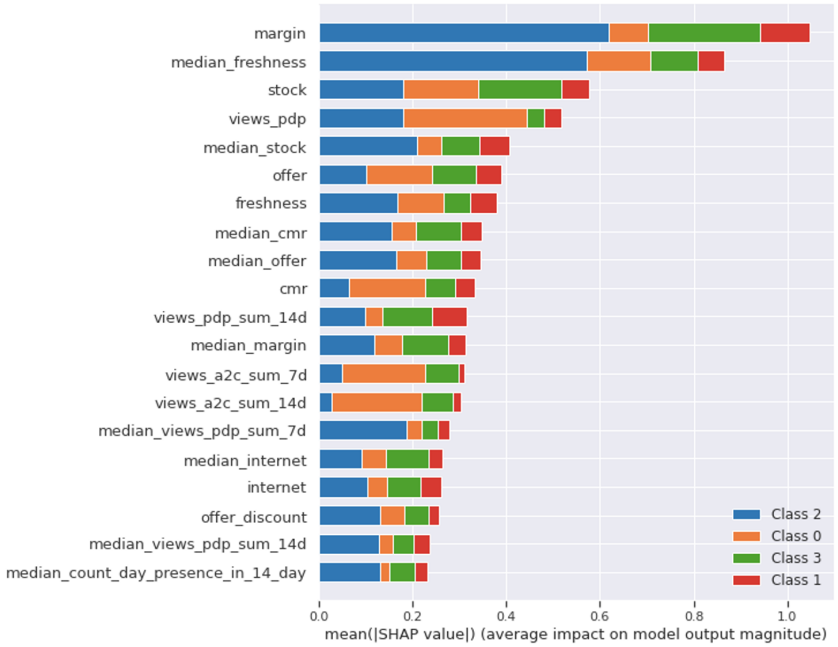
Feature Importance para clúster 12. Fuente: Elaboración propia



Feature Importance para clúster 13. Fuente: Elaboración propia



Feature Importance para clúster 14. Fuente: Elaboración propia



Feature Importance para clúster 15. Fuente: Elaboración propia

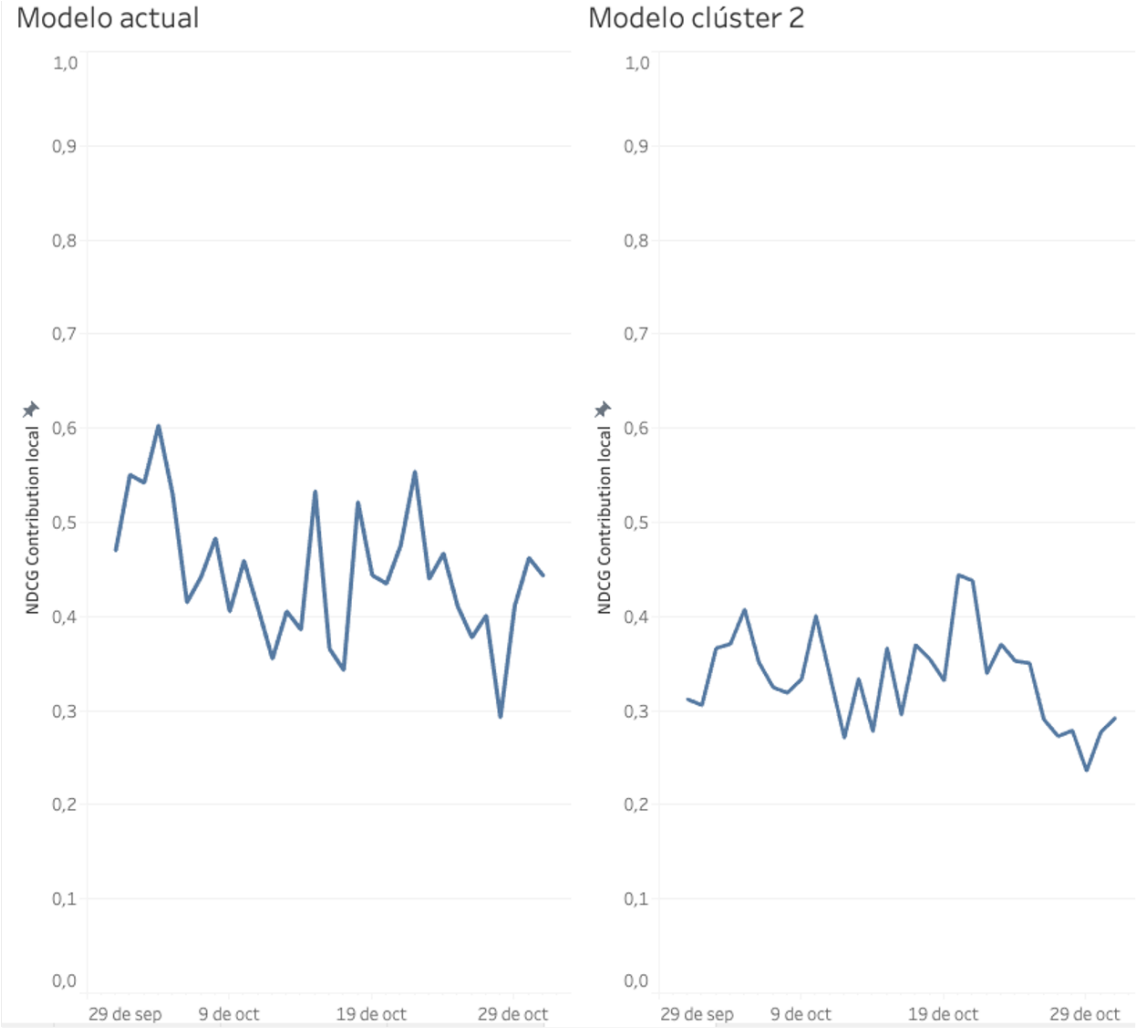
E.2. Anexo 11

E.2.1. Análisis de NDCG

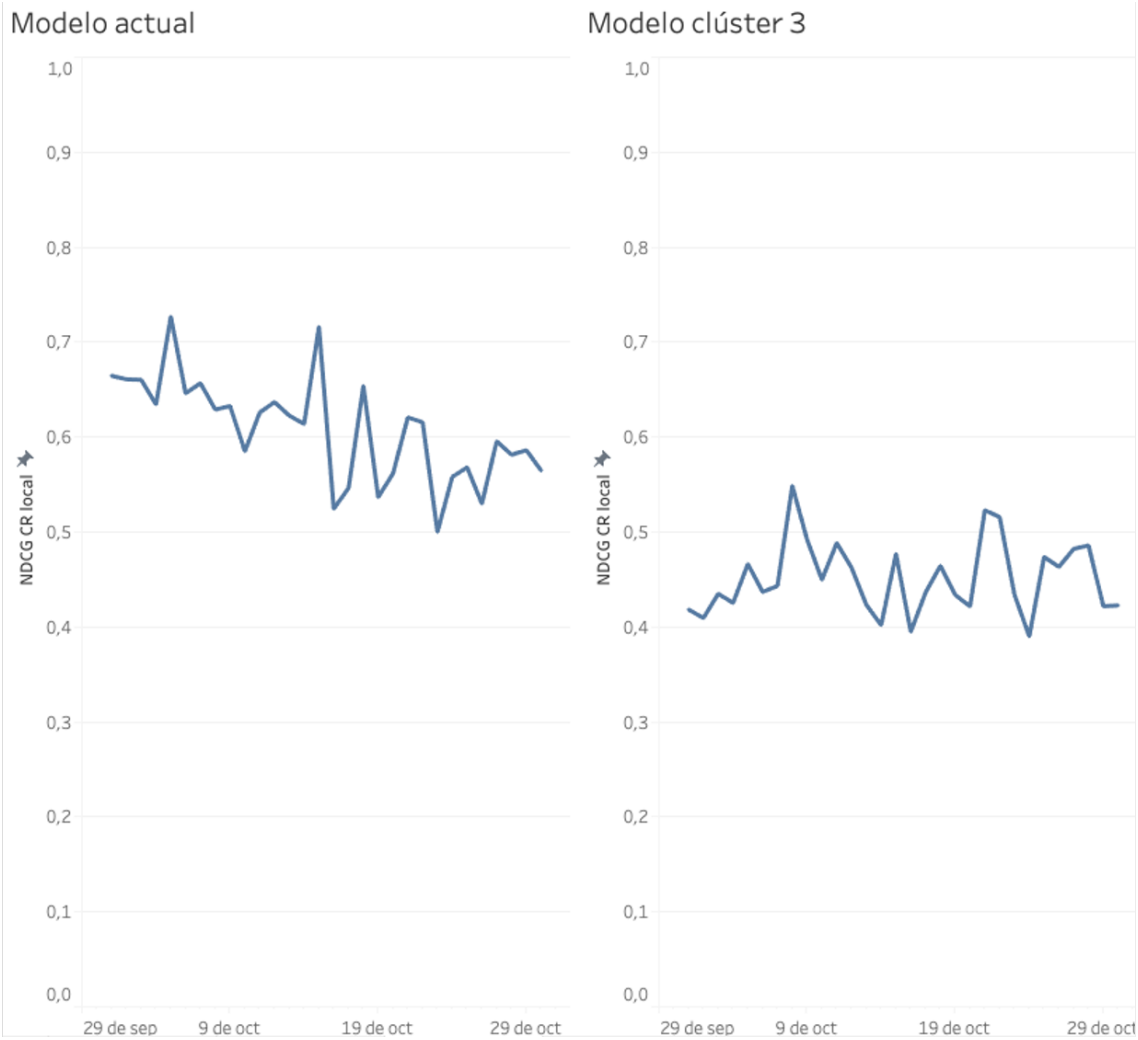
En la imagen E.1 se muestra la correlación que presenta el valor de NDCG contribución local con la contribución online, para todas las PLP activas en el sitio, en el periodo abril - octubre 2020

Figura E.1: Correlación para NDCG contribución local y KPI online. Fuente: Elaboración propia

