Imagen que contiene Texto

Descripción generada automáticamente

Modelo de marketing para MODA SIN LÍMITE con metodología crisp-dm

Por:

Daniel Esteban González Zuluaga

Profesor a cargo:

Jorge Andrés Alvarado Valencia

Pontificia Universidad Javeriana

Maestría de Analítica para la Inteligencia de Negocios

Métodos y aplicaciones de analítica1

2024

**1. Business Understanding**

**1.1 Background**

La empresa 'Moda sin Limite' se encuentra en un momento crucial de su trayectoria comercial, enfrentando desafíos y oportunidades en un mercado dinámico y competitivo. Ante la necesidad de comprender mejor su entorno operativo y maximizar su potencial de crecimiento, 'Moda sin Limite' ha decidido contratar nuestros servicios de consultoría estratégica. Nos han encomendado la tarea de realizar un análisis exhaustivo tanto del entorno externo como de las capacidades internas de la empresa, con el objetivo de desarrollar estrategias efectivas para pronosticar las ventas anuales de su línea de ropa femenina. A través de marcos reconocidos como PESTLE, Fuerzas de Porter y SWOT, buscaremos identificar las tendencias del mercado, evaluar la competencia y descubrir las fortalezas y debilidades internas de 'Moda sin Limite', para así proporcionar recomendaciones sólidas que impulsen su éxito futuro.

El marco PESTLE ofrece una perspectiva holística de los factores externos que podrían influir en la industria de la moda. En el ámbito político, se debe considerar el impacto de las regulaciones comerciales y las políticas fiscales en las operaciones del negocio. Económicamente, es crucial evaluar las tendencias del mercado, las tasas de interés y el poder adquisitivo del consumidor, la moda en Colombia obtuvo un comercio en enero de 2024 de 2,73 billones de pesos, un 7,8% más que en 2023. En términos reales se registra un aumento de 2,56% y un crecimiento de 5,13% en los precios, lo que genera un buen momento para la inversión en textiles [1]. A nivel social, se deben analizar las tendencias culturales y sociales que puedan afectar las decisiones de compra, mientras que, en el ámbito tecnológico, se debe prestar atención a la digitalización de las ventas y la innovación en la cadena de suministro. Legalmente, se deben considerar las normativas relacionadas con la seguridad de productos y las leyes laborales, mientras que, en términos ambientales, es fundamental abordar la sostenibilidad y la gestión de residuos.

Por otro lado, el análisis de las Fuerzas de Porter proporciona una comprensión más profunda de la dinámica competitiva de la industria. La rivalidad entre competidores, determinada por la concentración de empresas y la diferenciación de productos, puede influir en la estrategia de precios y marketing. La amenaza de nuevos entrantes es otro aspecto relevante, ya que la facilidad de ingresar al mercado puede afectar la competencia existente. Además, el poder de negociación de proveedores y compradores, así como la amenaza de productos sustitutos, son factores que deben tenerse en cuenta al evaluar la posición competitiva de la empresa en el mercado, en este momento Gef, Punto Blanco, Leonisa, Tania, Cahet Tall, Laura, Touché y OndadeMar, en este orden de importancia, son las marcas colombianas que puntean las cifras [2].

**1.3 Business goal**

1. **Maximizar la rentabilidad y eficiencia operativa:** Optimizar la asignación de recursos para la línea de ropa de mujer. Al comprender mejor cómo las variables de los servicios ofrecidos impactan en las ventas, así la empresa puede asignar sus recursos de manera más efectiva, lo que se traduce en una mayor rentabilidad y eficiencia en las operaciones.

**1.4 Business success criteria**

1. **Margen de contribución por los servicios ofrecidos:** Este KPI permite medir el margen de contribución generado por cada servicio ofrecido en la línea de ropa de mujer. Al entender cómo cada servicio impacta en las ventas y la rentabilidad, la empresa puede asignar sus recursos de manera más eficiente, centrándose en aquellos servicios que generan mayores márgenes de contribución.
2. **Porcentaje de utilización de recursos:** Este KPI evalúa la eficiencia en la asignación de recursos al medir el porcentaje de tiempo y recursos dedicados a cada servicio ofrecido en comparación con su capacidad total. Un alto porcentaje de utilización indica una asignación eficiente de recursos, lo que contribuye a una mayor rentabilidad y eficiencia operativa en la línea de ropa de mujer.

