Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos II

Evidencia de Reto con Arca Continental

Edgar González Fernández
Mauricio González Soto

Autores:
Cleber Gerardo Pérez Galícia - A01236390
Juan Pablo Bernal Lafarga - A01742342
Jacobo Hirsch Rodríguez - A00829679
Eryk Elizondo González - A01284899

Noviembre 2024

Introducción

Arca Continental es la segunda embotelladora de Coca-Cola más grande de América Latina y una de las más importantes del mundo. La compañía no solo se especializa en la producción y distribución de bebidas gaseosas, sino que también ofrece una amplia gama de productos, como aguas, jugos y bebidas energéticas. Su compromiso con la sostenibilidad, la innovación y la responsabilidad social la ha posicionado como un líder en la industria de bebidas, contribuyendo al desarrollo económico de las regiones en las que opera.
En el competitivo entorno del mercado de bebidas, el análisis predictivo se ha convertido en una herramienta crucial para el éxito de las empresas. La ciencia de datos desempeña un papel fundamental en este proceso, ya que permite a las organizaciones como Arca Continental analizar grandes volúmenes de datos históricos de ventas, así como datos demográficos y de comportamiento del consumidor con el propósito de realizar predicciones de las ventas alrededor de nuevos y novedosos productos de futuro lanzamiento en el mercado.

Objetivos

El objetivo de este análisis es proporcionar a Arca Continental un conocimiento profundo sobre el desempeño de sus nuevos productos dentro de su base de clientes históricos. Mediante el análisis del historial de ventas de años anteriores, se busca identificar patrones de comportamiento de los clientes que han demostrado una adopción favorable hacia productos de la familia CocaCola. Además, se pretende construir un perfil de cliente óptimo basado en características de consumo y de fidelidad, para identificar a aquellos con mayor probabilidad de adoptar exitosamente productos nuevos.

Nuestro enfoque se centra en determinar qué características constituye un "producto exitoso" mediante el análisis de estos con técnicas de análisis de datos y machine learning , definido como aquel que un cliente consume en los primeros dos meses de prueba y continúa comprando al menos una vez al mes durante los tres meses posteriores. Esta segmentación permitirá a Arca Continental realizar predicciones informadas y direccionar estrategias de marketing más eficaces hacia clientes potenciales, optimizando recursos y aumentando la tasa de adopción de sus lanzamientos.

Antecedentes

La inteligencia artificial ha transformado diversos sectores, ofreciendo herramientas para optimizar procesos, mejorar la toma de decisiones y personalizar productos y servicios. En el sector de la salud, la IA ha impulsado diagnósticos más precisos y tratamientos personalizados, aunque plantea preocupaciones éticas sobre la privacidad de los datos y el sesgo algorítmico. En finanzas, se ha mejorado la detección de fraudes y la gestión de riesgos, pero también se debate el impacto en la seguridad de los datos y la equidad en los algoritmos de crédito. En educación, la personalización de la enseñanza ha aumentado, aunque existe preocupación por el uso de datos sensibles de estudiantes. En todos los sectores, el cumplimiento de normativas como el GDPR y principios éticos como la transparencia es crucial.
En el ámbito específico de nuestro proyecto con Arca Continental, la predicción del éxito de ventas mediante herramientas de IA implica trabajar con datos demográficos y de consumo, lo cual puede beneficiar a la empresa al optimizar la oferta de productos y mejorar la satisfacción del cliente. Sin embargo, hay desafíos éticos, como garantizar la privacidad de los clientes y evitar sesgos que puedan surgir al segmentar grupos de clientes de manera injusta. A nivel normativo, es fundamental cumplir con leyes de protección de datos personales y considerar el impacto de las decisiones automatizadas en las prácticas de negocio y la percepción pública de la empresa.

Recursos

Dado que el proyecto implica un análisis de datos, se decidió desarrollar la solución en Python, ejecutándose de manera local en Visual Studio Code para cumplir con los términos del acuerdo de confi-

dencialidad (NDA) firmado. Para ello, se emplearon tres librerías clave: Pandas, para gestionar los DataFrames; Numpy, para realizar operaciones sobre arrays; y Scikit-Learn, para el preprocesamiento y otras transformaciones de datos.