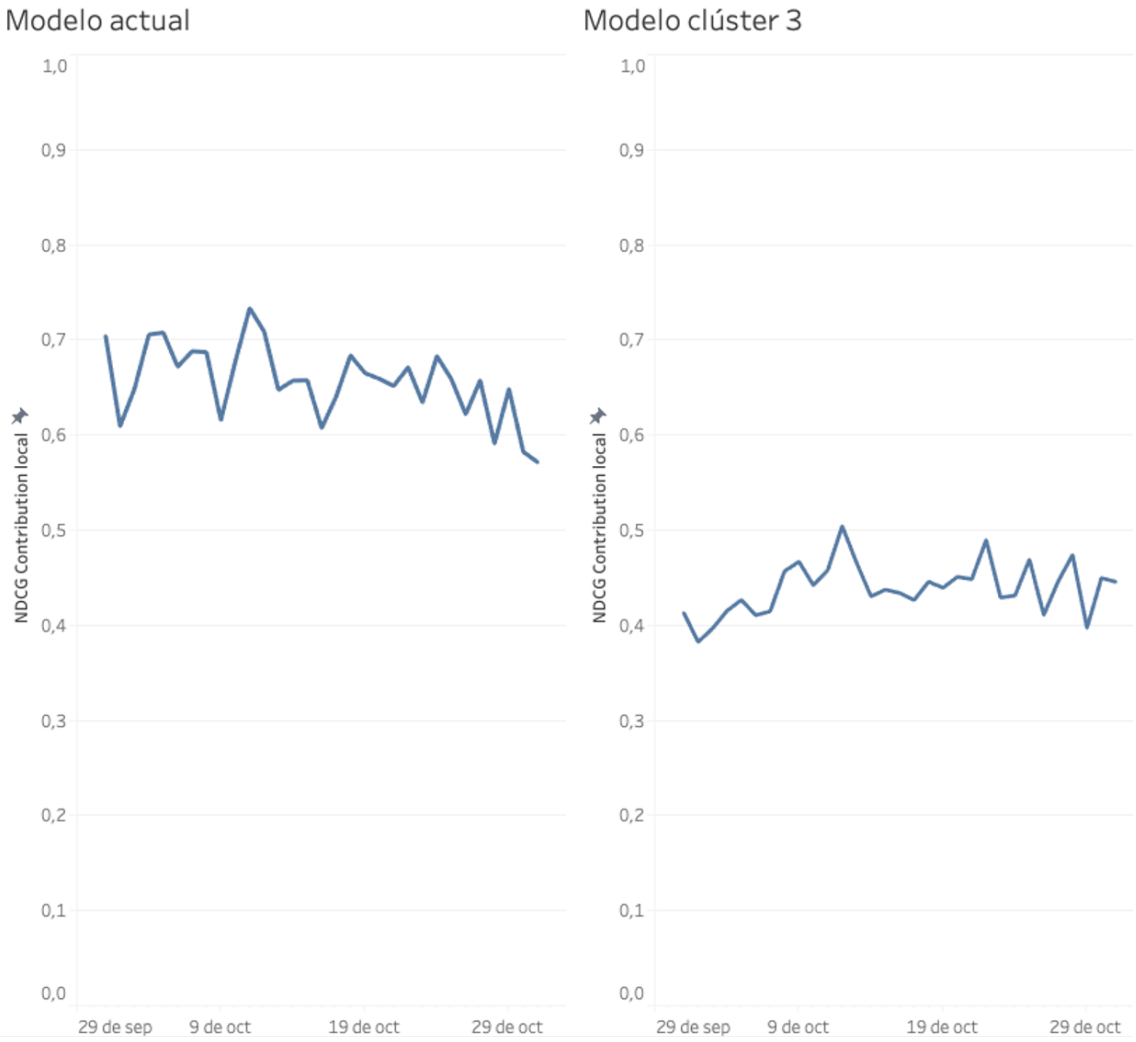
E.2.2. Comparación de NDCG entre modelos



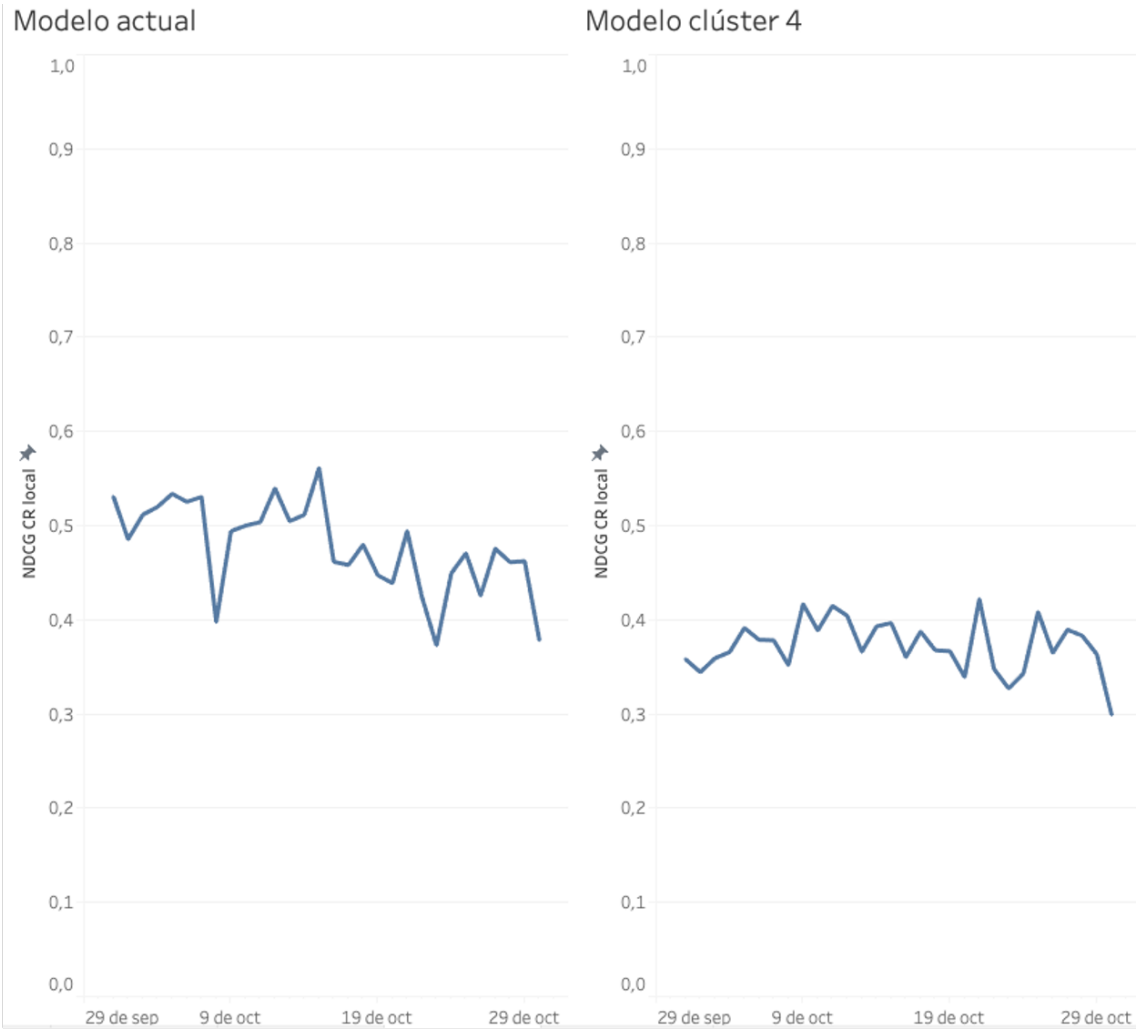
Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 2. Fuente: Elaboración propia



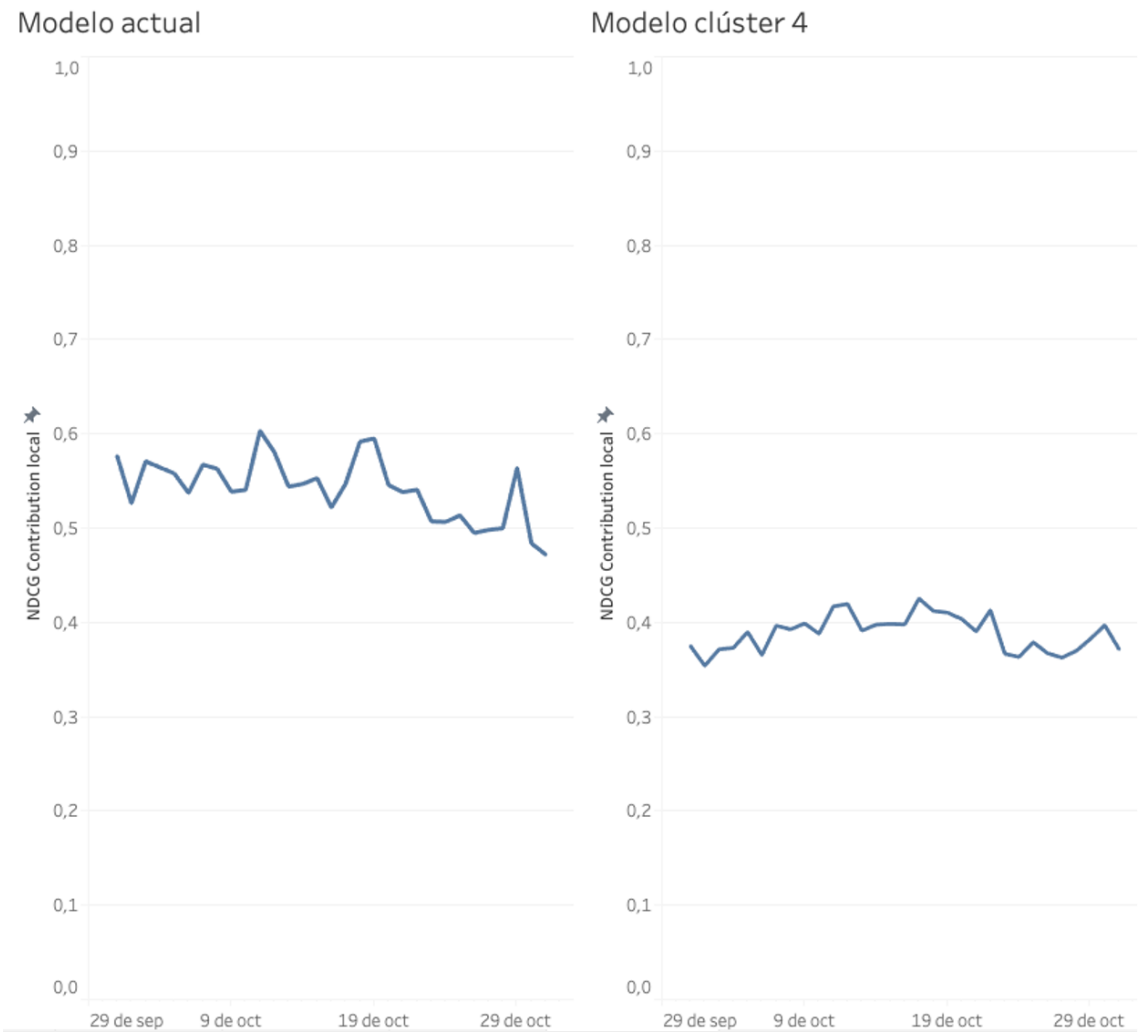
Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 3. Fuente: Elaboración propia



