Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Equipo 4

"Momento de Retroalimentación: Reto Metodología"

Integrantes:

Eliezer Cavazos Rochin A00835194

Facundo Colasurdo Caldironi A01198015

Saul Francisco Vázquez del Río A01198261

José Carlos Sánchez Gómez A01174050

Campus Monterrey **Fecha: 22/11/2024**

Índice

Introducción

Actualmente, se vive en una sociedad donde el aumento de la población, ha generado un incremento masivo en el consumo de productos, debido a la constante y creciente demanda de los mismos. En este contexto, Arca Continental, quien se puede considerar como una de las embotelladoras más grandes de América Latina, enfrenta el reto de adaptarse a estos desafíos cambiantes, optimizando su producción y estrategias comerciales para satisfacer las necesidades de sus clientes.

Planteamiento del problema

Arca continental planteó el reto de poder identificar de manera confiable los clientes potenciales para sus nuevos productos de lanzamiento.

Objetivo

Identificar potenciales clientes cuyas preferencias se ajusten a las características de los productos de lanzamiento.

Subobjetivos

- Segmentar a los clientes en función de sus patrones de compra, permitiendo una mayor personalización de estrategias de marketing y ventas.

- Agrupar los productos según su rendimiento en el mercado, con el fin de optimizar la oferta y la disponibilidad de los mismos en función de su demanda.

Objetivo

Predecir la venta de los próximos cinco meses de un producto nuevo por características y por nivel socioeconómico.

Subobjetivos

- Generar modelos de predicción para predecir la venta de un producto nuevo de manera general

- Usando los clusters de clientes por nivel socioeconómico predecir la venta de un nuevo producto por cada cliente, para ver los comportamientos de este nuevo producto en los diferentes clientes que se tienen.

Antecedentes y proyectos relacionados:

La inteligencia artificial (IA) ha transformado múltiples sectores, desde el comercio minorista y el entretenimiento hasta la salud y el transporte, facilitando el análisis y la predicción de patrones de comportamiento de los consumidores. No obstante, el uso de IA también plantea desafíos éticos y normativos, particularmente en cuanto a la privacidad, la transparencia de los modelos predictivos y el manejo responsable de los datos.

La aplicación de IA en el análisis de consumo y ventas requiere una consideración cuidadosa de estos aspectos, ya que el uso de datos personales y de consumo debe ajustarse a regulaciones de protección de datos, como el Reglamento General de Protección de Datos (GDPR) en Europa y sus equivalentes en América Latina.

Desde una perspectiva ética, el uso de la IA tiene grandes responsabilidades esto debido a que implica que se manejen los datos de manera segura y que se cumplan con las normativas internacionales de seguridad, ademas, demás, es fundamental promover la transparencia algorítmica, explicando a los usuarios cómo se utilizan sus datos y cómo funcionan los modelos de predicción, ya que esto no solo genera confianza en los usuarios, sino que también evita posibles malentendidos en el uso de la información.

Durante el desarrollo del proyecto se decidió que se va a segmentar los clientes por patrones de compra para facilitar la recomendación de productos de lanzamiento parecidos a su patrón de compra y también segmentar los productos por sus características principales.

Entre las herramientas relevantes para realizar esta segmentación se encuentran los algoritmos de clustering, como el k-means, en donde el uso de k-means ha demostrado su eficacia en la identificación de patrones clave en el comportamiento del consumidor dentro del sector minorista, permitiendo personalizar estrategias de marketing como se describe en un análisis práctico de segmentación disponible en la investigación realizadas por la analista Jennifer salazar: “Por otro lado, se considera de suma importancia efectuar una segmentación de clientes buscando detectar aquel segmento que genera mayor rentabilidad para el Banco. “

Asimismo, los modelos de clasificación, como los árboles de decisión y las redes neuronales, ayudan a predecir las probabilidades de que un cliente opte por un producto específico, proporcionando insights valiosos sobre sus preferencias. El análisis de patrones secuenciales es otra técnica

que permite estudiar el orden y frecuencia de las compras, apoyando así en la creación de recomendaciones más efectivas.

