INFORME TÉCNICO – ENTREGA FINAL

CONCEPTUALIZACIÓN DEL PROBLEMA

En un entorno empresarial colombiano marcado por la hipercompetencia, la saturación de oferta y la transformación digital, la capacidad de una organización para entender y anticiparse a las necesidades del cliente se ha vuelto una necesidad estratégica. El problema central que se aborda en este reto es la búsqueda de una solución automatizada para suplir la limitada capacidad de las empresas para personalizar la experiencia del consumidor. Esta problemática compromete no solo la satisfacción y fidelización del cliente, sino también el aprovechamiento pleno del valor de los datos transaccionales acumulados, generando ineficiencias operativas y oportunidades de venta desaprovechadas. En este contexto, la ausencia de un enfoque inteligente en la recomendación de productos impide la consolidación de relaciones duraderas con los usuarios y dificulta la optimización de portafolios en función de patrones reales de comportamiento y alineación estratégica.

La problemática tratada también trasciende el ámbito tecnológico para situarse en una intersección entre la analítica de datos, estrategia comercial y diseño de experiencias. Por ende, se plantea la necesidad de desarrollar un sistema robusto que articule capacidades de *machine learning* con objetivos de negocio claramente definidos: incrementar la conversión de ventas, facilitar la labor de los asesores en punto de venta y generar *insights* accionables para la toma de decisiones. La solución debe ser capaz de absorber, interpretar y capitalizar datos heterogéneos de distintas fuentes (transacciones, cotizaciones y ventas B2B), para derivar recomendaciones personalizadas que no solo respondan al historial del cliente, sino que estén alineadas con el portafolio estratégico de la compañía. En suma, se trata de superar las limitaciones actuales de la gestión comercial a través de un modelo de recomendación inteligente que funcione como catalizador de valor integral para el negocio.

ANÁLISIS DESCRIPTIVO

Base Transaccional B2C (base_1_transaccional)

Para comenzar el desarrollo del proyecto, se llevó a cabo un análisis exploratorio detallado utilizando la base de datos B2C, compuesta por 2.099.836 registros y 18 variables explicativas. Esta base contiene información clave sobre el comportamiento de compra del consumidor final, incluyendo datos como el tipo de producto, su categoría y subcategoría, la fecha de transacción, el valor total de la compra y su alineación con el portafolio estratégico B2C.

La calidad de los datos fue satisfactoria en general. No se detectaron registros duplicados y el porcentaje de datos faltantes no fue superior al 2% en la variable precio y al 0.008% en la

variable zona, estos problemas de completitud fueron resueltos eliminando dichos registros ya que la ausencia de esas dos variables los convierte en información no relevante para el proyecto. Con respecto a la preparación de los datos, la variable “fecha_factura” fue transformada correctamente al formato *datetime*, lo que facilita la exploración de tendencias temporales y la identificación de patrones estacionales.

En cuanto a las variables numéricas, la variable “valor” presentó una distribución claramente sesgada hacia la derecha. El 75% de las transacciones registran valores iguales o inferiores a aproximadamente 37 unidades monetarias, lo que indica una alta concentración de compras de bajo valor. Sin embargo, se observan valores extremos que alcanzan hasta 56,876 unidades, lo cual sugiere la existencia de adquisiciones puntuales de alto valor, posiblemente relacionadas con productos premium, compras por volumen o campañas especiales. La mediana se sitúa en 13.32 unidades, reflejando una diferencia significativa frente al valor máximo y reafirmando el sesgo positivo. Por otro lado, la variable “edad” muestra una distribución relativamente simétrica, con una media de aproximadamente 41.9 años y una mediana cercana (43 años). Sin embargo, hay una altísima presencia de compradores al inicio de sus 30, estos podrían representar a los clientes que están independizando y por ende realizan muchas compras para la adecuación de su nuevo hogar.

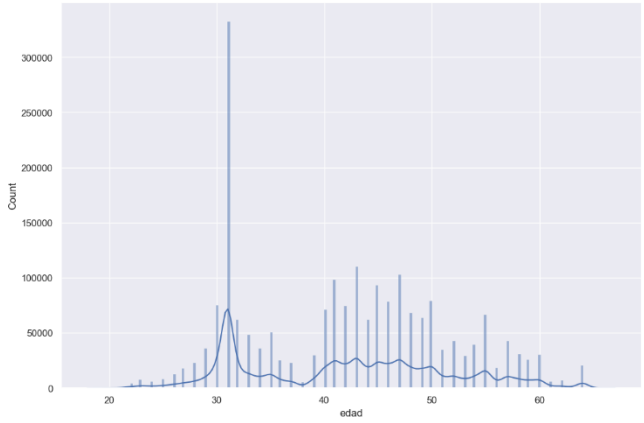

Ilustración 1 Histogramas de frecuencia de edad

La variable “alineación con portafolio estratégico” mostró una alta dispersión, con valores entre -24,187 y 4,162. Aunque la media (3.95) y la mediana (1.41) indican una leve alineación positiva en la mayoría de las transacciones, la presencia de valores negativos y extremos sugiere comportamientos comerciales dispares. Esto resalta la necesidad de guiar a los clientes hacia productos más alineados con la estrategia de la compañía.

El análisis de correlación reveló una relación moderada y positiva entre el “valor” y la “alineación con portafolio estratégico” (r = 0.53), lo que sugiere que las transacciones de mayor valor tienden a estar más alineadas con los objetivos estratégicos de la compañía. También se observa una correlación moderada entre “valor” y “cantidad” (r = 0.51), indicando que el monto total está fuertemente influenciado por el volumen de productos adquiridos.

En general, el resto de las variables presentan correlaciones débiles o casi nulas entre sí, lo que indica una baja multicolinealidad y una relativa independencia entre los atributos.

En lo que respecta a las variables categóricas, se evidencia una fuerte concentración de transacciones en un número reducido de municipios y zonas. El municipio de Curití representa más del 27% del total de transacciones, seguido por Natagaima y Villanueva. A nivel de zona, Santander domina ampliamente con más de 700 mil registros, seguido de Tolima y Antioquia. Esta concentración geográfica plantea oportunidades para diversificar la cobertura comercial y fortalecer estrategias locales en zonas con menor participación.

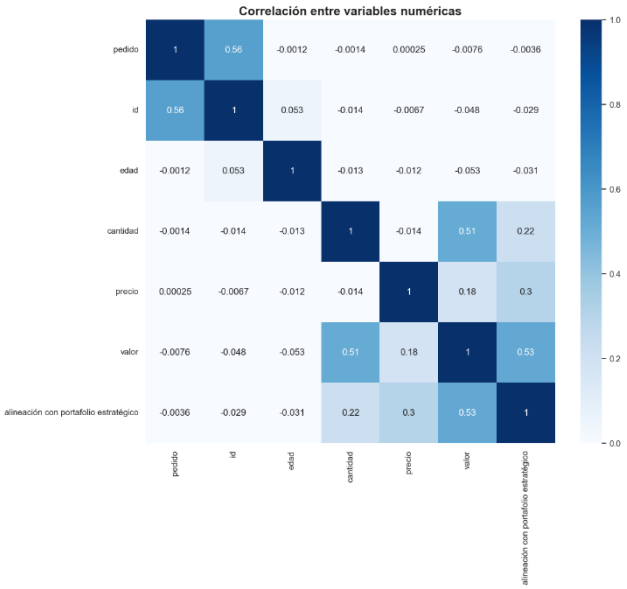

Ilustración 2 Correlación de variables numérica

En cuanto a los puntos de venta y asesores, los cinco principales puntos concentran una proporción significativa del volumen total, con “punto_venta_7” a la cabeza (132,386 transacciones). De manera similar, algunos asesores como “asesor_137” y “asesor_7” superan las 14 mil ventas, mientras que más de 300 asesores tienen apenas una transacción registrada, lo que refleja una distribución altamente desigual en el desempeño comercial.