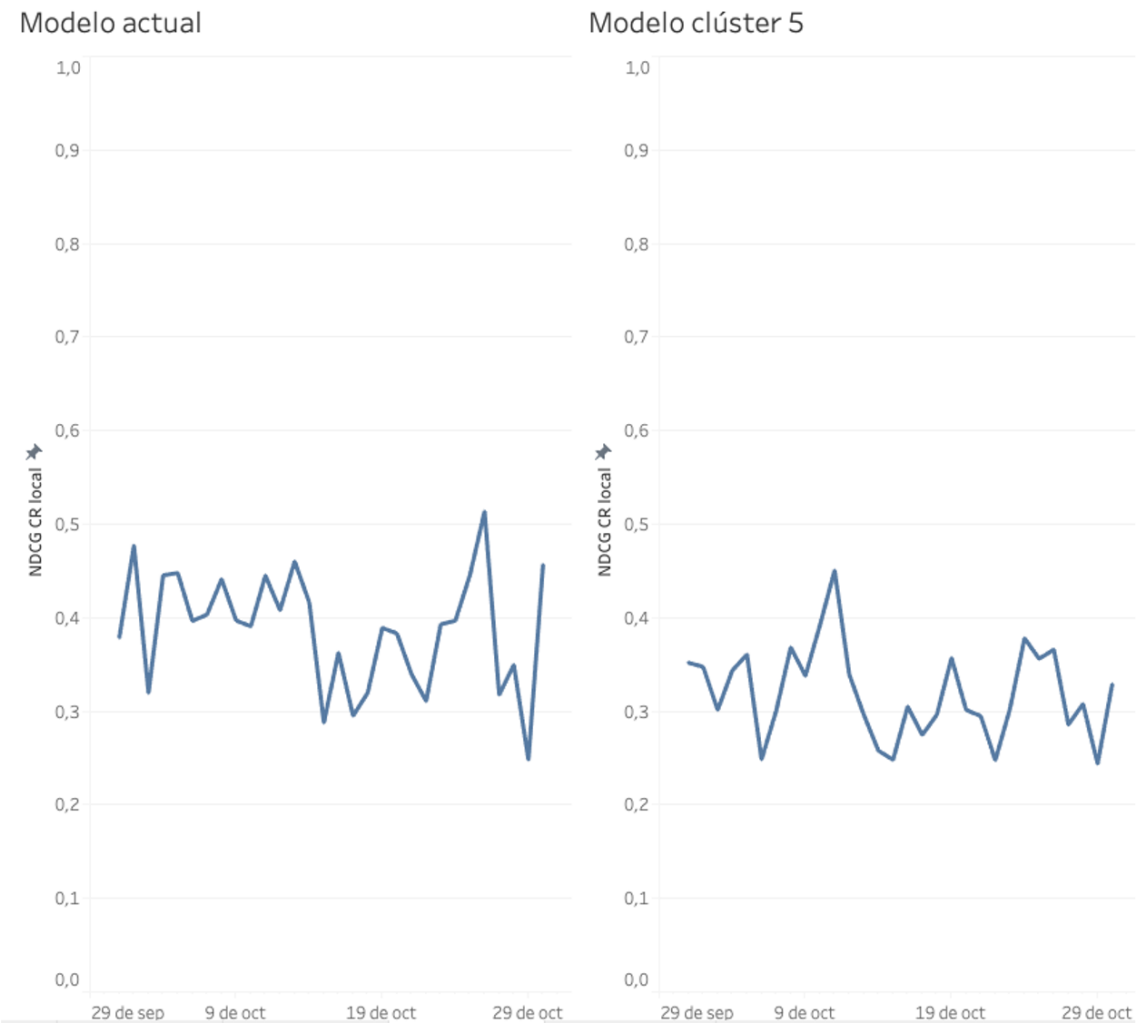
Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 3. Fuente: Elaboración propia



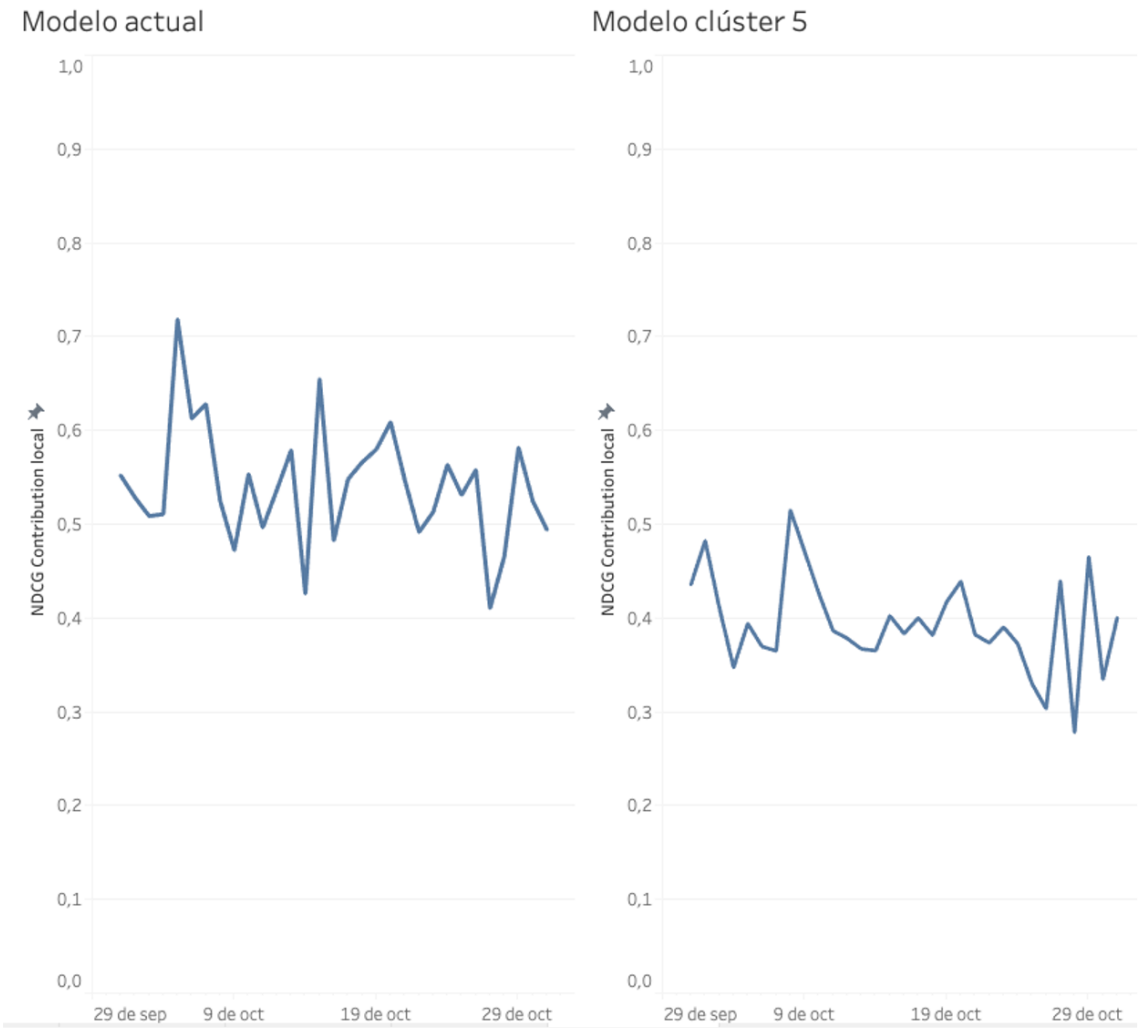
Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 4. Fuente: Elaboración propia