En el contexto específico de este proyecto, la IA será utilizada para analizar patrones de consumo de los clientes de Arca Continental y de esta forma, apoyar la toma de decisiones estratégicas que maximicen las ventas de productos de lanzamiento.

Para entender de mejor maner a Arca Continenta, es necesario comprender que es una de las embotelladoras más grandes del mundo, la cual se caracteriza por usar estrategias avanzadas de mercado para sus productos, según el Reporte de Estrategias de Segmentación 2023, la empresa aprovecha los datos de consumo para diseñar campañas altamente efectivas.

A través de la segmentación de clientes y el agrupamiento de productos según su rendimiento en el mercado, el proyecto busca maximizar las ventas de nuevos productos.

Herramientas y recursos a usar:

Para el desarrollo del proyecto se utilizaron una gran variedad de herramientos que no solo facilitaron el análisis de los datos, sino que también, ayudaron a la construcción de los modelos predictivos, estas siendo: PowerBi, Pandas, Matplotlib, Tensor Flow, SKLearn, google colab, drive, los datasets dados por Arca Continental

Power bi fue utilizado durante la exploración inicial para poder mostrar los resultados de nuestra investigación de una manera más visual y sencilla.

Por otra parte, Pandas fue una herramienta esencial para lograr analizar los datos durante el proyecto, ya que esta nos permite trabajar con los datos dentro de python que nos permiten una mejor interacción con los datos de Arca Continental, logrando así eliminar datos innecesarios y logrando manejar grandes cantidades de información de manera eficiente.

SkLearn fue una librería de python usada para la clusterización de los datos tanto de los clientes como de los productos para poder agrupar y segmentar por los parecidos que habían entre los registros.

Tensor Flow fue usado para poder crear modelos predictivos, logrando construir árboles de decisión y modelos de inteligencia artificial usadas dentro del proyecto, con la finalidad de ser usadas para obtener los patrones de

comportamiento de los clientes y las compras de los productos, logrando de esa manera predecir qué productos serán exitosos y con qué grupos de clientes.

Google Colab y Google Drive, fueron de vital importancia, ya que nos permitieron almacenar los datos y los resultados del proyecto, a su vez, que nos facilitaron la carga de trabajo al permitir a todos los miembros del equipo poder trabajar de una manera más organizada.

Metodología

Este proyecto se divide en cinco diferentes etapas que pasan desde el entendimiento de los datos, la limpieza de datos, creación de transformaciones, clusterización/segmentación y predicciones.

1-. Exploración de los datos

Cada miembro de equipo fue explorando los datos proporcionados por Arca Continental, lo que más nos llamó la atención fue el dataset de clientes ya que este contenía muchas tablas con diferentes tipos de información como los niveles socioeconómicos alrededor del cliente, las zonas que estaba cerca del cliente en unos 300 metros, si cerca del cliente hay parques, supermercados, escuelas, hospitales, además de esto había varios indicadores de velocidad de lo que creíamos que eran los camiones de Arca continental dejan los productos a los clientes, siguiendo en los datos que había columnas que nos indican si los clientes eran tiendas, si estos eran nos indican sus gatos, que tantos autos pasaban, que población estaba más cerca al igual si tenían ocupaciones.

Una vez que terminamos de explorar los datos de clientes nos surgieron varias preguntas como ¿Necesitamos tanta información para el proyecto?, ¿Podemos quitar columnas del dataset?, ¿Podemos hacer más compacto el dataset?, etc… estas preguntas la mayoría fueron contestadas por los profesores que nos indicaron porque camino ir, al igual que socio formador contestando preguntas que tuvieran que ver con el entendimiento de los datos. Una vez establecidas las preguntas no fuimos a explorar los otros dataset de ventas y de productos pero en estos no teníamos tantas preguntas como en el dataset de clientes ya que este fue el dataset con más información que se nos presentó. Decidimos usar la herramienta de Power bi para tener una mejor visualización de los datos para tener una toma de decisiones de qué haríamos con las columnas y el cómo los usuarios para entrenar el modelo

2-. Limpieza de datos

Para nuestra limpieza de datos, decidimos seguir pasos para los tres dataset que se nos proporcionaron está siendo.