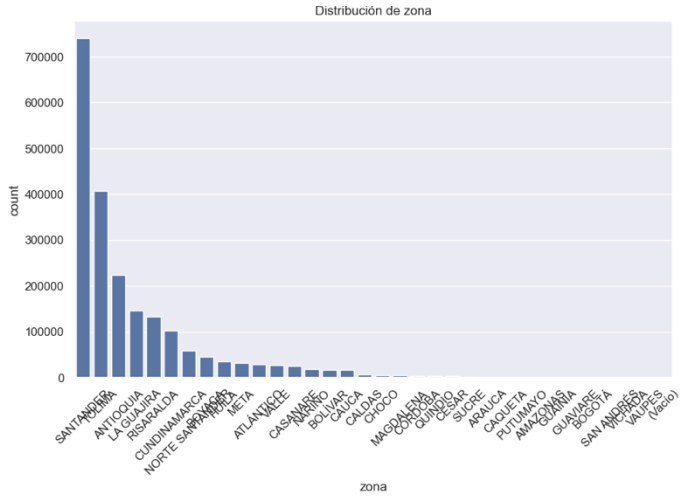

Ilustración 3 Frecuencia de departamentos de los registros

Respecto a los clústeres, más del 80% de las ventas están concentradas en “cluster_tienda_3” y “cluster_tienda_2”, lo cual indica una segmentación de mercado dominada por unos pocos perfiles de tienda. A nivel de producto, se identificaron más de 7,200 referencias únicas, aunque el producto más vendido (“producto_49”) representa apenas el 2.56% del total, lo que indica una alta dispersión en las preferencias del consumidor. Este patrón se mantiene a nivel de subcategorías y categorías, con más de 100 subcategorías y 27 categorías, pero una fuerte concentración en las más populares como “subcategoria_5” y “categoria_3”.

En cuanto a las categorías macro, cinco agrupan la totalidad de las transacciones, destacándose “categoria_macro_2” con más de 1.3 millones de registros (62% del total), seguida por “categoria_macro_4” y “categoria_macro_1”. Finalmente, la variable “color” también muestra un patrón concentrado: más del 47% de las referencias no tienen color identificado, y entre las restantes predominan tonos neutros como gris, blanco y beige. Estos

hallazgos destacan la necesidad de mejorar la calidad del dato en ciertas dimensiones y de adoptar estrategias de diversificación que reduzcan la dependencia de territorios, asesores y productos específicos.

Base de Cotizaciones B2C (base_2_cotizaciones)

Para comprender la dinámica de cotización del segmento B2C, se llevó a cabo un análisis exploratorio exhaustivo sobre la base de datos proporcionada, la cual contiene 180.387 registros distribuidos en 11 variables explicativas. Esta base refleja el comportamiento de los clientes en procesos de solicitud de información o intención de compra, permitiendo identificar patrones de interés sobre el portafolio comercial y la interacción con los productos ofrecidos. Con respecto a la calidad de datos se encontró que la base de datos estaba completa, sin embargo, se evidenció un problema de duplicados que fue resuelto eliminándolos. La consistencia de las variables numéricas y categóricas, junto con la validez de estos fueron coherentes con el contexto y los datos. Los datos de fecha no estaba en formato *datetime* lo que podía perjudicar el manejo de los mismo y por ende se cambió al formato esperado.

El análisis de las variables numéricas en las cotizaciones B2C muestra un comportamiento altamente asimétrico y disperso. El valor total (valor) tiene una media de 36,9 pero un máximo superior a 366.000, evidenciando *outliers* significativos y un sesgo positivo. El 75% de los registros están por debajo de 27,6, lo que indica un mercado dominado por productos de bajo costo, aunque existen casos excepcionales de productos premium. La variable cantidad presenta una alta dispersión (std = 1.128) y un valor máximo extremo de 438.000 unidades, lo cual también sugiere la existencia de registros atípicos o compras mayoristas aisladas.

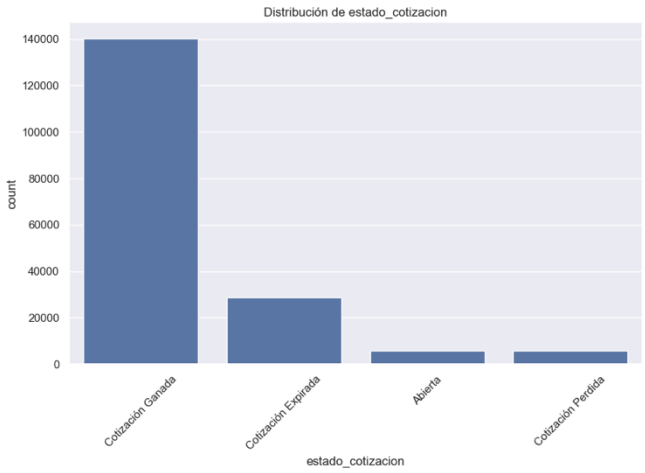

Ilustración 4 Histograma de distribución de estado de cotización

En cuanto a las variables categóricas, específicamente "estado_cotizacion", se observa una clara predominancia de cotizaciones ganadas, que representan cerca del 77% del total. Le siguen en frecuencia las cotizaciones expiradas, mientras que las abiertas y perdidas tienen una participación marginal y muy similar entre sí.

Este patrón sugiere un proceso comercial con alta efectividad de conversión desde la etapa de cotización, aunque también podría reflejar una sobre clasificación automática como "ganada" si no hay actualización posterior.

En cuanto a correlaciones, valor se relaciona débilmente con cantidad (r = 0.10) y precio (r = 0.056), lo cual es esperado ya que valor = cantidad * precio, pero las correlaciones bajas indican alta variabilidad en estas relaciones. El resto de las variables no muestra correlaciones relevantes con los datos económicos y operativos, y pueden considerarse identificadores sin valor predictivo. Este comportamiento refuerza la necesidad de limpieza de *outliers*, especialmente en cantidad, y de modelar considerando la alta varianza y fragmentación del mercado.

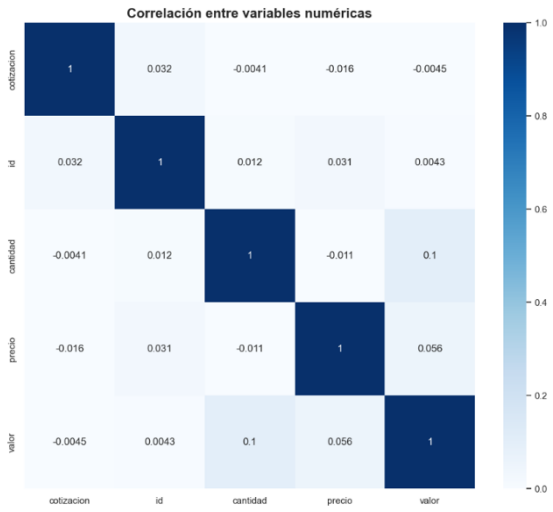

Ilustración 5 Correlación entre las variables numéricas del cotizaciones

Base Transaccional B2B (base_3_transaccional_b2b)

Para comprender la dinámica comercial del segmento B2B, se realizó un análisis exploratorio detallado sobre la base de datos brindada, la cual contiene 25,866 transacciones con 10 variables explicativas. Esta base representa información transaccional relevante, destacando aspectos clave como el tipo de producto adquirido, su categoría y subcategoría, la fecha de compra, el valor total de la transacción y su alineación con el portafolio estratégico B2B. Es importante mencionar que se evaluó la calidad de los datos, destacando que no se identificaron valores nulos ni registros duplicados. Esto facilitó un análisis directo, sin necesidad de limpieza adicional. La variable "fecha_factura" fue correctamente transformada al formato *datetime*, permitiendo futuros análisis temporales como identificación de estacionalidades o tendencias comerciales.