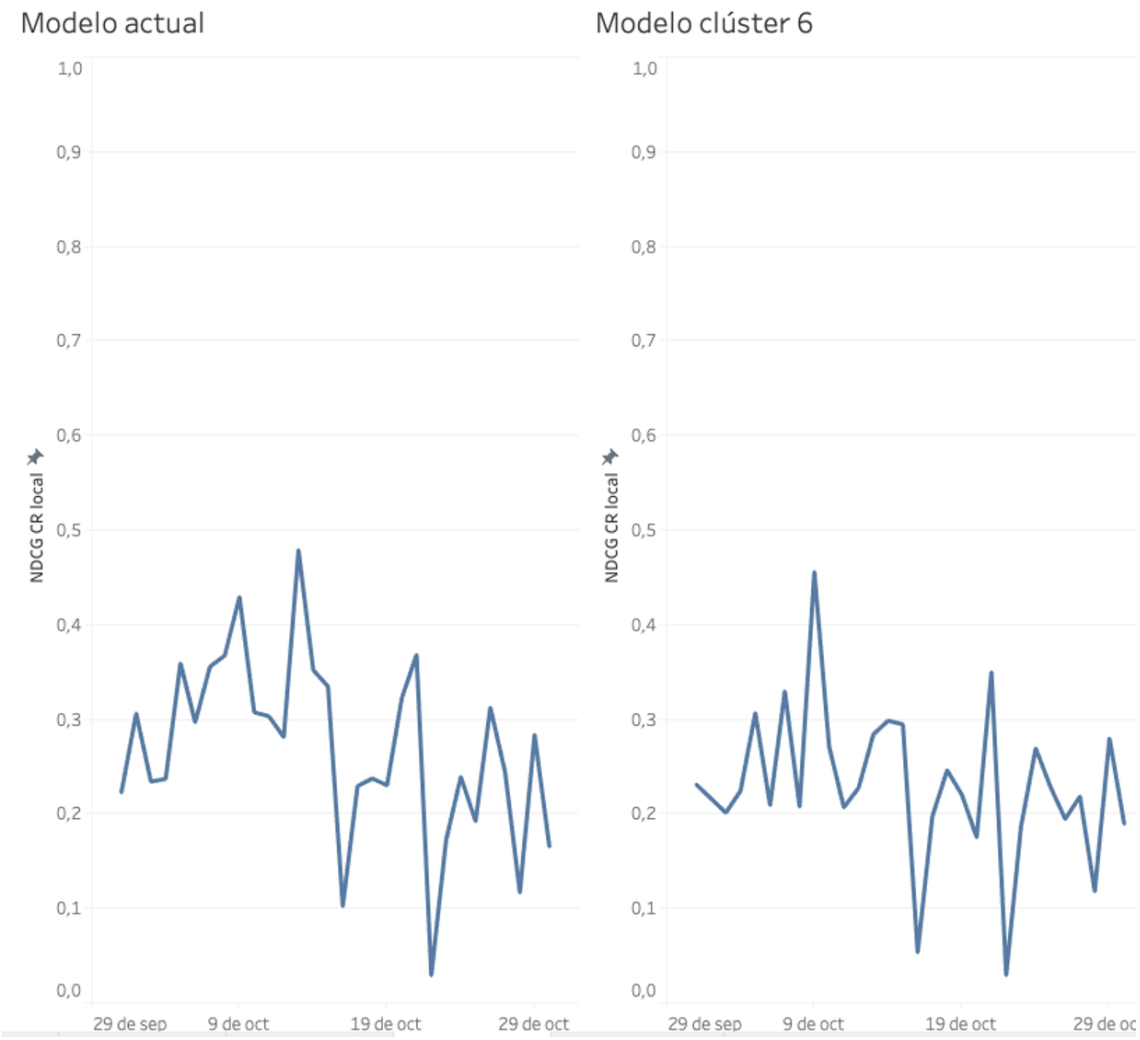
Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 4. Fuente: Elaboración propia



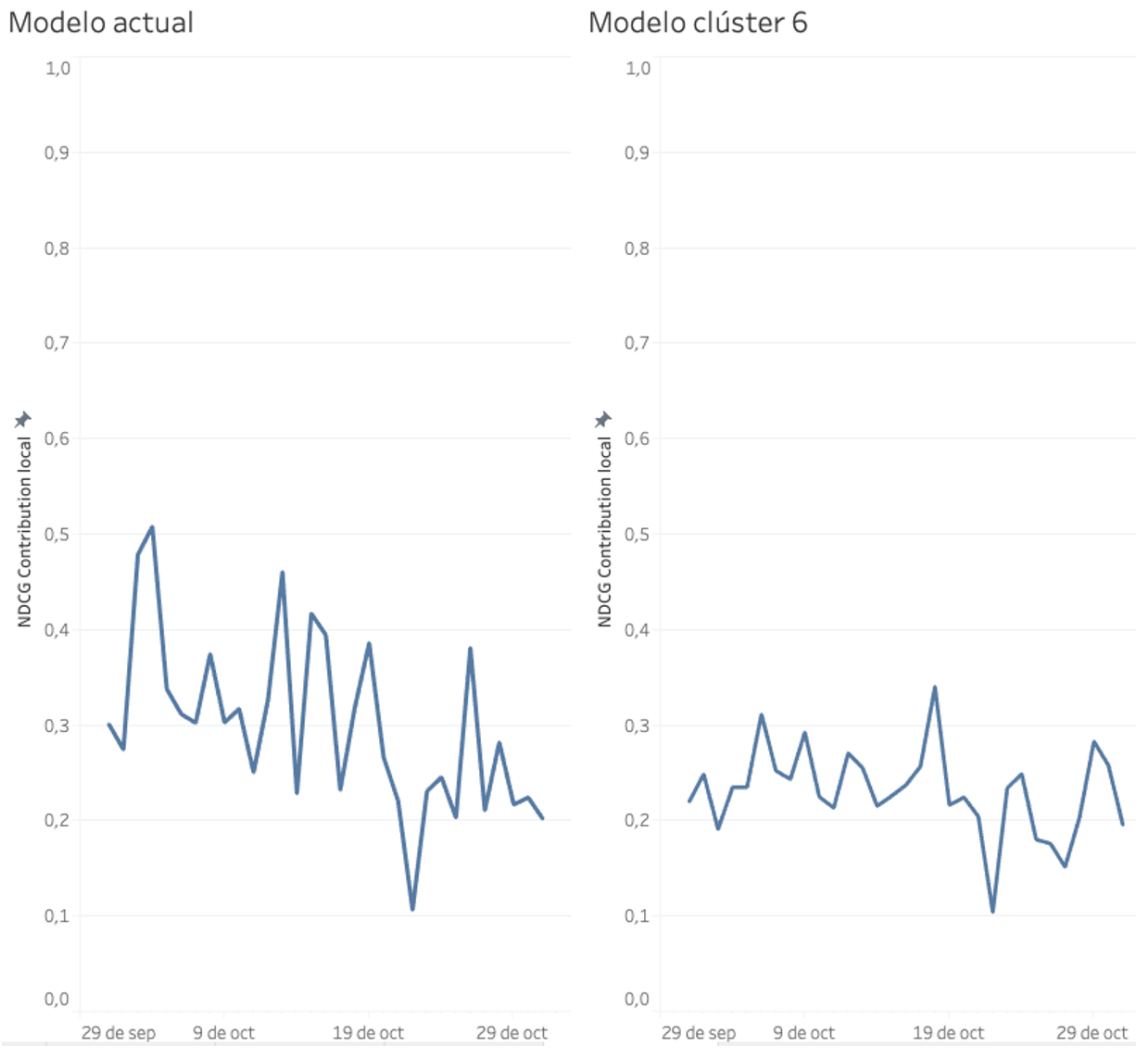
Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 5. Fuente: Elaboración propia



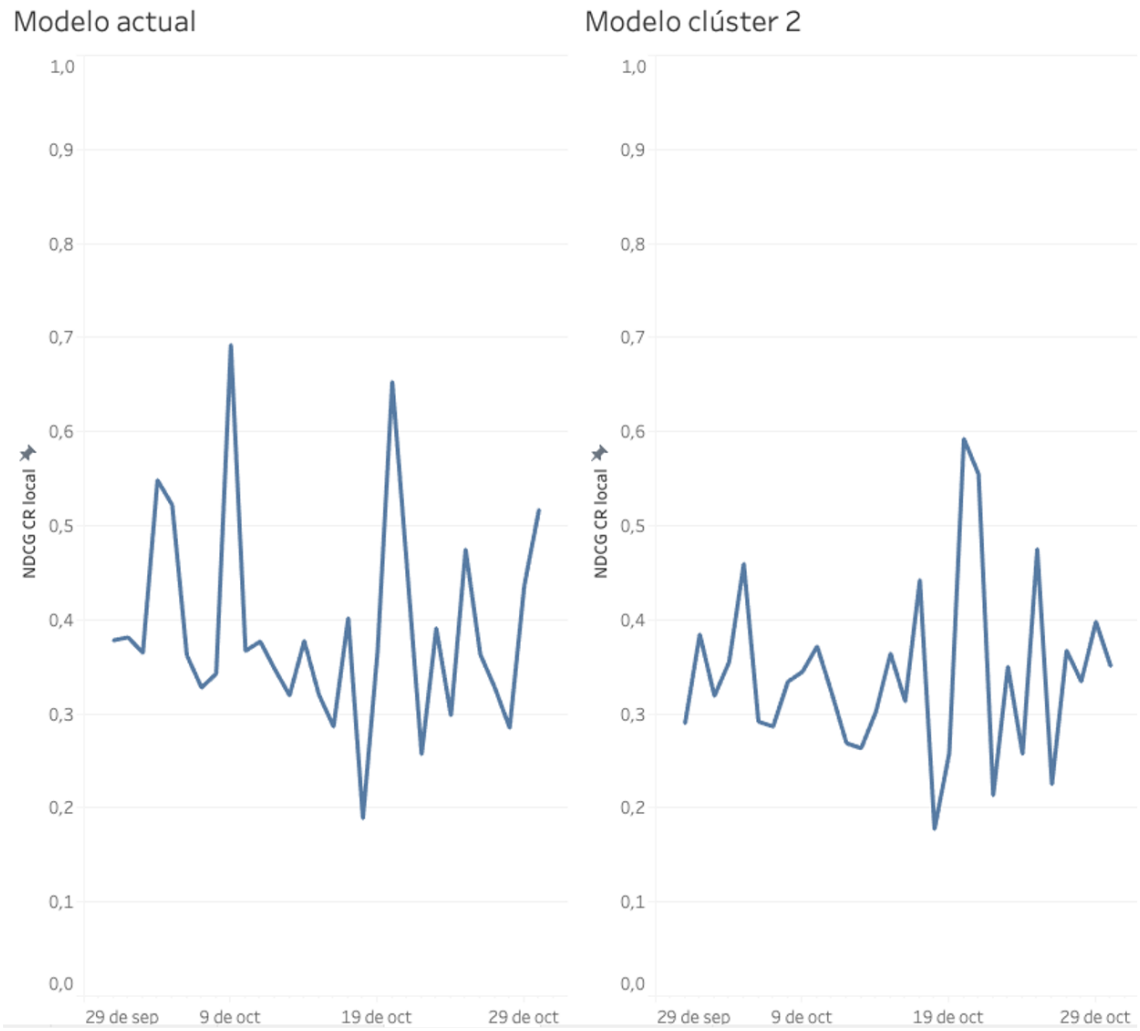
Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 5. Fuente: Elaboración propia