**1.5 Data mining goal**

1. **Desarrollar un modelo de regresión lineal utilizando las variables proporcionadas**: Este objetivo implica la aplicación de técnicas de análisis estadístico para construir un modelo predictivo que estime las ventas de la línea de ropa de mujer basándose en las variables proporcionadas.
2. **Optimizar el modelo de regresión:** Este objetivo implica la selección y ajuste del modelo de regresión lineal, como la selección de variables o la regularización, con el fin de mejorar la precisión y la generalización del modelo.

**1.6 Data mining success criteria**

Desarrollar un modelo de regresión lineal utilizando las variables proporcionadas:

1. **Error cuadrático medio (RMSE):** Este indicador mide la diferencia entre los valores reales y los valores predichos por el modelo de regresión. Un RMSE bajo indica que el modelo tiene una buena capacidad de predicción de las ventas de la línea de ropa de mujer.

Optimizar el modelo de regresión:

1. **Selección de variables:** Este indicador evalúa la importancia de cada variable en el modelo de regresión. Utilizando técnicas como stepwise, se pueden identificar las variables más relevantes para mejorar la precisión del modelo.
2. **Validación cruzada:** Este indicador evalúa la capacidad de generalización del modelo de regresión. Al dividir los datos en conjuntos de entrenamiento y prueba múltiples veces y calcular el rendimiento del modelo en cada iteración, la validación cruzada proporciona una estimación más robusta del rendimiento del modelo.

**2. Data Understanding**

**2.1 Describe data**

En la base de datos se identificaron 1600 datos para training y 400 para test, además las variables correspondientes de la base datos son:

**idloc**: Es una variable única que representa el id de la tienda.

**edadloc**: Variable numérica que representa los años transcurridos de la tienda en la zona.

**correo**: Variable numérica que representa el número de catálogos enviados durante el año por correo.

**paginas:** Variable numérica que representa el número de páginas promedio de los catálogos.

**telefono:** Variable numérica que representa el número de líneas promedio abiertas para pedidos.

**impresa:** Variable numérica que representa la cantidad gastada en publicidad impresa.

**servicio:** Variable numérica que representa el número de representantes de servicio al cliente.

**nomina:** Variable numérica que representa el valor total de la nómina durante el año.

**idmercado:** Variable categórica que identifica un tipo de mercado que se desea atacar.

**tamamer:** Variable ordinal que indica el tamaño potencial del mercado, proyectado por las cifras de marketing.

**promo:** Variable categórica que identifica un tipo de promoción que se llevó a cabo durante el año.

**ropamujer:** Variable que numérica que representa las ventas de prendas para mujer en millones de pesos durante el año.

**2.2 Explore data**

Para explorar los datos proporcionados por 'Moda sin Límite', se realizaron análisis descriptivos univariados y multivariados sobre una partición de la base de datos que incluye 1600 datos para entrenamiento y 400 para pruebas. Estos análisis se llevaron a cabo con el objetivo de abordar los objetivos del análisis previo, priorizando precisión y confianza en los resultados obtenidos. Se examinaron características clave como la distribución de las variables numéricas y categóricas, las relaciones entre las variables y su posible impacto en las ventas de la línea de ropa femenina. Los resultados de este análisis proporcionarán una base sólida para el desarrollo de estrategias efectivas y la optimización de modelos predictivos para pronosticar las ventas futuras de la empresa.