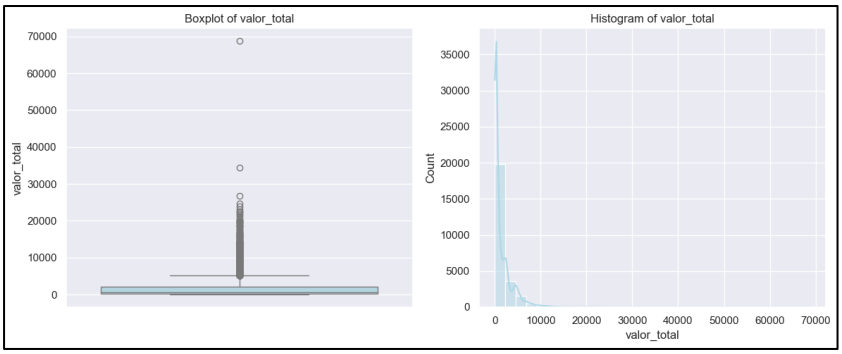

Ilustración 6 Histograma y boxplot de "valor_total"

Respecto a las variables numéricas, la variable "valor_total" mostró una distribución claramente sesgada hacia la derecha, con numerosos valores atípicos, como se evidencia en los histogramas y *boxplots* realizados. El 75% de las transacciones están por debajo de aproximadamente 2,200 unidades monetarias, destacando una concentración de valores bajos a medios, mientras que algunos valores extremadamente altos alcanzan hasta cerca de 68,750 unidades monetarias. Estos valores atípicos podrían estar asociados a transacciones puntuales de alto valor,

posiblemente correspondientes a productos premium o adquisiciones de clientes corporativos clave.

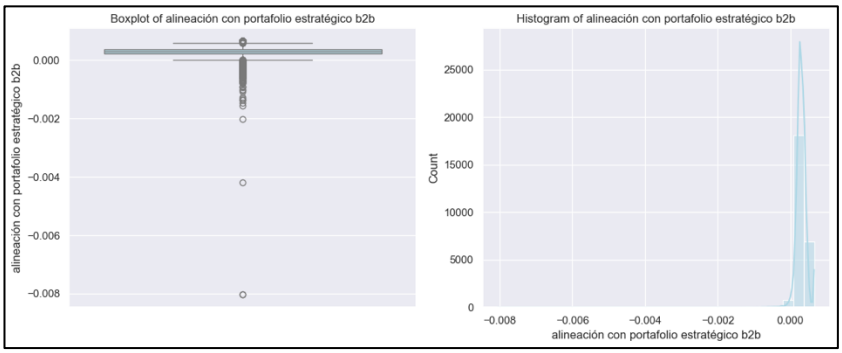

Ilustración 7 Histograma y boxplot de "alineación con portafolio estratégico B2B"

La variable “alineación con portafolio estratégico B2B” mostró una distribución altamente concentrada alrededor de cero con una desviación estándar muy baja (0.000155), indicando que la mayoría de las transacciones están relativamente cercanas a un valor neutro. Sin embargo, la media (0.000294) y el valor máximo observado (0.000654) son tan bajos que, desde un punto de vista práctico, se interpretan como una baja alineación con el portafolio estratégico. Esto implica que muchas de las compras realizadas, aunque no necesariamente negativas, no parecen ser las más efectivas ni acordes con la estrategia definida por la empresa.

La presencia adicional de valores negativos, llegando hasta -0.008, enfatiza aún más la existencia de transacciones que claramente se alejan del enfoque estratégico deseado. En términos prácticos, esto representa una importante oportunidad para reevaluar y ajustar las estrategias comerciales, asegurando que las futuras transacciones prioricen productos con mayor valor estratégico para la compañía, optimizando así el uso de recursos y aumentando la efectividad comercial general. No obstante, se entiende que, al tratarse del segmento B2B, existe una limitación inherente al control directo sobre los productos que terceros comercializan. Por ello, se podría pensar en implementar campañas dirigidas específicamente a incentivar o patrocinar productos más alineados con la estrategia de la empresa, motivando así a los aliados comerciales a promover aquellas referencias que mejor se ajustan a los objetivos estratégicos corporativos.

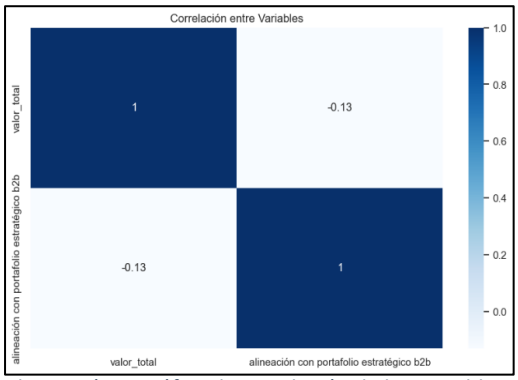

Ilustración 8 Gráfica de correlación de las variables numéricas B2B

En el análisis de correlación, se identificó una correlación negativa débil (r = -0.13) entre “valor_total” y “alineación con portafolio estratégico b2b”. Esta asociación, aunque pequeña, indica una leve tendencia en que transacciones de mayor valor podrían estar menos alineadas con el portafolio estratégico. Esto podría significar que los clientes con transacciones más altas podrían estar adquiriendo productos menos prioritarios desde la perspectiva estratégica de la empresa. Tal hallazgo, aunque limitado en magnitud, ofrece información valiosa para futuras

revisiones estratégicas y comerciales que permitan aumentar la alineación estratégica de las transacciones más significativas en términos monetarios.

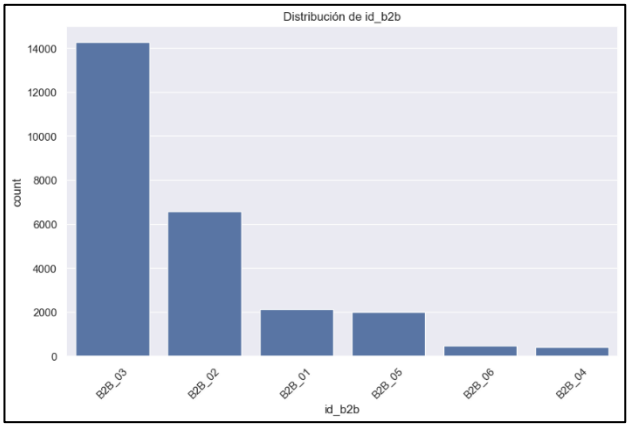

Ilustración 9 Distribución de los compradores B2B

Desde el punto de vista categórico, se observa una concentración significativa en pocos actores. A nivel de clientes B2B, más del 50% de las transacciones están concentradas en el cliente B2B_03, seguido por B2B_02 y B2B_01. Esto implica una fuerte dependencia comercial en unos pocos aliados estratégicos. Por municipios, destaca Fusagasugá con más del 62% de las transacciones, seguido por Villa de Leyva y Madrid. Un pequeño porcentaje (1.5%) tiene registros con municipio desconocido ("#"), lo cual debería revisarse si se desea realizar análisis geográficos más precisos. En cuanto a zonas, Cundinamarca representa el 71% de las ventas, seguida de Boyacá con un 27%. En cuanto a las categorías de productos, tanto a nivel macro como específico, la distribución también es bastante desigual. Las cinco principales categorías macro concentran más del 78% del total de transacciones, siendo la categoría "cat_b2b_macro_1" la más destacada con el 22.3%. A nivel más granular, se identifican más de 130 subcategorías y más de 2,500 productos únicos, aunque nuevamente, una minoría de ellos acumula la mayoría del volumen de transacciones. Por ejemplo, el producto más vendido apenas representa el 1.11% del total.