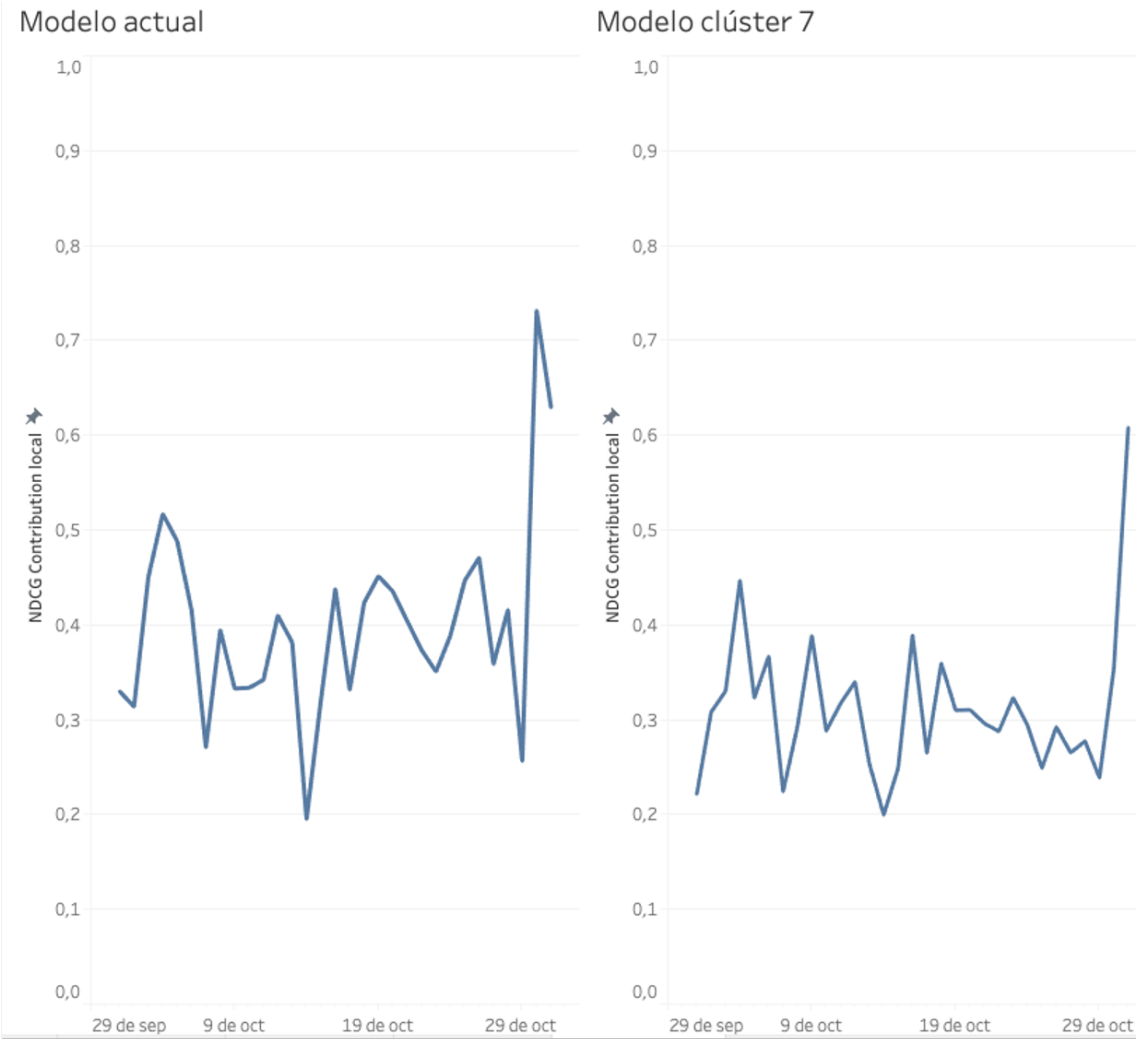
Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 6. Fuente: Elaboración propia



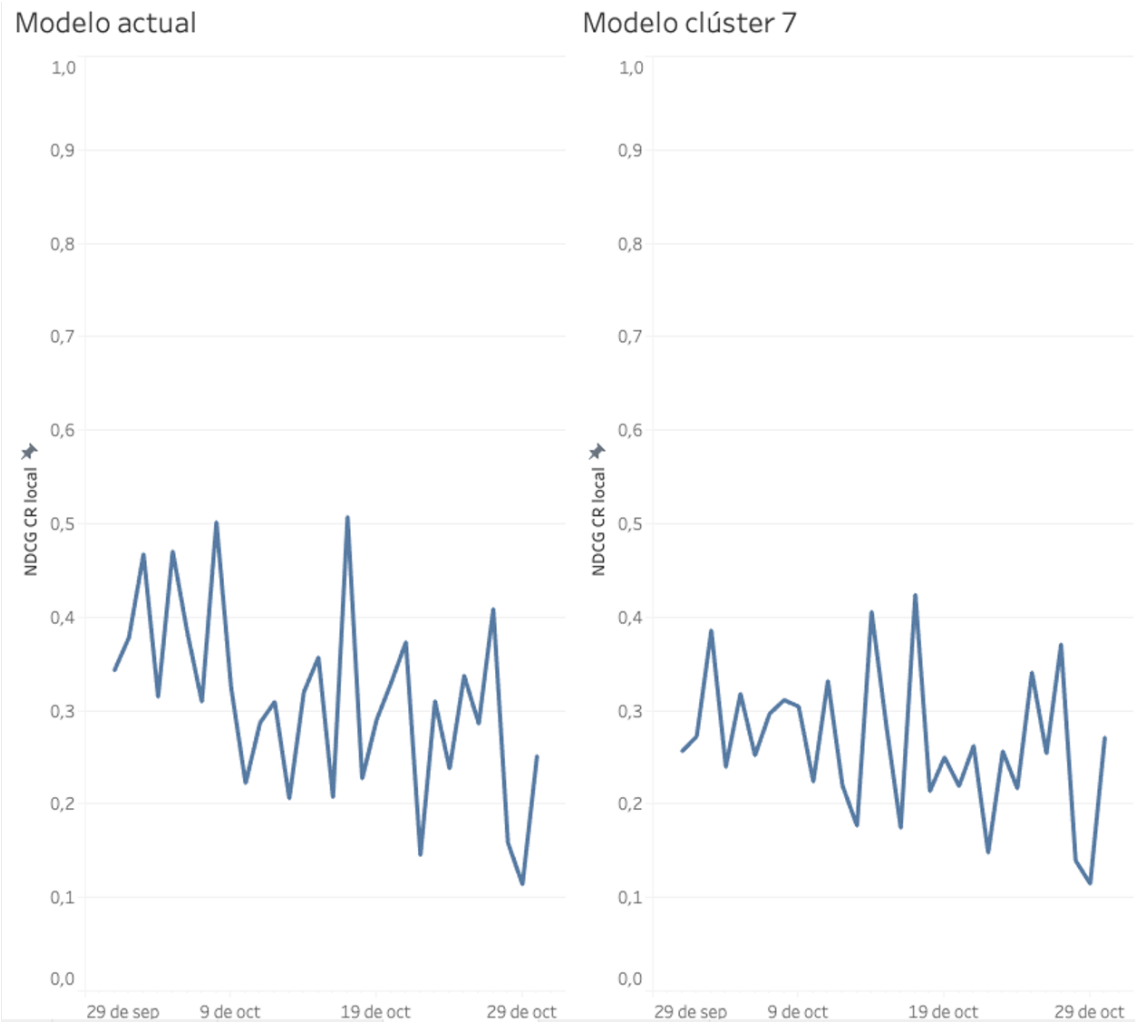
Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 2. Fuente: Elaboración propia