- Resaltar variables que creíamos importantes
- Buscar valores con NaN
- Buscar relaciones entre los datasets

Primeramente se empezó en el dataset de clientes resaltando las variables de sub_canal_economico, población alrededor de 300 m, status económico, zonas alrededor de 300m y los nivel de ocupación de las personas alrededor, luego se buscaron los valores que tuvieran NaN por el momento se decidió que esos valores con NaN serían sustituidos con el valor numérico 0 para no tener que borrar los datos que si tuvieran informacion. Por último se decidió buscar la relación entre algún dataset que tuviera la variable de CustomerId para hacer la conexión.

Ante esto el dataset que tuvo una conexión con la variable de CustomerId fue el de ventas entonces decidimos hacer lo mismo que en el dataset de clientes, ahora en este dataset todas las variables tuvieron importancia así que pasamos a lo siguiente buscar los valores con NaN sustituyendolos con el valor numérico de 0. Antes de buscar la siguiente conexión entre los datasets nos dimos cuenta que al conectar el dataset de clientes con el de ventas vimos que había clientes que no estaban en el dataset de ventas.

Esto indicaba que no habían realizado una compara, decidimos preguntarle los profesores y al socio formador sobre esto y se nos recomendó que los identificamos y los borraremos ya que no aportan para la solución que estábamos buscando, una vez identificado los clientes y haberlos borrado del dataset de clientes ahora sí se buscó la forma de conectar el dataset de productos con el ventas.

Se logró identificar la variable de Material en ambos datasets de ventas y productos logrando hacer la conexión de ambos, una vez realizado la conexión paso el mismo suceso de que en la tabla de ventas había productos que no estaban en el dataset de productos. Con las recomendaciones anteriores de los profesores y el socio formador se decidió hacer lo mismo de

identificar las ventas de los productos que no estaban en el dataset de producto y para el dataset de ventas se borraron las ventas de productos sin relación con el dataset de productos.

Una vez resuelto el problema de las ventas de productos, se identificaron las variables que pensamos que eran importantes en el dataset de productos como su tipo, envase, tamaño, sabor, marca y categoría. Además de buscar los valores con NaN y sustituirlos con el valor numérico de 0.

Terminado la limpieza de datos mostramos el avance al salón y al profesor y se nos comentó que quitamos los clientes con compras a partir de Septiembre 2022. Y por petición del socio formador decidimos también quitar todas las transacciones que iniciaron en 2019.