Los hallazgos revelan una estructura de ventas altamente concentrada, tanto en un número reducido de clientes como en un subconjunto limitado del portafolio de productos, lo que resalta la necesidad de adoptar estrategias segmentadas más precisas. Una vía clave para abordar esta situación es diversificar el portafolio transaccional, impulsar la rotación de productos estratégicamente alineados y disminuir la dependencia comercial de un grupo tan acotado de clientes y referencias. No obstante, considerando que en el modelo B2B la empresa no ejerce control directo sobre las decisiones de compra de sus aliados, la implementación de un sistema de recomendación debe enfocarse en una lógica de acompañamiento inteligente. Aprovechando los datos detallados disponibles tanto del canal B2B como del B2C —incluyendo información geográfica como municipios y departamentos— es posible identificar patrones de consumo específicos por zona. Con base en esta evidencia, el sistema podrá sugerir al aliado comercial productos con alta demanda en su región, motivándolo a incorporar dichas referencias para responder mejor a las necesidades locales. Así, lejos de imponer decisiones, el modelo actúa como un facilitador estratégico que orienta al cliente hacia decisiones comerciales más informadas.

Enriquecimiento de la base de datos

En el contexto de este reto, donde se busca desarrollar un modelo de recomendación inteligente capaz de personalizar la oferta de productos y mejorar la experiencia del cliente, se identificó la necesidad de complementar las bases de datos transaccionales con información adicional que permitiera captar mejor el contexto y las necesidades de los consumidores. Aunque inicialmente se evaluó la posibilidad de incorporar variables individuales como género, ocupación o preferencias personales, las bases disponibles —particularmente la B2C, que cuenta con el mayor nivel de granularidad— no incluyen este tipo de atributos. Si bien contiene datos como edad, municipio, punto de venta y producto adquirido, estos no son suficientes para capturar completamente los factores que inciden en la decisión de compra. Frente a esta limitación, se optó por enriquecer la información con variables externas de carácter demográfico, que pudieran contextualizar la demanda a nivel territorial. Se incorporaron así datos agregados por municipio y/o departamento, como porcentaje urbano, nivel de empleabilidad y otras condiciones socioeconómicas.

Este enfoque permite al modelo de recomendación ir más allá de los historiales de compra y considerar factores que inciden en las necesidades reales de los consumidores según su entorno. En lugar de imponer decisiones dentro del canal B2B —donde el control directo sobre lo que el tercero decide comprar es limitado—, el sistema podrá sugerir productos relevantes basados en comportamientos observados en la misma zona geográfica. Por ejemplo, si en determinado municipio los datos B2C muestran alta demanda de un producto específico, se puede recomendar al cliente B2B que lo incluya en su portafolio para satisfacer la demanda local. Con incluir esta información extra, se busca que el modelo no solo mejore la conversión de ventas y facilite el trabajo de los asesores comerciales, sino que también funcione como una herramienta estratégica para alinear las decisiones comerciales con los objetivos de negocio, maximizando así el valor de los datos y la eficiencia operativa.

Base Transaccional B2C (base_1_transaccional)

Con el fin de fortalecer el modelo de recomendación y capturar mejor el comportamiento del consumidor final, se decidió enriquecer la base transaccional del canal B2C a través de la construcción de nuevas variables derivadas directamente del historial de compras. Estas métricas permiten caracterizar a cada cliente desde múltiples dimensiones: intensidad de consumo (“total_productos”, “total_gasto”), sensibilidad al precio (“precio_promedio”), frecuencia de compra (“num_pedidos”), monto promedio por transacción (“ticket_promedio”), profundidad por pedido (“cantidad_promedio”) y diversidad en el portafolio consumido (“categorias_diferentes”). En conjunto, estas variables no solo mejoran la granularidad del perfilamiento, sino que habilitan una segmentación más precisa para personalizar las recomendaciones según patrones de comportamiento reales.

Adicionalmente, se incorporaron variables externas provenientes de fuentes oficiales como el DANE, con el propósito de capturar el contexto demográfico y socioeconómico de los territorios en los que se realizan las compras. Se incluyeron atributos como la edad e ingreso laboral promedio, el porcentaje urbano, el índice de pobreza multidimensional (IPUG), el coeficiente de GINI y el total de edificaciones en obra por municipio o departamento. Estas variables permiten interpretar la demanda no solo desde la perspectiva individual, sino también desde las condiciones estructurales del entorno. Por ejemplo, el ingreso promedio y la urbanización pueden influir directamente en la capacidad adquisitiva y en la naturaleza de los productos más demandados, mientras que las edificaciones en obra ofrecen señales de crecimiento territorial que podrían anticipar futuras oportunidades de consumo.

Esta integración de información transaccional con variables contextuales responde a la necesidad de desarrollar un sistema de recomendación más robusto, que no se limite a repetir patrones de compra, sino que sea capaz de "entender" al cliente desde una lógica territorial, económica y social. Con ello, el modelo no solo personaliza de forma más efectiva, sino que también se convierte en una herramienta estratégica para la empresa, al alinear sus decisiones comerciales con el potencial real de cada zona y con el perfil específico de sus consumidores.

Base Transaccional B2B (base_3_transaccional_b2b)

En el marco del desarrollo del modelo de recomendación, y reconociendo las particularidades del canal B2B — donde la compañía no tiene control directo sobre las decisiones de compra de sus aliados comerciales—, se incorporaron variables externas de carácter territorial con el fin de contextualizar mejor el entorno económico y urbano en el que estos operan. A diferencia del canal B2C, donde se cuenta con información granular por cliente, en el B2B el enriquecimiento se concentró en datos agregados a nivel municipal y departamental, lo que permite capturar dinámicas estructurales clave para orientar recomendaciones estratégicas.

Se incluyeron variables como el total de edificaciones en obra, el promedio y la tasa de edificaciones por manzana (normalizadas por tamaño territorial), y la tasa por habitante, todas orientadas a medir el dinamismo urbano y el potencial de crecimiento físico del territorio. Estas métricas permiten identificar zonas en expansión o transformación urbana, donde es esperable una mayor demanda de productos para el hogar, la construcción o la remodelación, alineados con el portafolio de la compañía. Adicionalmente, se integraron variables económicas como el total de unidades económicas activas y la participación económica relativa del municipio dentro del total nacional. Estos indicadores permiten estimar la densidad empresarial y el peso económico de cada zona, facilitando la

priorización comercial en función de su relevancia estratégica y su potencial de transacciones.

La inclusión de estas variables transforma el modelo de recomendación en una herramienta propositiva y sensible al contexto territorial. Lejos de imponer productos, el sistema puede sugerir aquellos con mayor probabilidad de aceptación local, basándose en evidencia concreta del entorno donde opera cada aliado. Este enfoque mejora la pertinencia de las recomendaciones, fortalece la relación comercial y permite una alineación más efectiva con la demanda real, lo que puede traducirse en mayores niveles de rotación, mejor experiencia para el consumidor final y una mayor fidelización al canal.

PRINCIPALES HALLAZGOS

En el canal B2B se exploraron técnicas de agrupamiento para segmentar aliados comerciales según patrones de compra. Inicialmente se aplicó **KMeans**, pero el resultado fue insatisfactorio (coeficiente de silueta de 0.13), lo que indicaba agrupaciones poco claras. Se optó entonces por **DBSCAN**, que arrojó un coeficiente de silueta de **0.78**, revelando segmentos bien definidos. Sin embargo, el algoritmo generó **1,178 clusters**, muchos con muy pocos registros, lo que dificultaba extraer conclusiones prácticas.

Ilustración 10 PCA MetaClusters

Para consolidar esta segmentación, se calcularon los centroides de los clusters de DBSCAN (sin outliers) y se aplicó KMeans nuevamente sobre estos puntos, sintetizando los grupos en **10 meta-clusters** más interpretables. Este enfoque permitió identificar perfiles de clientes con comportamientos comerciales diferenciados.