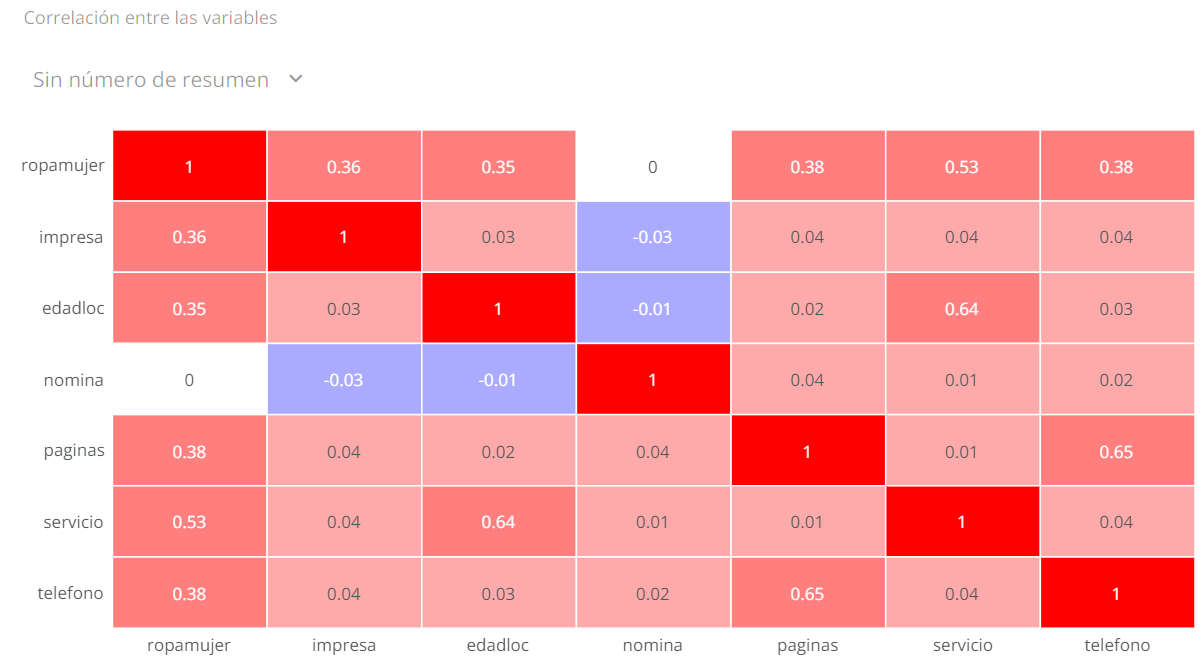
Para evitar sesgos en nuestros análisis, se enfoca exclusivamente en la base de datos de entrenamiento. La Gráfica 1 ilustra la relación entre la variable de promoción y las ventas de ropa para mujer (ropamujer), dado que se planea realizar una regresión lineal basada en la variable “promo”, más adelante se podrá entender mejor este contexto.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

*Gráfico 1. Comportamiento de variable promo vs ropamujer.*

Por último, en el contexto de este análisis, la matriz de correlaciones revela aspectos esenciales sobre la interrelación entre las variables en un modelo de regresión lineal. Cuando las correlaciones entre las variables son mínimas, es decir, cercanas a cero, esto indica una baja dependencia lineal entre ellas. Esta observación es fundamental, ya que sugiere que las variables predictoras son independientes entre sí, lo que puede conducir a estimaciones más precisas de los coeficientes de regresión. Esta independencia disminuye el riesgo de multicolinealidad, un factor que puede distorsionar los resultados del modelo. Por lo tanto, una matriz de correlaciones con valores bajos señala la robustez y la confiabilidad del modelo de regresión lineal, en el gráfico 2 podemos observar la matriz de correlación de las variables numéricas.



*Gráfico 2. Matriz de correlación*

Se nota que no hay correlaciones significativamente altas, lo cual es prometedor para llevar a cabo una regresión lineal sin preocupaciones de sobreajuste (overfitting) o multicolinealidad. Sin embargo, para asegurar la selección de las variables más relevantes en nuestra regresión, planeamos aplicar diversos métodos de análisis en etapas posteriores del estudio.