Los datos utilizados provienen de tres archivos .CSV proporcionados por Arca Continental, que contienen registros de ventas de 2019 a 2022, descripciones de productos y un diccionario con características demográficas de sus clientes.

Metodología

Extracción de datos

Para la solución de la problemática nos enfocaremos en dos fuentes principales: el historial de ventas y las características de los productos. Para su análisis se convirtieron los archivos en formato CSV a un dataframe de Pandas. Como convencion se estará refiriéndo al daraset extraído de productos como "productos" y al dataset extraído de ventas como "ventas". A continuación se detalla una descripción de las caracterísitcass generales que se encontraron para los dos dataframes que vamos a utilizar:

Ventas

proporciona un registro histórico de transacciones, donde cada entrada corresponde a la venta de un producto específico a un cliente en un mes dado. Este archivo contiene los siguientes campos relevantes:

- ***CustomerId*** : identificador único del cliente
- ***material*** : código del producto vendido
- ***calmonth*** : fecha en la que se vendió el producto en formato yyyymm y *unibox* : cantidad de unidades vendidas en formato decimal

Productos proporciona información detallada sobre las características de cada producto disponible para la venta. Cada entrada corresponde a un producto específico y contiene los siguientes campos relevantes:

- ***Material:*** Código único que identifica el producto.
- ***Materialdesc***: Descripción detallada del producto, que incluye marca, sabor, tamaño y presentación.
- ***ProductosPorEmpaque***: Cantidad de unidades contenidas en el empaque del producto).
- ***BrandPresRet***: Combinación de la marca y el tipo de presentación (retornable o no retornable) del producto.
- ***ProdKey***: Llave que agrupa los productos en categorías amplias.
- ***Brand***: Marca del producto, que identifica el fabricante o marca comercial.
- ***Presentation***: Tipo de presentación y volumen del producto.
- ***MLSize***: Tamaño en mililitros del producto.
- ***Returnability***: Indicador de retornabilidad del envase.
- ***Pack***: Presentación general del tamaño del producto.
- ***Size***: Tamaño del producto, simplificado en distintas formas de presentación.
- ***Flavor***: Sabor del producto.
- ***Container***: Tipo de envase del producto.
- ***Ncb***: Indicador de negocio para productos no carbonatados (1) o carbonatados (0).
- ***ProductType***: Tipo de producto.
- ***ProductCategory***: Categoría general del producto.
- ***SegAg***: Segmento agregado del producto, que lo clasifica en categorías de mercado amplias.
- ***SegDet***: Segmento detallado del producto, que permite subcategorizar dentro de SegAg.

- ***GlobalCategory***: Clasificación global del producto, indicando la categoría general a la que pertenece.
- ***GlobalSubcategory***: Subcategoría global del producto, una clasificación más específica dentro de GlobalCategory.
- ***BrandGrouper***: Agrupador de marca que clasifica productos de marcas similares en una categoría.
- ***GlobalFlavor***: Sabor global del producto, que permite estandarizar la clasificación de sabores.

La cantidad de columnas y de filas para los datasets es: ventas tiene 2,347,110 filas y 4 columnas, mientras que productos tiene 793 filas y 22 columnas. La columna material sirve como llave foránea para conectar con ventas, por lo que se comprobó que para cada fila que hay en productos el valor de material es único.

En ventas hay 3 variables de tipo entero, dos de las cuales fungen como llave compuesta siendo CustomerId y material sus respectivas claves. la variable restante es de tipo decimal.

En productos hay 4 columnas de tipo entero y 18 columnas de tipo object, de antemano sabemos que las que son de tipo object pueden ser convertidas a cadenas de texto, proceso que se hará

Limpieza de datos

0.0.1 Manejo de datos nulos

Transformacion de datos

Dichas variables sobrevivientes a la limpieza fueron aquellas que cubren desde el número de contenedores por empaque, el volumen de los productos, si es retornable, el tamaño como "familia" o "individual", el sabor del producto, el material del contenedor, el tipo de producto, si el producto es un líquido o comida, la marca a la que pertenece y el grupo de dicha marca.