Entre los más destacados está el **Meta-Cluster 0**, con 8,416 registros y alta concentración en tres productos (Producto_54, _55 y _56), pero con bajo valor promedio de compra, ideal para estrategias de aumento del ticket promedio. El **Meta-Cluster 1** (1,533 registros) muestra fuerte dependencia de Producto_377 y un ticket promedio medio, lo que lo hace apto para programas de fidelización. Por su parte, el **Meta-Cluster 3** (2,457 registros) presenta alta rotación y volumen en productos económicos,

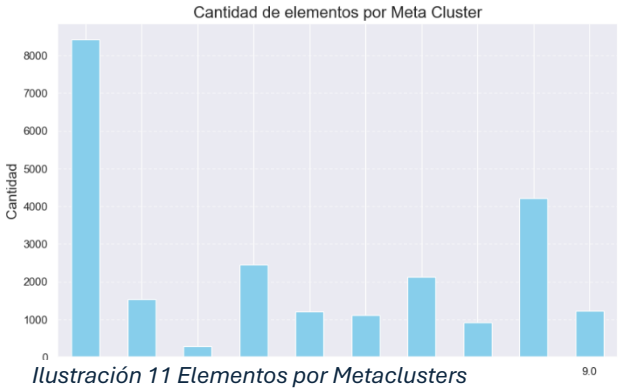

Ilustración 11 Elementos por Metaclusters

perfil típico de distribuidores, ideal para estrategias centradas en disponibilidad y eficiencia logística.

Se intentó replicar este análisis en el canal B2C, pero debido a la alta dimensionalidad y volumen de datos, los algoritmos no convergieron por limitaciones computacionales. Por ello, se priorizó el análisis de B2B, donde los datos permitieron resultados consistentes y aplicables desde el punto de vista comercial.

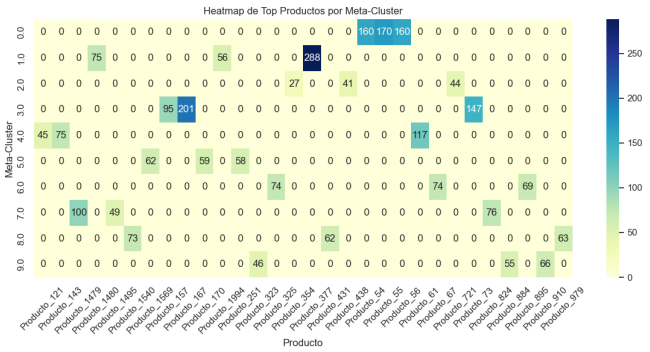

Ilustración 12 Distribución de productos de Metaclusters

METODOLOGÍA IMPLEMENTADA

Para seleccionar la metodología adecuada, partimos del reconocimiento de que se trata de un problema implícito, es decir, sin acceso directo a las preferencias declaradas por los clientes. Esto nos obligó a centrar los esfuerzos en comprender al usuario a través de su historial de compra y su contexto. Este enfoque se mantuvo a lo largo de todo el proceso, desde el enriquecimiento de datos hasta la elección de los algoritmos. Dada la alta dimensionalidad de las variables y el volumen considerable de registros, se concluyó que ningún modelo por sí solo sería capaz de capturar la complejidad del problema. Además, debido a las limitaciones de recursos computacionales, fue necesario dividir el reto en subproblemas más manejables, evitando así la construcción de un modelo monolítico sobredimensionado y propenso al sobreajuste.

Con esto en mente se decidió implementar un algoritmo de LightFM, uno de clasificación supervisada XGBoost y finalmente un modelo de recomendación Híbrido que pudiera representar una combinación ponderada entre los algoritmos, para de alguna forma dar otra voz al problema. Esta solución se desarrolló para ambos contextos, con ciertas diferencias en su implementación, pero bajo la misma lógica.

Algoritmo LightFM: Se optó por este algoritmo ya que es un modelo de recomendación híbrido donde se combinan las ventajas del **filtrado colaborativo** (útil debido a la cantidad de clientes y compras realizadas) junto con el uso de la información de contenido (como las características de los productos).

- Marco Teórico

Es un algoritmo basado en ***embeddings*****,** esto hace que las recomendaciones sean personalizadas, también con altas dimensionalidades. Se genera una función matricial que implementa un *ranking*. Para dicho *ranking* utilizamos la función WRAP, que penaliza los primeros lugares del *ranking*, útil en este contexto donde hay un grandísimo número de

productos a ranquear. Se genera un producto punto entre el producto comprado con los demás vectores (otros productos) junto con el aprendizaje colaborativo:

$$\hat{r}ui = \langle p_u, q_i \rangle + b_u + b_i$$

Donde, p_u es un vector de usuario, q_i es el vector de producto y $b_u + b_i$ son los sesgos aprendidos.

- **Implementación (se dividió en tres etapas):**
  a. *Construcción del Dataset*

Se mapearon los usuarios y los productos junto con sus características (debido a la alta dimensionalidad no fue posible usar todas las variables seleccionadas pero las que no se utilizaron en este modelo se usaron el en siguiente). Las variables seleccionadas fueron:

Tabla 1 Variables consideradas en cada modelo

Modelo	**Usuarios**	**Productos**
B2C	Clúster, municipio, asesor, punto de venta, zona	Categoría macro, subcategoría, color
B2B	Municipio, zona, Total de unidades, Total de edificaciones en obra	Categoría macro, categoría, subcategoría

  b. *Generación de Interacciones y Features*

Las interacciones se construyeron a partir del historial de compras. En el caso B2C, se utilizó una señal binaria (compra/no compra), mientras que en B2B se empleó el valor monetario total de compra como señal de intensidad. Se emplearon funciones personalizadas para construir listas de atributos por usuario y producto, que luego fueron vectorizadas mediante build_user_features() y build_item_features().

  c. *Entrenamiento del modelo*

El modelo se entrenó utilizando el algoritmo WARP con una configuración estándar de no_components=16 y epochs=5. Se utilizó num_threads=4 para paralelizar el entrenamiento y mejorar el rendimiento. Para cada cliente, se utilizó el método predict() sobre todos los ítems no comprados, y se ordenaron por *score* descendente. Se devolvieron los *Top N* productos recomendados.

Algoritmo XGBoost: Se seleccionó este algoritmo ya que le daba una perspectiva distinta al modelo generando un modelo de aprendizaje directo a partir de atributos explícitos de los usuarios y los productos.

- Marco Teórico

XGBoost (*Extreme Gradient Boosting*) es un algoritmo de aprendizaje supervisado basado en árboles de decisión. Combina múltiples árboles secuenciales, donde cada árbol corrige

los errores del anterior. Esto genera modelos precisos, robustos frente a *overfitting* (problema que ya estábamos experimentando con el anterior modelo) y eficientes en tareas de clasificación binaria, lo que buscábamos: predecir la probabilidad de que un cliente compre un producto. Este algoritmo optimiza una función objetivo que combina la pérdida logística con una penalización de regularización que controla la complejidad de los árboles:

$$\iota(\phi) = \sum_i \iota(y_i, \hat{y}_i) + \sum_k \Omega(f_k)\,, \qquad \Omega(f) = \Upsilon\mathrm{T} + \frac{1}{2}\lambda|\omega|^2$$

Donde, ι es la función de pérdida logística, $\hat{y}_i$ es la predicción del modelo y $\Omega(f_k)$ penaliza árboles con demasiadas hojas (T) o pesos grandes.