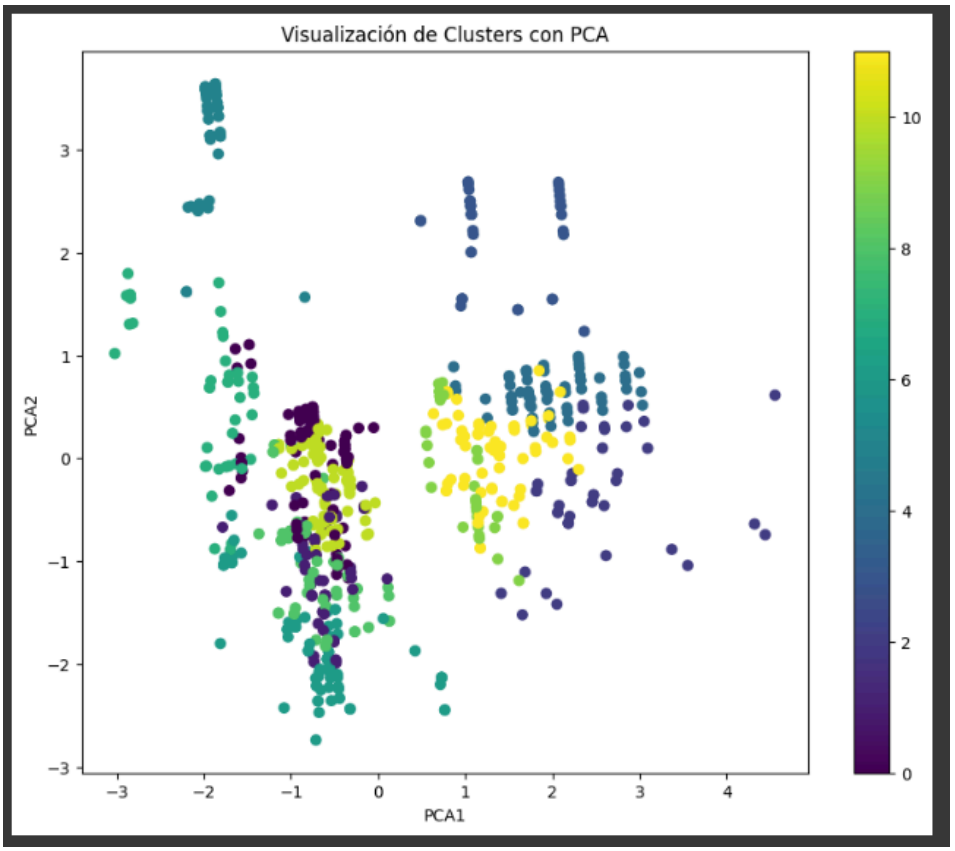
3. Transformación de datos

Se optó por un modelado estrella de datos para separar las categorías de los clientes en diferentes dimensiones, al igual que las categorías de los productos para la optimización y calidad de los datos. Una vez tomada la decisión se crearon las diferentes tablas con las características de los data frame de clientes dimensionado las columnas de zona, nivel Socioeconómico y Subcanal Para la tabla de productos se dimensionan las columnas de marcas, contenedor del producto, tamaño del envase, si es retornable o no, la categoría del producto y el tipo del producto.

Además de esto se crearon las tablas de hechos qué son las métricas que deseamos medir y analizar, se generaron 3 tablas de hechos la primera tabla de hechos se creó para visualizar cuanto compraba cada cliente por categoría de producto y la segunda tabla se creó para ver cuantas variaciones de producto compraron cada cliente y la última que es la más recientes es usada para identificar los productos exitosos de cada cliente.

4-. Selección y entrenamiento de modelos

Se realizó la clusterización de clientes y productos, esto nos ayudó a determinar en que cluster iba a ir los productos de lanzamiento: el cluster de productos siendo el primer filtro que determina en qué categoría irá el producto de lanzamiento, en este clusters se usaron las columnas de contenedor, categoría del producto, el tamaño de su empaque, sabor, productos por empaque y si este es retornable o no. Para determinar la categoría del producto teniendo un total de diez clusters.



Para la clusterización de clientes se optó por la creación de un cluster de clientes determinando su nivel socioeconómico usando las columnas de nivel socioeconómico y de zonas alrededor a 300 metros teniendo tu total de tres clusters. Además de esto se crearon sub clusters en donde se muestran

los patrones de compra de cada clase viendo cuáles productos compran más y cuales productos compran variedad de este usando las columnas de las categoría de los productos creando un total de doce sub clusters.

Se segmentan los clusters creados previamente para identificar mejor los clusters tanto de productos como de clientes para identificar que se está agrupando en estos clusters.