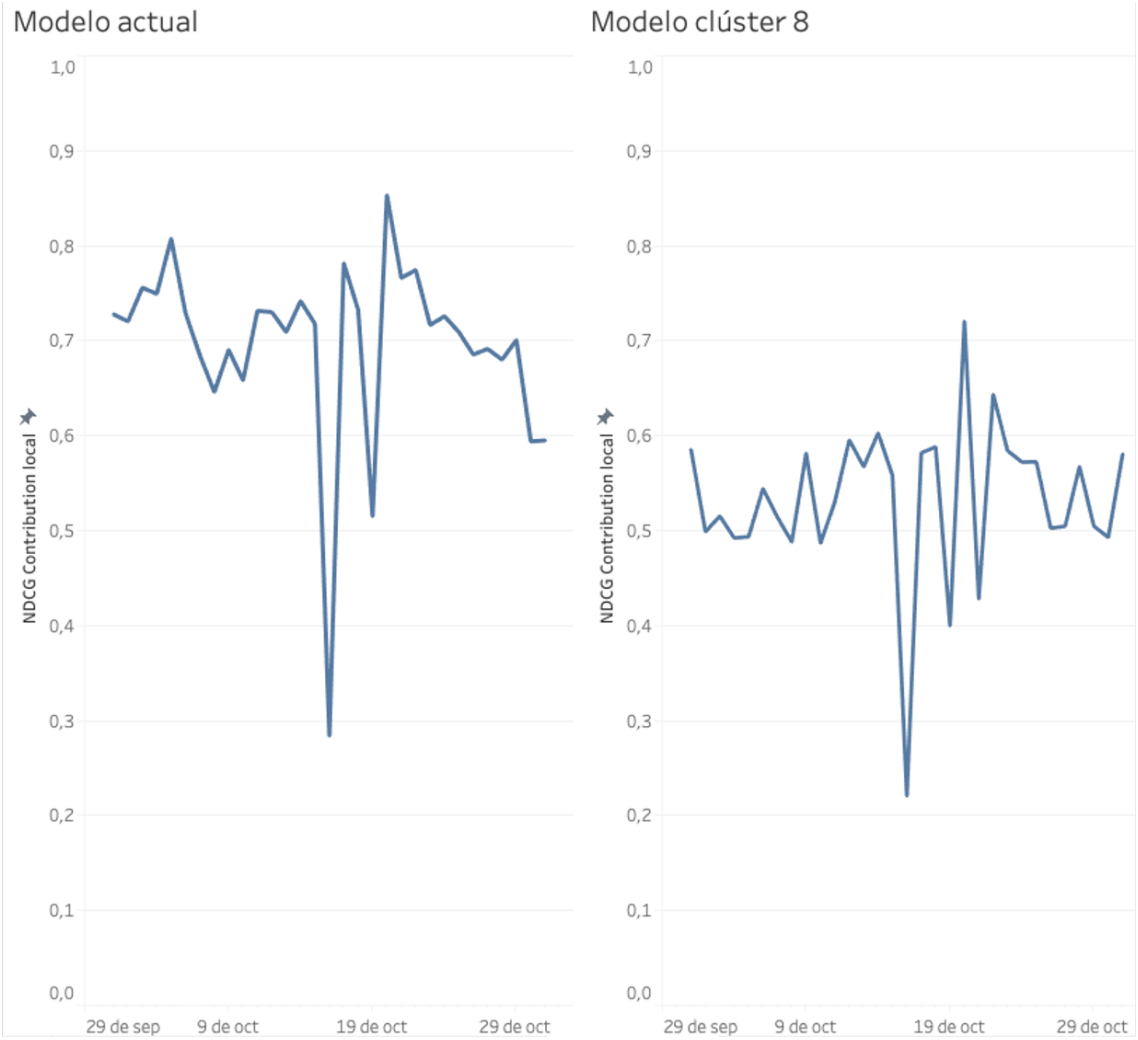
Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 6. Fuente: Elaboración propia



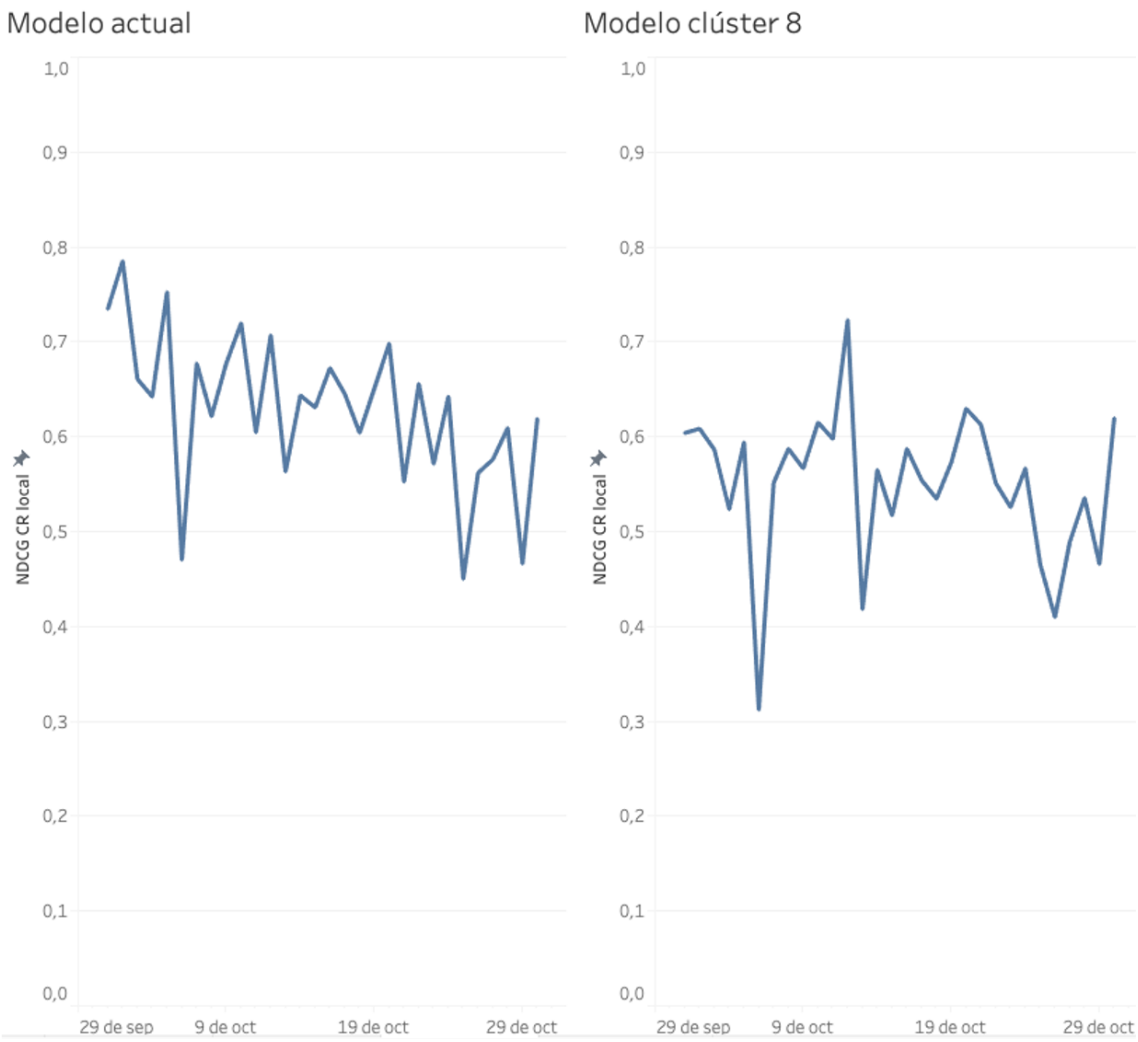
Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 7. Fuente: Elaboración propia



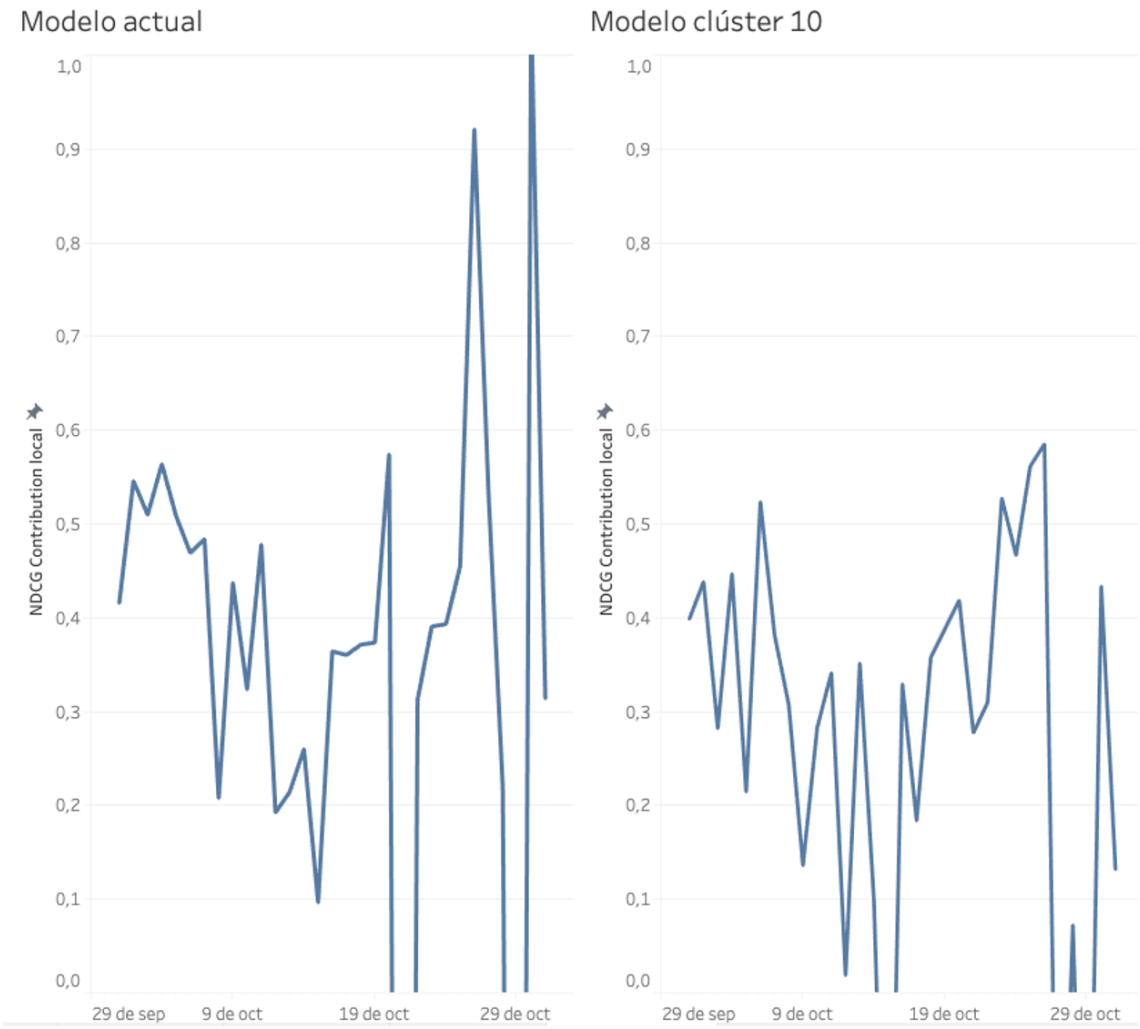
Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 7. Fuente: Elaboración propia