- Implementación

Se buscó implementar este clasificador binario que pudiera estimar la probabilidad de compra para cada cliente y cada producto, para esto se generaron también tres etapas:

a. Construcción del dataset Tabular

Se generaron combinaciones cliente-producto con etiquetas 1-0 de compra y no compra, generando así valores negativos. Dichos valores se balancearon con los positivos.

b. Enriquecimiento de variables

Para la base b2b fue necesario manejar las variables categóricas con un *Label Encoder* distinto a los *dummies* ya que estos aumentaban la dimensionalidad de manera exponencial. Finalmente, las variables seleccionadas fueron:

Tabla 2 Tabla de variables consideradas para el XGB

Modelo	**Usuarios**	**Productos**
B2C	edad, edad_promedio, ingreso_laboral_promedio, GINI, IPUG, cluster	Categoría macro, subcategoría, color, precio, alineación con portafolio estratégico
B2B	Municipio, zona, Total de unidades, Total de edificaciones en obra	Categoría macro, categoría, subcategoría

c. Entrenamiento

Se utilizó XGBClassifier con hiperparámetros conservadores, priorizando interpretabilidad y estabilidad del modelo. El entrenamiento se realizó sobre un *split* estratificado de 80/20, separando entrenamiento y evaluación final (*hold-out*). También se aplicó validación cruzada en 5 pliegues para garantizar la generalización del modelo.

Modelo de recomendación Híbrido

Teniendo la implementación de estos dos modelos complementarios, pero con limitaciones, ya que uno interpreta bien las preferencias implícitas (LFM) pero tiene dificultadas con la grande dimensionalidad, el otro balancea, pero tiene un alcance limitado de predicción. Por esto, se generó otro algoritmo de recomendación que combina las señales colaborativas y de contenido, ponderándolas para obtener un *score* final de recomendación más robusto.

Con la siguiente ecuación se busca establecer una combinación lineal entre ambos modelos,

$$score_{híbrido} = \alpha * score_{lfm} + (1 - \alpha) * score_{xgb}$$

Donde, α es un hiperparámetro ajustable entre 0 y 1 que controla la influencia relativa de cada modelo, $score_{lfm}$ es la predicción del modelo LightFM (colaborativo), $score_{xgb}$ es la probabilidad estimada por el modelo XGBoost (contenido).

A partir de unos productos candidatos, se manejan las predicciones de ambos modelos y ya con ambos scores resultantes se ajusta el peso α de la ecuación y se retorna el *ranking*. Dicho peso es ajustable ya que se considera que, a partir del análisis de los modelos, el equipo de analítica de Corona puede determinar el peso dependiendo cual modelo se ajusta de mejor manera al contexto y a las necesidades de la empresa retornando predicciones más precisas y coherentes.

Análisis Computacional

Desde una perspectiva computacional, la implementación del sistema de recomendación demostró ser eficiente, considerando las capacidades del equipo utilizado. El entrenamiento y compilación de los tres modelos (LightFM, XGBoost e híbrido) para el escenario B2B tomó **23.94 segundos**, mientras que para el escenario B2C (más complejo en términos de volumen y dimensionalidad) tomó **17.57 minutos**. Estas ejecuciones se realizaron sobre un MacBook Pro con chip Apple M1, 8 núcleos (4 de rendimiento y 4 de eficiencia) y 8 GB de memoria RAM, con los datos ya limpios y preprocesados. Los tiempos obtenidos validan que la solución puede ejecutarse en entornos de recursos moderados, lo que favorece su replicabilidad y despliegue operativo sin requerir infraestructura de alto costo.

VALIDACIÓN DE LA HERRAMIENTA

Para cuantificar el impacto real de la herramienta de recomendación desarrollada para Corona, se diseñó un proceso de validación estructurado, riguroso y alineado con el contexto operativo y estratégico de la compañía.

A partir de la base de cotizaciones, se identificaron clientes que presentaron tanto cotizaciones exitosas como fallidas (expiradas o cerradas). Esta selección permitió evaluar el comportamiento del modelo en escenarios reales de intención de compra, incluyendo tanto casos de conversión como de pérdida. La lista resultante se cruzó con la base transaccional

B2C enriquecida, asegurando que cada cliente contara con todas las variables necesarias para alimentar el modelo de manera completa, incluyendo dimensiones demográficas, comportamentales y estratégicas.

Sobre esta muestra se aplicó el sistema de recomendación híbrido. Para cada cliente se obtuvo el producto efectivamente comprado, el producto fallido, el producto recomendado por el modelo, su *score* de recomendación, el precio estimado del producto sugerido y su nivel de alineación con el portafolio estratégico de la compañía. Estos datos permitieron construir una base de indicadores clave que reflejan con precisión no solo la capacidad técnica del modelo, sino sobre todo su aporte directo al negocio. Se calcularon indicadores como el **valor esperado total de ingresos**, este es la suma ponderada del *score* por el precio de cada producto recomendado. Este valor representa una estimación del ingreso potencial si las recomendaciones fueran adoptadas, modelando así distintos escenarios de **retorno sobre la inversión (ROI)** y justificando la implementación del sistema desde una perspectiva financiera. También, se calculó el **valor esperado promedio por cliente**, lo que ayuda a estimar el beneficio incremental por usuario y evaluar la escalabilidad del sistema.

Desde una perspectiva estratégica, se midió la **tasa de alineación**, es decir, el porcentaje de recomendaciones que superan el umbral de 70% en su coincidencia con las prioridades del portafolio. Esta métrica permite asegurar que las recomendaciones no solo sean relevantes para el cliente, sino también coherentes con la dirección comercial de Corona. Complementariamente, se calculó la **tasa de mejora estratégica**, comparando la alineación del producto fallido con la del producto sugerido por el modelo. En la mayoría de los casos, el sistema logró proponer alternativas mejor alineadas, lo que evidencia su capacidad para corregir desviaciones y aumentar la efectividad comercial.

Finalmente, se analizaron la **distribución de los *scores* de recomendación**, útil para interpretar el grado de confianza del sistema y ajustar umbrales de acción, así como el **ranking de productos más recomendados**, lo que permite focalizar acciones de marketing, gestión de inventarios y optimización del portafolio en productos de alta tracción. Este proceso de validación, completamente anclado en datos reales, demuestra que el sistema de recomendación no solo es técnicamente sólido, sino que representa una herramienta con alto potencial de monetización, alineación estratégica y eficiencia operativa. Su implementación permite transformar datos en decisiones comerciales concretas, maximizando el retorno y fortaleciendo la ventaja competitiva de Corona.

MÉTRICAS E IMPACTO

Para medir la validez y la utilidad de la herramienta se calcularon las siguientes métricas específicas de rendimiento para cada modelo implementado.

AUC: Mide la capacidad del modelo para distinguir entre las clases, los productos que serán comprados y los que no.

Precisión: De los productos que el sistema predijo como "relevantes", ¿cuántos realmente lo fueron? Esto se mide con esta fórmula:

$$Precisión = \frac{Positivos}{Positivos + Falsos\ Positivos}$$

Recall: De todos los productos que son relevantes, ¿cuántos se recomendaron?

$$Recall = \frac{Positivos}{Positivos + Falsos\ Negativos}$$

F1-Score: Es una métrica que pondera y combina las dos anteriores.

$$F1 = 2 * \frac{Precisión * Recall}{Precisión + Recall}$$

Las métricas de los algoritmos implementados con los datos generados fueron:

Tabla 3 Métricas de todos los modelos construidos

	B2B				B2C			
	Enriquecida		No Enriquecida		Enriquecida		No Enriquecida	
	LightFM	XGBoost	LightFM	XGBoost	LightFM	XGBoost	LightFM	XGBoost
Preción	93.33%	80%	86.67%	77%	0.11%	84%	0.11%	83.05%
Recall	0.40%	81%	0.34%	76%	0.64%	86%	0.64%	80.64%
F1	0.80%	79%	0.67%	78%	0.50%	85%	0.50%	81.82%
AUC	87.79%	87.35%	85.49%	87.07%	93.62%	91.38%	93.62%	83.54%

La comparación de métricas evidencia que el enriquecimiento de datos mejora significativamente el rendimiento de los modelos en ambos canales. En **B2B**, LightFM incrementa su precisión de 86.67% a 93.33% y su F1 de 0.67% a 0.80% con datos enriquecidos, mientras que XGBoost mejora su recall de 76% a 81%, reflejando una mayor cobertura de productos relevantes. En **B2C**, el impacto es más evidente en XGBoost: la F1 sube de 81.82% a 85% y el AUC de 83.54% a 91.38%, lo que confirma una mejor capacidad para discriminar entre productos adecuados y no adecuados. Estas mejoras demuestran que integrar variables adicionales permite a los modelos hacer recomendaciones más acertadas, con mayor retorno potencial para Corona y que recomiendan de forma acertada y precisa productos a los distintos tipos de clientes.

