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: 1/11/2024**

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 cluster de clientes, 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.

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 y el análisis jerárquico, que agrupan a los consumidores en categorías homogéneas. 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. 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, entendiendo lo que significa cada columna tanto su nombre y lo datos que esta contiene, al igual que se usó la herramienta de Power bi para tener una mejor visualización de los datos.

2-. Limpieza de datos

Para nuestra limpieza de datos, primeramente se empezó rellenando todos los valores NaN con valor numérico de 0 de los datos de clientes por Arca Continental. Además de esto en el data frame de clientes identificamos y borramos los clientes sin ninguna compra en la tabla de ventas y se quitaron los clientes con compras a partir de Septiembre 2022. Para la tabla de ventas se identificó y quitaron las ventas de productos sin relación con la tabla de productos.

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-. Clusterización y Segmentación de Datos

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 quince 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 ocho subclusters.

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 kmeans 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 sobreajusten. 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.

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/