Dada la naturaleza categórica de todas las features que definen un producto, las variables con 2 categorías se transformaron en variables binarias manualmente, mientras que a las variables multiclase se les aplicó la función One-Hot Encoding de Scikit Learn, la cual convierte variables categóricas en formato numérico, creando una nueva columna para cada posible categoría y asignando un valor binario (0 o 1) para indicar si una instancia pertenece o no a esa categoría. Para finalizar con la limpieza de los productos se estandarizaron los nombres de todas las columnas y se eliminó la columna de identificadores de los productos, creando así un espacio vectorial en que cada producto es representado por un vector de características único.

Una vez terminado de preparar los datos del archivo de productos, ahora pasamos a la preparación del archivo de ventas, donde principalmente hicimos una transformación a la variable de Date Time donde lo organizamos en formato de día, mes y año, para después hacer la separación por mes, esto para un mejor análisis, y realizamos un re nombramiento de las columnas donde obtenemos una columna que nos es de gran importancia sobre los galones vendidos.

Para elevar la significancia sobre las ventas de cada cliente por mes sobre los galones que se compraron de un producto específico, se realizó una transposición de los meses para que estos se conviertan en columnas y muestren como valor la cantidad de galones que el cliente compro de ese producto, si hubo meses donde no se compró ese producto a la columna del mes donde no se compró se deja con un valor de 0.

Creamos una nueva columna llamada "proporción" donde calculamos la proporción de los galones comprados de productos de un mismo cliente con base en los meses, esto quiere decir que tenemos como resultado una proporción de un producto que compro un cliente con base en las demás compras que hecho ese mismo cliente de otros productos. Y registramos las ventas unitarias por mes de ese producto de la misma manera que la anterior, con base en su proporción con las ventas de los demás productos.

Producto Nuevo

Para introducir un nuevo producto, se crea un vector de características igual al de la base de datos de los productos con la única condición de que sea diferente a todos los productos existentes por al menos una característica. Por ejemplo:

$$Existente = [Fanta, Naranja, 600ml, Refresco]$$

$$Nuevo = [Fanta, Limón, 600ml, Refresco]$$

Indicador de éxito

Definido un producto nuevo, se calcula su similitud coseno con todos los otros productos

Resultados

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla nec malesuada libero, et accumsan orci. Etiam tempus, eros id rutrum rhoncus, tortor nisi ultrices lacus, eu rutrum sem orci sed velit. In rhoncus, tellus eu suscipit scelerisque, urna libero pharetra ante, sit amet porta tortor lacus ut lacus. Aliquam fringilla nec nibh eleifend suscipit. Etiam vulputate elit a pellentesque accumsan. Nulla leo velit, viverra ut odio sit amet, congue vehicula enim. Sed nec volutpat ipsum, sed venenatis quam. Sed iaculis gravida finibus. Fusce laoreet convallis felis, sed fringilla tellus auctor eu.

Conclusiones

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla nec malesuada libero, et accumsan orci. Etiam tempus, eros id rutrum rhoncus, tortor nisi ultrices lacus, eu rutrum sem orci sed velit. In rhoncus, tellus eu suscipit scelerisque, urna libero pharetra ante, sit amet porta tortor lacus ut lacus. Aliquam fringilla nec nibh eleifend suscipit. Etiam vulputate elit a pellentesque accumsan. Nulla leo velit, viverra ut odio sit amet, congue vehicula enim. Sed nec volutpat ipsum, sed venenatis quam. Sed iaculis gravida finibus. Fusce laoreet convallis felis, sed fringilla tellus auctor eu.

Discusiones

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla nec malesuada libero, et accumsan orci. Etiam tempus, eros id rutrum rhoncus, tortor nisi ultrices lacus, eu rutrum sem orci sed velit. In rhoncus, tellus eu suscipit scelerisque, urna libero pharetra ante, sit amet porta tortor lacus ut lacus. Aliquam fringilla nec nibh eleifend suscipit. Etiam vulputate elit a pellentesque accumsan. Nulla leo velit, viverra ut odio sit amet, congue vehicula enim. Sed nec volutpat ipsum, sed venenatis quam. Sed iaculis gravida finibus. Fusce laoreet convallis felis, sed fringilla tellus auctor eu.

References

[1] P. Virtanen et al., "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261-272, 2020.

https://www.arcacontal.com/