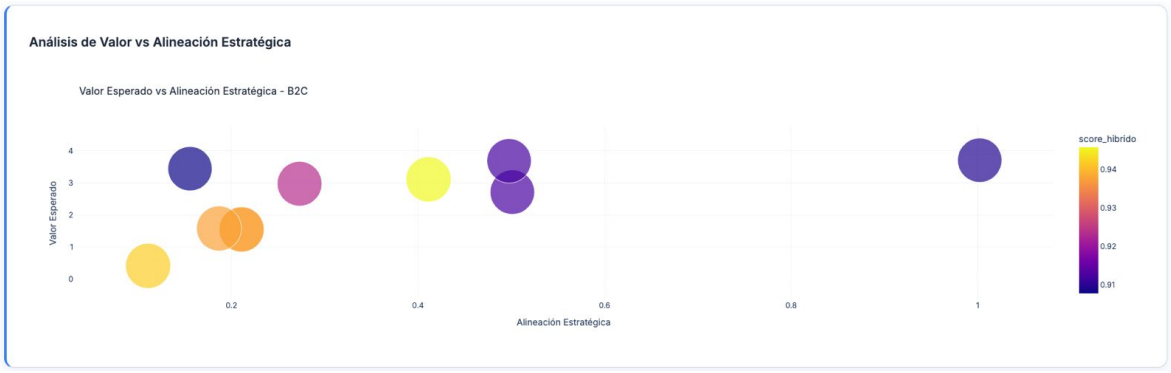

Ilustración 13 Relación entre Alineación Estratégica y Valor esperado

Esta métrica es valiosa porque conecta el valor económico potencial de una recomendación con su nivel de alineación estratégica. Permite identificar qué sugerencias del modelo no solo son rentables, sino también coherentes con las prioridades del portafolio de la compañía. Sirve para priorizar acciones comerciales: por ejemplo, enfocarse en los puntos que combinan alto valor esperado y alta alineación estratégica, optimizando tanto ingresos como dirección comercial. Además, el color indica el nivel de confianza (*score*), ayudando a decidir en qué recomendaciones confiar más.

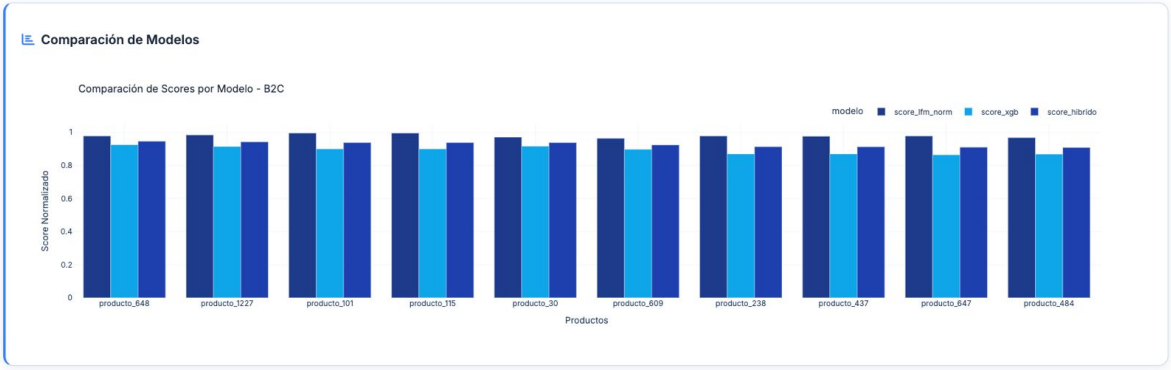

Ilustración 14 Productos recomendados, comparativa entre los modelos

Finalmente se visualiza el resultado esperado de la herramienta: las recomendaciones por cliente (B2B y B2C) con todas las respectivas estadísticas para que Corona tenga la información para recomendar el producto más adecuado a sus clientes y así cumplir con sus objetivos.

RECOMENDACIONES Y CONCLUSIONES

Los resultados obtenidos, junto con la arquitectura desarrollada y el proceso de validación realizado, constituyen una base sólida para que Corona avance hacia una estrategia comercial fundamentada en personalización inteligente y toma de decisiones basada en datos. La validación del sistema, aplicada sobre una muestra realista de clientes con cotizaciones ganadas y fallidas, confirmó no solo la capacidad técnica del modelo, sino

también su potencial para generar ingresos y fortalecer la alineación estratégica de las decisiones comerciales.

Concretamente, el modelo arrojó un **valor esperado total de ingresos de $1,328.82**, con un **promedio de $6.07 por usuario**, lo cual permite cuantificar de manera directa el retorno potencial que tendría la implementación de esta herramienta. Estos valores no solo permiten estimar escenarios de rentabilidad y retorno sobre la inversión (ROI), sino que sirven como base para priorizar el despliegue operativo del sistema.

Desde el punto de vista estratégico, el **54.34% de las recomendaciones** generadas por el modelo presentaron una alta alineación con el portafolio estratégico de Corona, lo que indica que el sistema no solo predice productos relevantes para el cliente, sino también coherentes con las metas comerciales de la compañía. Además, en el **23.74% de los casos**, el modelo logró recomendar productos con **mejor alineación estratégica que aquellos que fueron fallidos en intentos previos**, evidenciando su capacidad para corregir desviaciones y orientar la demanda hacia opciones de mayor valor estratégico.

Para maximizar el impacto de esta solución en el mediano y largo plazo, se recomienda fortalecer tres frentes clave: capacidad computacional, calidad de datos y profundidad analítica. En primer lugar, aumentar los recursos computacionales permitirá explorar y manejar modelos más complejos, adaptables y escalables. En segundo lugar, consolidar nuevas fuentes de datos, como la inclusión de señales de preferencias explícitas por parte del cliente, elevará la potencia predictiva y precisión del sistema. Y finalmente, profundizar en el análisis causal de las variables explicativas mediante segmentaciones dinámicas y experimentos controlados permitirá perfeccionar la comprensión de los factores que inciden en la decisión de compra.

En conjunto, el sistema validado no solo aporta precisión técnica, sino valor estratégico y financiero comprobado. Su implementación representa una oportunidad concreta para que Corona fortalezca su ventaja competitiva, optimice su gestión comercial y avance hacia una organización verdaderamente orientada a datos para la toma de decisiones.

PERSPECTIVAS FUTURAS DE NEGOCIO

La implementación del sistema de recomendación en Corona no representa un punto final, sino más bien un punto de partida hacia una transformación más profunda en la manera en que la empresa se relaciona con sus consumidores, asesores y aliados estratégicos. Las perspectivas futuras se estructuran en tres dimensiones complementarias: técnica, comercial y estratégica.