**2.3 Verify data quality**

Para verificar la calidad de los datos, es crucial identificar posibles valores atípicos, así como considerar posibles variables que no sirvan para la regresión. Los valores atípicos pueden manifestarse como observaciones que se desvían significativamente de la distribución esperada, lo que podría deberse a errores de medición, ingreso incorrecto de datos o eventos inusuales. Por otro lado, los valores perdidos pueden surgir por diversos motivos, como errores de registro, fallas en la recolección de datos o simplemente porque ciertas variables no se registraron para algunas observaciones. Es fundamental examinar detalladamente estas anomalías. Esta evaluación inicial de la calidad de los datos proporciona una base sólida para abordar y corregir posibles problemas, garantizando así la integridad y confiabilidad de la regresión. A continuación, en el gráfico 3 observamos el boxplot de cada variable numérica para observar si tiene datos atípicos.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

*Gráfico 3. Boxplot de cada variable numérica.*

En la gráfica presentada, no se observan valores atípicos o datos imposibles en las variables, lo cual es un indicador positivo para llevar a cabo la regresión lineal. Sin embargo, es importante tener en cuenta que aunque no se detecten anomalías en la distribución de los datos, aún podría haber puntos atípicos en el contexto de la regresión, es decir, observaciones que se encuentren notablemente alejadas de la línea de regresión. Es esencial considerar esta posibilidad durante el análisis y evaluar cómo estos puntos atípicos podrían afectar la precisión y la interpretación del modelo de regresión.

**3. Data preparation**

**3.1 Clean data**

Durante la limpieza de datos, se enfocó en identificar y eliminar valores atípicos que pudieran distorsionar la precisión del modelo de regresión. Se realizó una minuciosa revisión de las observaciones que se apartaban notablemente de la tendencia general de los datos, evaluando su posible influencia en el modelo. Aquellas observaciones con residuos estandarizados considerablemente altos fueron eliminadas, con el propósito de mejorar la calidad y confiabilidad del análisis. Se mantuvo un registro detallado y justificado de este proceso, asegurando que las decisiones estuvieran respaldadas por el contexto específico de los datos y los objetivos del análisis. En la Gráfica 4 se exhiben los residuos estandarizados de una regresión simple con el método stepwise, sin realizar alteraciones en la base de datos original. Se determinó que los residuos estandarizados superiores a 2 indicaban una discrepancia significativa, por lo que se optó por eliminar estas observaciones, dado que carecíamos de certeza o conocimiento para ajustarlas de manera adecuada.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

*Gráfico 4. Residuos estandarizado de una regresión con stepwise.*

**3.2 Construct data**

Se realizaron cambios significativos en el conjunto de datos con el objetivo de mejorar su utilidad y relevancia para el análisis final. Estos cambios incluyeron procesos de transformación de variables y la creación de nuevas variables mediante feature engineering. Cada modificación se llevó a cabo con trazabilidad y justificación. Las variables originales se ajustaron y combinaron de manera cuidadosa para capturar mejor las relaciones y patrones subyacentes en los datos. Este enfoque de construcción de datos busca optimizar la calidad y la capacidad predictiva de la base de datos final. A continuación, se mencionarán cada uno de los cambios.  
  
**One hot encoding:** Esta técnica se usa para transformar variables categóricas en donde cada categoría se convierte en una columna y los valores pasan a ser 1 o 0 dependiendo de que la fila haya tenido esa categoría, ahora bien, este método se usó en una variable “idmercado”.  
  
**Tamamer:** Esta variable es de tipo ordinal, sus valores son pequeño, median y grande, por lo que los convertimos en 1,2 y 3 respectivamente dependiendo de su tamaño.  
  
**Reducción de atípicos:** Además de la previa eliminación de atípicos se optó por usar la 3 formas de reducción de atípicos más conocidas que son la logarítmica, la raíz cuadrada y la inversa. En todas las variables numéricas se usó cada una de las opciones mencionadas para que el stepwise o la regularización escojan las más indicadas.  
  
**Servicio por teléfono:** Esta variable se creó a partir de saber cuántas personas estaban disponibles por cada servicio telefónico que había, el método de cálculo fue (servicio/telefono).  
  
**Regresión por promo:** Dado el conocimiento sobre regresión, decidimos dividir el conjunto de datos en tres subconjuntos basados en la variable 'promo'. Esta estrategia de clasificación previa podría contribuir a la reducción de la dimensionalidad y mejorar la capacidad predictiva de la regresión. Dividir el dataset de esta manera nos permite explorar y modelar diferentes comportamientos o tendencias asociadas con distintos tipos de promociones, lo que potencialmente mejora la precisión de nuestras predicciones.