5-. Predicción de Clientes Potenciales para nuevo producto

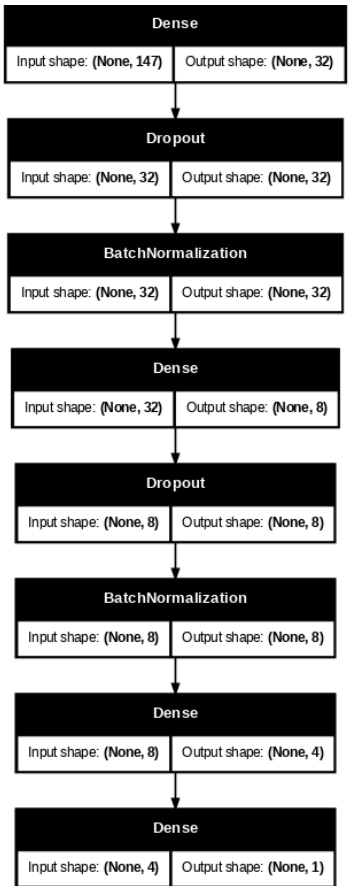
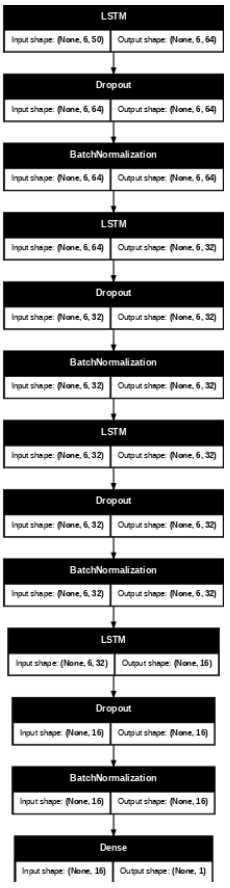
Para la predicción, utilizamos la librería de python tensor flow para crear de modelos de inteligencia artificial que sean capaces de predecir la venta de un nuevo producto en los próximos cinco meses, además utilizando los clusters que se generaron con k means determinar los clientes a los que más les vendería dependiendo del cluster de producto donde entre el nuevo producto, y también analizar cómo se desempeñaría la venta del producto dependiendo del nivel socioeconómico.

Estos modelos son redes neuronales, las cuales cuentan con capas de Dropout y BatchNormalization para garantizar que el modelo pueda abstraer

la información del conjunto de datos y no se sobreajuste. Para la función de pérdida se usa la función huber, la cual es menos sensible a los datos atípicos que la de error cuadrado promedio. De igual manera usamos las funciones EarlyStopping y ReduceLROnPlateau para asegurarnos que el modelo pudiera obtener un valor menor en la pérdida de los datos de validación.

La entrada del modelo tiene un formato diferente al que tienen nuestros set de datos. Utilizamos la técnica de OneHotEncoding para poder utilizar nuestras variables categóricas sin darles un peso mayor a una que la otra. Para nuestras variables numéricas, hicimos un escalado de estas con el objetivo de que estuvieran en un rango más limitado, y fuera más sencillo para el modelo aprender de este.

Nuestra solución ofrece dos modelos, para predecir valores diferentes. El modelo 1 (uno) se enfoca en predecir la venta promedio de un artículo en un mes en específico, independientemente del cliente. El modelo 2 (dos) tiene un propósito similar. Este último predice la venta promedio en un mes que va a tener un producto con un cliente en específico.

Arquitectura Modelo 1	Arquitectura Modelo 2
	

La razón por la que decidimos optar en el segundo modelo por un modelo de LSTM (Long Short Term Memory) es debido a la temporalidad que había en los datos, y la cantidad de estos que habían. Hicimos una prueba comparativa entre una arquitectura similar a la del Modelo 1 para el problema de predecir la venta promedio de un producto en un mes con un cliente específico, sin embargo el modelo de LSTM tuvo un mejor rendimiento. Sucedió el caso contrario con el problema de encontrar la venta promedio en un mes de un producto. La primera arquitectura tuvo un mejor desempeño que la de LSTM.