1. Perspectiva técnica: hacia una arquitectura evolutiva e inteligente: Desde el plano técnico, se abren múltiples líneas de desarrollo. Una de las prioridades es el refinamiento

continuo del modelo híbrido implementado. Esto implica optimizar el ajuste de hiperparámetros, incorporar nuevas funciones de pérdida adaptativas, e incluso explorar arquitecturas de metamodelos, que combinen los *outputs* de LightFM y XGBoost mediante redes neuronales o modelos de *boosting* adicionales. Siguiendo el ejemplo de empresas como Netflix, que ha evolucionado hacia un modelo fundacional centralizado capaz de transferir aprendizajes entre distintas tareas (Hsiao et al., 2025), Corona podría avanzar hacia una arquitectura unificada de recomendación capaz de adaptarse dinámicamente al canal, al contexto geográfico y al perfil del usuario.

Adicionalmente, la incorporación de aprendizaje continuo permitiría que el modelo se mantenga actualizado frente a cambios en el comportamiento de los usuarios, sin requerir reentrenamientos completos. Esto cobra especial relevancia en contextos como el B2B, donde los ciclos de compra pueden ser estacionales y los patrones de demanda más sensibles a eventos externos como licitaciones, obras o decisiones gubernamentales.

2. Perspectiva comercial: integración *real-time* y estrategias postventa: Desde una visión de negocio, el sistema puede escalarse más allá del canal analizado. Una evolución natural sería su integración en otros puntos de contacto como lugares de cotización web, canales de atención al cliente o aplicaciones móviles. Esto permitiría capturar información en tiempo real (clics, búsquedas, abandono de carrito) y retroalimentar el sistema con señales implícitas y explícitas para mejorar la precisión y personalización de las recomendaciones.

Otra línea clave es la implementación de estrategias postventa. Permitir que los clientes califiquen productos o experiencias abre la puerta a enriquecer los modelos con *feedback* explícito, habilitando la transición hacia modelos de aprendizaje reforzado o sistemas basados en satisfacción percibida. Empresas como Netflix han demostrado que redefinir sus métricas clave (de “vistas” a “minutos vistos”) permitió alinear mejor sus modelos con el éxito real de sus contenidos (O'Brien, 2024). De forma similar, Corona podría redefinir su métrica de conversión efectiva, incorporando no solo compra sino repetición, recomendación o satisfacción.

3. Perspectiva estratégica: un motor de decisiones inteligente y transversal: La visión de largo plazo posiciona este sistema como un motor de decisiones automatizado para múltiples áreas del negocio. Su alcance no se limita a ventas: puede guiar decisiones de surtido, campañas de marketing, gestión de inventario, expansión territorial o incluso desarrollo de producto. Por ejemplo, si el sistema detecta alta demanda de una categoría en una zona específica, puede sugerir ajustes en el abastecimiento o incluso en la estrategia de precios y promociones. En este sentido, el sistema trasciende su rol de recomendador y se convierte en un orquestador de valor.

Inspirándose en casos como el de Netflix, donde el sistema de recomendación no solo guía lo que el usuario ve, sino qué contenido se produce, cómo se lanza y dónde se promociona (Quin, 2024), Corona puede adoptar un enfoque similar en su ecosistema de productos y canales. La clave está en convertir los datos transaccionales y contextuales en inteligencia estratégica, no solo para personalizar experiencias, sino para anticipar movimientos del mercado y operar con ventaja.

Finalmente, como señala Ravindran (2023), los sistemas de recomendación efectivos no se construyen solo con buenos modelos, sino con claridad en los objetivos y en el impacto deseado. En el caso de Corona, el objetivo es claro: optimizar la oferta, personalizar la experiencia y maximizar ingresos. Pero el verdadero diferencial será lograr que este sistema evolucione al ritmo del negocio, incorporando cada nueva interacción como una oportunidad de aprender, adaptarse y servir mejor.

Anexos

1. Video YouTube del despliegue de la herramienta: https://youtu.be/2X_YzRFQ9DI
2. Presentación Ejecutiva: https://www.canva.com/design/DAGoqDP8fUo/mgCKgXHgmqi0_esi1bkW7w/edit?utm_content=DAGoqDP8fUo&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton
3. Organización en GitHub con la implementación e iteraciones de la herramienta: https://github.com/VVDataInsights

Bibliografía

Adhikari, S. (2019, 27 de febrero). *Building a Movie Recommendation Engine in Python using Scikit-Learn*. Recuperado de https://medium.com/@sumanadhikari/building-a-movie-recommendation-engine-using-scikit-learn-8dbb11c5aa4b

Cámara de Comercio de Bogotá (s.f). *Mercado laboral*. Org.co. Recuperado de https://www.ccb.org.co/informacion-especializada/observatorio/analisis-economico/mercado-laboral

Chandra, R. (2024, 3 de abril*). 9 machine learning algorithms for recommendation engines*. Daffodilsw.com. Recuperado de https://insights.daffodilsw.com/blog/machine-learning-algorithms-for-recommendation-engines

DANE. (s.f). *Empleo y desempleo*. Recuperado de https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo

DANE. (s.f). *Gran Encuesta Integrada de Hogares - GEIH - 2024*. Recuperado de https://microdatos.dane.gov.co/index.php/catalog/819/get-microdata

Datascientest. (2022, 30 de noviembre). *Machine Learning & Clustering: el algoritmo DBSCAN*. Formación en ciencia de datos. Recuperado de https://datascientest.com/es/machine-learning-clustering-dbscan

Hsiao, K.J., Feng, Y., & Lamkhede, S. (2025, March 21). *Foundation Model for Personalized Recommendation*. Netflix Tech Blog. https://netflixtechblog.com/foundation-model-for-personalized-recommendation-1a0bd8e02d39

Kavlakoglu, E. (2025, 10 de febrero). ¿Qué es el filtrado basado en contenido?. *Ibm.com*. Recuperado de https://www.ibm.com/mx-es/think/topics/content-based-filtering

KMeans. (s.f). Scikit-Learn. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Longo, M. (2017, 8 de septiembre). *The What, Why and How of Recommendation Systems*. Medium. Recuperado de https://medium.com/retargetly/the-what-why-and-how-of-recommendation-systems-810d98789f83

Nvidia. (s.f.). *Recommendation System*. Recuperado de https://www.nvidia.com/en-us/glossary/recommendation-system

O'Brien, C. (2024, 8 de mayo). *The Unstoppable Success of Netflix*. Digital Marketing Institute. Recuperado de https://digitalmarketinginstitute.com/blog/the-unstoppable-success-of-netflix

Ravindran, R. (2023, 1 de enero). *What are Recommendation Systems and How are Companies Using Them?* Recuperado de https://rishika-ravindran.medium.com/what-are-recommendation-systems-and-how-are-companies-using-them-a5b08ff4df42

RPubs - Clustering Jerárquico en R. (s.f). Rpubs.com. Recuperado de https://rpubs.com/mjimcua/clustering-jerarquico-en-r

Seeda, P. (2021, 31 de octubre). Towardsdatascience.com. Recuperado de https://towardsdatascience.com/a-complete-guide-to-recommender-system-tutorial-with-sklearn-surprise-keras-recommender-5e52e8ceace1/

Sierra, L. F. (2024, 3 de mayo). *CENU 2024 - inicio*. Todo Lo Que Necesita Saber Sobre El Censo Económico | DANE. Recuperado de https://censoeconomiconacionalurbano.dane.gov.co/

Quin, J. (2024, 25 de diciembre). *Netflix's Billion-Dollar Secret: How Recommendation Systems Fuel Revenue and Innovation*. Recuperado de https://www.linkedin.com/pulse/netflixs-billion-dollar-secret-how-recommendation-systems-qin-phd-7zece/

Villalonga, R. (2017, diciembre 1). *Sistemas recomendadores híbridos*. Medium. Recuperado de https://medium.com/@rvillalongar/sistemas-recomendadores-hibridos-93a6fff29500