Por último, la variable ‘idloc’ no se usó para este modelo de regresión ya que su valor no indica nada para el comportamiento de la variable ropamujer.

**3.3 Dataset description**

En este caso, como se mencionó anteriormente, disponemos de tres conjuntos de datos divididos según la categoría de promoción: para la promoción 1, contamos con 466 datos; para la promoción 2, con 695; y para la promoción 3, con 431. Es importante destacar que las variables en cada uno de estos conjuntos son idénticas, manteniendo consistencia en todo el análisis. En total, cada conjunto de datos consta de 34 variables, que incluyen todas aquellas mencionadas en la sección de construcción de datos. Este enfoque garantiza la coherencia y comparabilidad entre los conjuntos de datos, lo que facilita un análisis integral y riguroso.

**4. Modeling**

**4.1 Select modeling techniques**

Para el modelado de datos, se ha optado por utilizar principalmente técnicas de regresión lineal. Esta elección se basa en la capacidad de la regresión lineal para modelar la relación entre variables predictoras y una variable objetivo continua, como lo es las ventas de la línea de ropa femenina. La regresión lineal se considera apropiada debido a su simplicidad y facilidad de interpretación, lo que la hace adecuada para explorar relaciones lineales entre las variables en el conjunto de datos. Además, la regresión lineal permite analizar el impacto individual de cada variable predictora en la variable objetivo, lo que puede proporcionar información valiosa para la toma de decisiones. Por lo tanto, la elección de la regresión lineal se fundamenta en su idoneidad para el problema en cuestión y su capacidad para proporcionar insights claros sobre las relaciones dentro de los datos.

**4.2 Generate test design**

En este escenario, disponemos de tres conjuntos de datos, lo que implica que el proceso de entrenamiento puede variar entre ellos. En este caso, optamos por utilizar todos los datos disponibles en cada conjunto, aplicando tanto el método stepwise como la regularización. Esta elección se fundamenta en la posibilidad de que cada dataset presente diferentes parámetros o variables, o que ciertas técnicas se adapten mejor a conjuntos de datos específicos. El método stepwise se empleará para seleccionar de manera automática las variables más relevantes, con el fin de optimizar la precisión y generalización del modelo de regresión lineal en cada conjunto de datos. Asimismo, la validación cruzada se considera un componente esencial del proceso, ya que divide repetidamente los datos en conjuntos de entrenamiento y prueba para evaluar la capacidad de generalización del modelo en datos no vistos. Además, se aplicará regularización para controlar el sobreajuste del modelo, mejorando así su capacidad de generalización, lo cual es crucial en contextos con múltiples variables predictoras.

**4.3 Build model**

Para cada una de las promociones, se llevó a cabo un exhaustivo proceso de selección de variables predictoras y ajuste de hiperparámetros para determinar el modelo más adecuado. En el caso de Promo '1' y Promo '2', se encontró que el método más eficaz para identificar las mejores variables predictoras fue el enfoque stepwise. Este método demostró su eficacia al proporcionar un conjunto óptimo de variables para cada modelo utilizado. Por otro lado, para Promo ‘3’, se descubrió que el mejor enfoque para minimizar el error fue la combinación de validación cruzada y regularización, específicamente utilizando el método de regularización Lasso. Este proceso de selección y ajuste de modelos garantizó la utilización de la parametrización final más adecuada para cada contexto promocional, optimizando así el rendimiento predictivo de los modelos empleados.

**5. Evaluation**

**5.1 Assess model**

Teniendo en cuenta los objetivos y KPIs establecidos para maximizar la rentabilidad y eficiencia operativa en la línea de ropa de mujer, así como los objetivos y criterios de éxito en minería de datos, se realizaron las tres regresiones lineales ya mencionadas.  
   
El primer objetivo de minería de datos fue desarrollar un modelo de regresión lineal utilizando las variables proporcionadas. Este modelo se evaluó mediante el Error Cuadrático Medio (RMSE), donde un valor bajo indicaría una buena capacidad predictiva del modelo sobre las ventas de la línea de ropa de mujer.  
  