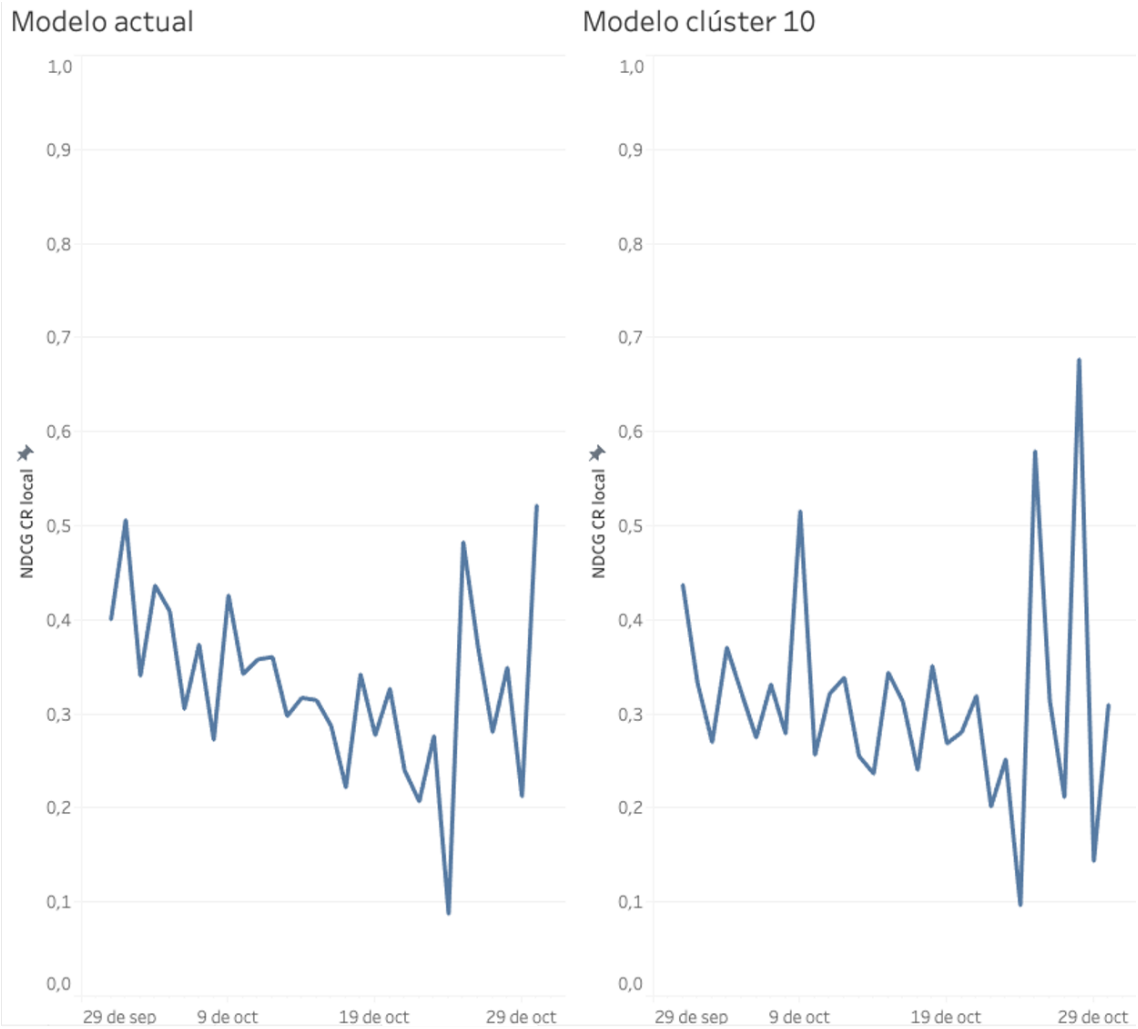
Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 8. Fuente: Elaboración propia



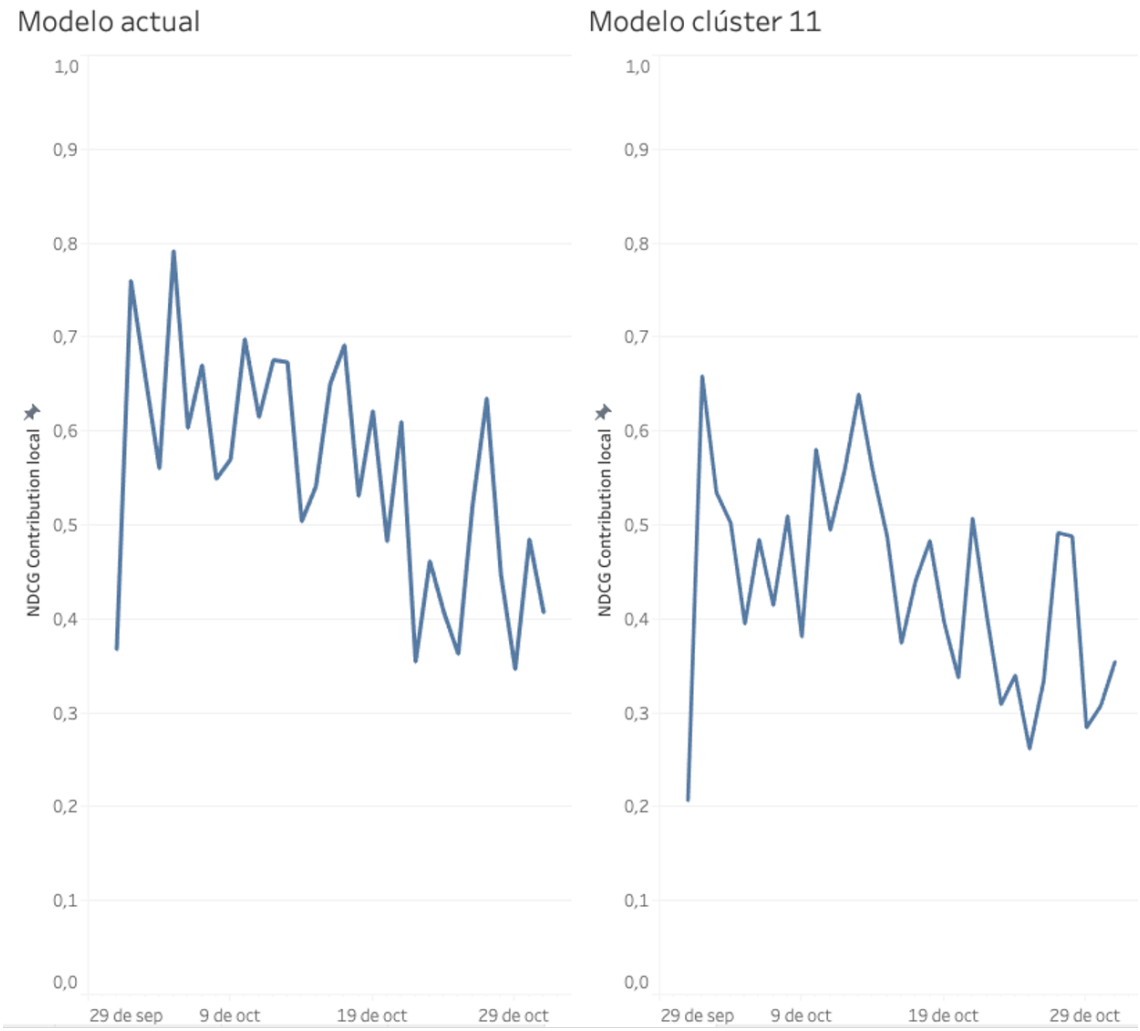
Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 8. Fuente: Elaboración propia



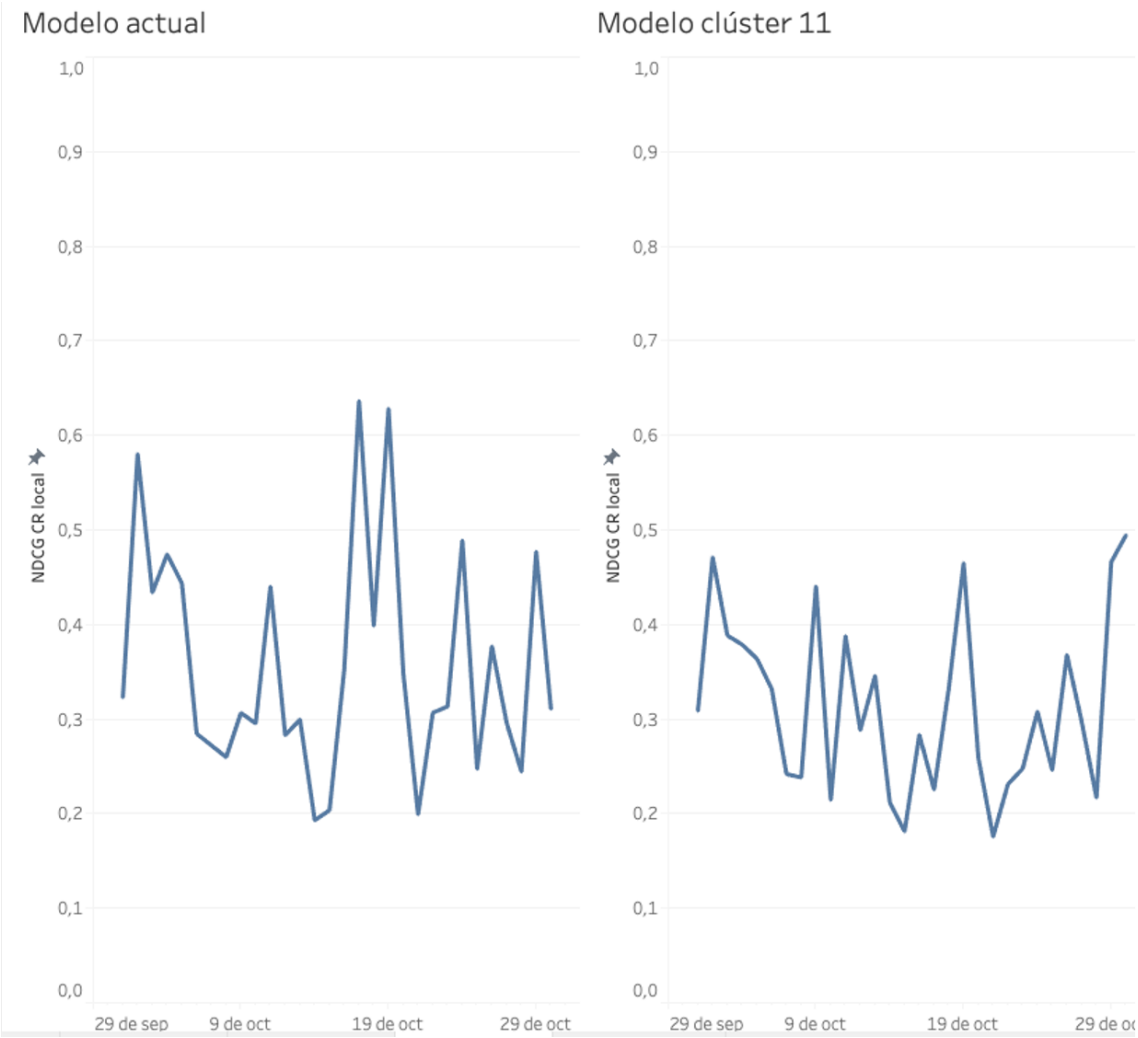
Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 10. Fuente: Elaboración propia