Modelo	Error Cuadrático Medio (MSE) - Problema 1	Error Cuadrático Medio (MSE) - Problema 2
Modelo1	8.31	18.43
Modelo 2	10.89	15.26

Resultados

Los resultados que nos han dado nuestras predicciones de clientes recomendados ha ido mejorando pero aún no está dando los resultados esperados. Usando los productos de pruebas que se nos proporcionaron donde se podrá ver el comportamiento que tienen nuestras predicciones en la siguiente tabla comparativa, además dejamos de considerar lo clientes exitosos para identificar las métricas ya que muchas veces estos valores no contaba muchos clientes que compraban mucha cantidad de estos productos:

Predicciones anterior:

Producto	Clientes Reales	Clientes Exitosos Reales	Clientes Propuestos	Clientes Exitosos que coinciden
MONSTER ENERGY 473 ML NO RETORNABLE	176 Clientes compraron este producto en los últimos 5 meses	Solo hay un cliente exitoso de este producto	Se proponen 471 clientes	De los clientes que se proponen ninguno coincide con el único cliente exitoso de este producto.
TCH HARD S	126 clientes	Solo hay un	Se proponen	De los clientes que

LIMA I 355 ML NR LAT 6 REFO	compraron este producto en los últimos 5 meses	cliente exitoso de este producto	471 clientes	se proponen ninguno coincide con el único cliente exitoso de este producto.
COCA COLA S/A MARS 355ML NR LSL 6B	295 clientes compraron este producto en los últimos 5 meses	No tiene ningún cliente exitoso	Se proponen 471 clientes	Como no tiene producto exitoso se compara cuantos clientes en general coinciden de lo propuesto de lo real que fueron 113 clientes

Predicciones nueva versión:

Producto	Clientes Reales	Clientes Exitosos Reales	Clientes Propuestos
MONSTER ENERGY 473 ML NO RETORNABLE	176 Clientes compraron este producto en los últimos 5 meses	Solo hay un cliente exitoso de este producto	Se proponen 491 clientes
TCH HARD S LIMA I 355 ML NR LAT 6 REFO	126 clientes compraron este producto en los últimos 5 meses	Solo hay un cliente exitoso de este producto	Se proponen 486 clientes
COCA COLA S/A MARS 355ML NR LSL 6B	295 clientes compraron este producto en los últimos 5 meses	No tiene ningún cliente exitoso	Se proponen 486 clientes

Con estas métricas podemos identificar que el modelo aún se tiene que mejorar para que los Clientes que se recomienda coincidan más con lo real. Las métricas de evaluación que estamos usando es identificar cuántos clientes recomendados por el modelo coinciden con la venta real del producto.

Conclusiones

En este proyecto se buscó implementar herramientas de analitica de datos y herramientas de inteligencia artificial para identificar clientes potenciales para nuevos productos de Arca Continental.

Aunque las predicciones muestran clientes potenciales para los nuevos productos, la predicción no es tan satisfactoria tenemos en mente mejoras que podrían ayudar a la mejora de este modelo de predicción de clientes potenciales. Cómo optimizar los sub clusters para generar mejores agrupaciones, mejorar el filtrado de los clientes por cluster de producto.

El uso de las herramientas como PowerBI, Pandas, SKLearn y Tensor Flow demostraron ser esenciales para la limpieza, análisis, segmentación y modelado de datos. Estas herramientas ayudaron para el procesamiento de grandes volúmenes de datos.

En conclusión, en este proyecto se nos mostró la importancia de la inteligencia artificial para la toma de decisiones estratégicas para el lanzamiento de productos. Sin embargo quedan mejoras para realizar para optimizar las predicciones y garantizar que estas sean más precisas y efectivas. Implementar estas mejoras propuestas nos permitirá realizar una solución que maximice la rentabilidad y satisfacción de los clientes con los productos de lanzamiento.

Referencias:

Arca. C. (2024) Balance General de Arca Continental. Recuperado de https://www.arcacontal.com/media/373544/2020_ac_consolidated_financial_statements.pdf

SalesForm. C.(2020) Clústeres: ¿qué son y para qué sirven?. Recuperado de https://www.salesforce.com/mx/blog/clusteres/