El segundo objetivo se centró en optimizar el modelo de regresión, incluyendo la selección de variables y la validación cruzada. La selección de variables permitió identificar las más relevantes para mejorar la precisión del modelo, mientras que la validación cruzada proporcionó una estimación robusta de su rendimiento.  
  
Estas regresiones lineales se realizaron con el fin de apoyar la maximización del margen de contribución por los servicios ofrecidos y el porcentaje de utilización de recursos, fundamentales para alcanzar los objetivos de rentabilidad y eficiencia operativa. A continuación, mencionaremos el error que obtuvimos en train y los resultados de cada dataset.

**Promo 1:** Para este dataset se uso stepwise, en la siguiente imagen se puede interpretar las variables que mas ayudan a predecir el valor de ropamujer.

Texto

Descripción generada automáticamente

Ahora bien, el error RMSE de este dataset fue de 9961, por último observamos en el gráfico 5 la relación entre los valores teóricos (eje horizontal) y los residuos estandarizados (eje vertical) de nuestro modelo.

Gráfico, Gráfico de líneas

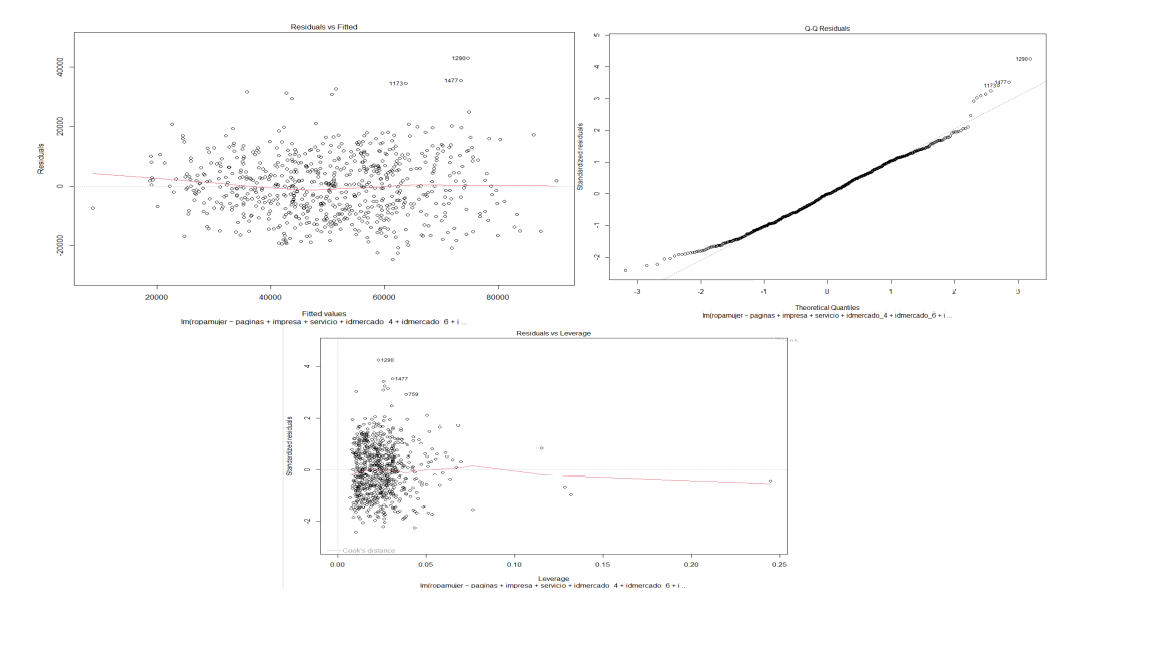
Descripción generada automáticamente

*Gráfico 5. Residuos vs valores reales.*

**Promo 2:** Para este dataset también se usó stepwise, en este caso fue el modelo que más error tuvo en la predicción con un RMSE de 10113. A continuación, podemos observar las variables predictoras y los gráficos residuales.

Texto

Descripción generada automáticamente



*Gráfico 6. Imágenes para comprension de errores.*

Podemos observar que en esta imagen los residuos si se alejan un poco más de la línea de regresión, no obstante, no observamos patrones y también no hay puntos influyentes que entren en la distancia de cook.