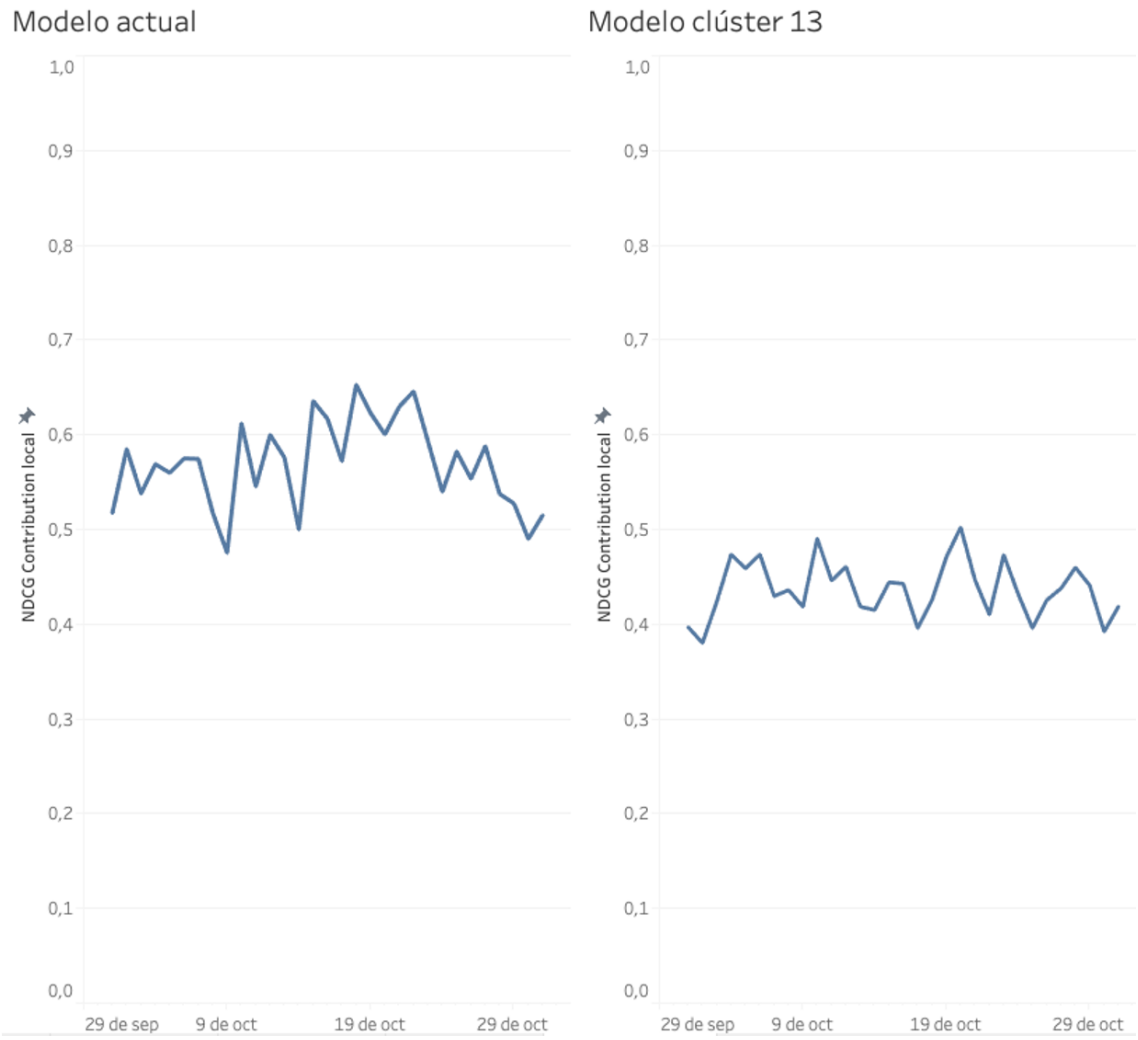
Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 10. Fuente: Elaboración propia



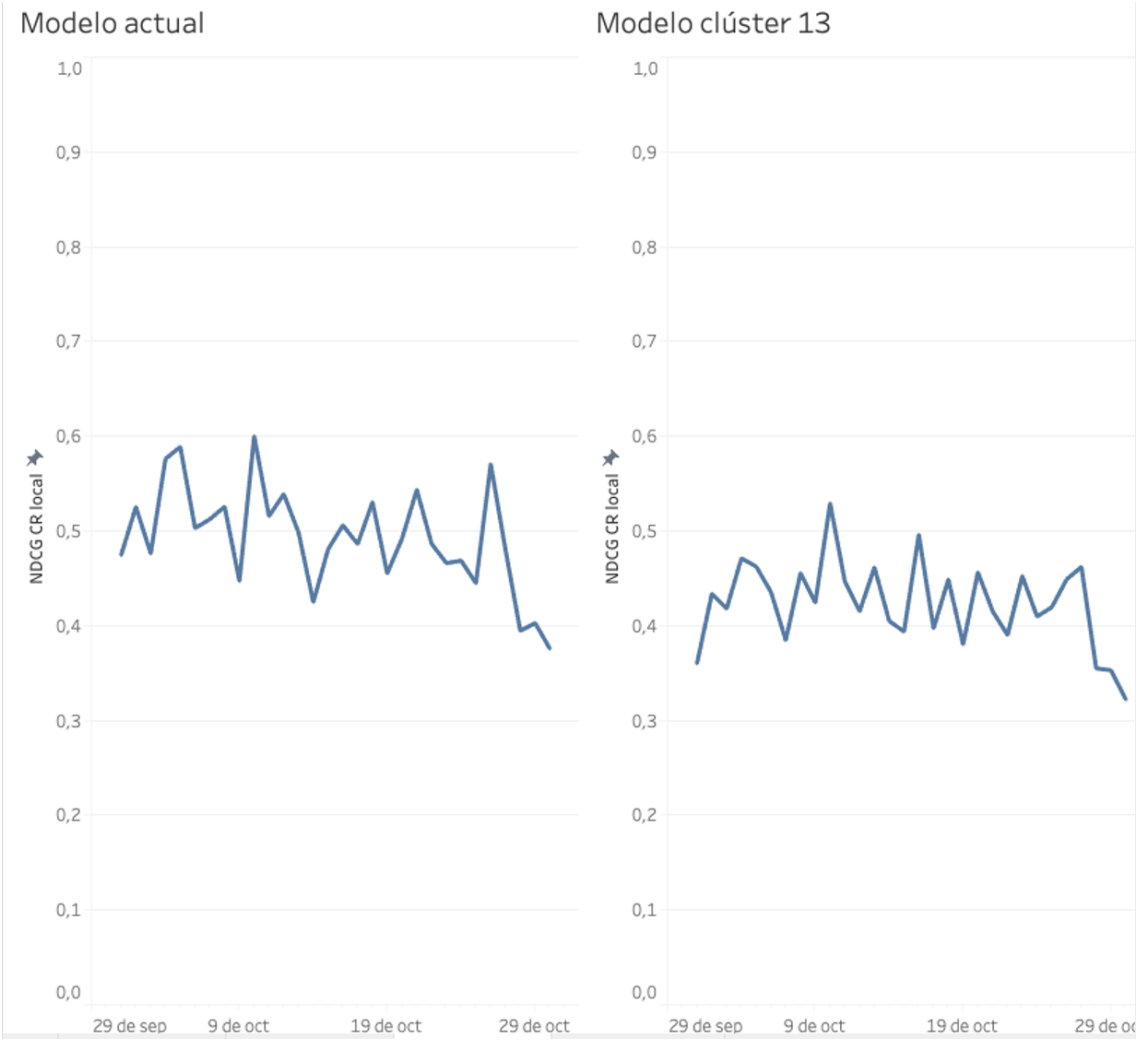
Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 11. Fuente: Elaboración propia



Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 11. Fuente: Elaboración propia



Comparación entre valores de NDCG de contribución local para modelo actual versus modelo del clúster 13. Fuente: Elaboración propia



Comparación entre valores de NDCG de conversión local para modelo actual versus modelo del clúster 13. Fuente: Elaboración propia