**Promo 3:** Este dataset fue el único que mejoró con la regularización obteniendo un error RMSE de 9753. A continuación, observamos en el gráfico 7 en que lambda se obtiene su mínimo error, vemos que la disminución de error no es sustancial pero se reduce, también observamos las variables más influyentes en ropamujer.

Interfaz de usuario gráfica

Descripción generada automáticamenteTabla

Descripción generada automáticamente con confianza media

*Gráfico 7. Error a partir de lambda y variables más influyentes.*

Para finalizar hay que tener en cuenta que para los datos de test también hicimos la misma separación de la variable promo y así poder entrenar cada modelo según su tipo de promoción.

**5.2 Produce final report**

Este informe gerencial tiene como objetivo presentar las conclusiones y mejoras derivadas de este proyecto. Comenzaremos identificando los beneficios de conocer las variables predictoras de cada tipo de promoción, si se quiere tener más información sobre cuáles son las variables por promoción se encuentran en la sección anterior:

1. **Mejor comprensión del proceso de ventas:** Identificar las variables que más influyen en las ventas de ropamujer, ayudará a comprender mejor el comportamiento de los clientes y los factores que impulsan las ventas. Esto te permitirá ajustar tus estrategias de marketing y operativas de manera más precisa.
2. **Optimización de recursos:** Al centrarte en las variables más predictoras, se pueden dirigir los recursos hacia áreas que realmente impacten en las ventas. Esto permite optimizar presupuesto y esfuerzos, evitando gastos innecesarios en áreas que tienen poco impacto en los resultados.
3. **Mejora en la toma de decisiones:** Contar con información sólida sobre las variables más importantes permite tomar decisiones más informadas y estratégicas.

Según estos beneficios proponemos las siguientes estrategias, basándonos en las variables más importantes como servicio, impresa y páginas:

1. **Mejora en la atención al cliente:** Dado que la variable "servicio" (cantidad de personal de atención al cliente) es una de las más predictivas, una estrategia clave podría ser mejorar la calidad y disponibilidad del servicio al cliente. Esto implica capacitación continua para el personal, implementación de sistemas eficientes de gestión de consultas y quejas, y una atención personalizada que responda a las necesidades específicas de los clientes.
2. **Optimización de la publicidad impresa:** Considerando que la variable "impresa" (inversión en publicidad impresa) es otra de las más influyentes, una estrategia importante sería optimizar el uso de la publicidad impresa. Esto podría incluir una evaluación cuidadosa de los canales de publicidad más efectivos, el diseño de anuncios atractivos y persuasivos para maximizar el retorno de la inversión en publicidad.
3. **Enfoque en contenido y diseño de catálogos:** Dado que la cantidad de páginas en los catálogos también es una variable significativa, una estrategia efectiva sería centrarse en mejorar el contenido y el diseño de los catálogos. Esto implica ofrecer una variedad atractiva de productos, presentados de manera visualmente atractiva y con información clara y relevante para los clientes. Además, se podría explorar la posibilidad de personalizar los catálogos según las preferencias y el historial de compras de cada cliente para aumentar su relevancia y efectividad.

Vale la pena destacar que el error promedio en train de todas las promociones fue de 10617, lo que sugiere la posibilidad de un margen de error considerable si se busca precisión absoluta en las predicciones. No obstante, es importante considerar que estos modelos nos brindan una estimación generalizada de las ganancias potenciales basadas en las variables conocidas de la empresa. Por ejemplo, si obtenemos un resultado de 70000 y tenemos en cuenta el error promedio, sabemos que nuestra cifra estimada oscila entre los 60000 y 80000, lo que sigue siendo un indicativo sólido de un negocio favorable.

**Bibliografía**

**1.** <https://www.inexmoda.org.co/wp-content/uploads/2024/02/Observatorio-Inexmoda-Febrero-2024-F.pdf>  
**2.** <https://pe.fashionnetwork.com/news/Colombia-ocho-empresas-manejan-la-industria-de-la-ropa-interior-,673137.html>