Tecnológico de Monterrey

CAMPUS MONTERREY

INTELIGENCIA ARTIFICIAL AVANZADA PARA LA CIENCIA DE DATOS II

TC3007C

RETO - ARCA CONTINENTAL

MOMENTO DE RETROALIMENTACIÓN: METODOLOGÍA

Prof. Edgar González Fernández

Daniela Jiménez Téllez - A01654798
Lautaro Gabriel Coteja - A01571214
Andrés Villareal González - A00833915
Héctor Hibran Tapia Fernández - A01661114

I. Introducción

Arca Continental es la segunda embotelladora más grande de América Latina, y una de las más relevantes a nivel global, con una gran trayectoria en la producción y distribución de bebidas de The Coca-Cola Company, así como de botanas saladas en países como México, Ecuador y Estados Unidos. Con marcas como Bokados, Inalecsa y Wise, la compañía ofrece una variedad en su oferta, y tiene un posicionamiento sólido en distintas áreas de consumo.

En este contexto, contar con una comprensión de cómo se comportan los clientes en sus compras es de suma importancia para la empresa, especialmente cuando se trata del lanzamiento de nuevos productos. La forma en la que responde el mercado a estos lanzamientos puede variar dependiendo de varios factores, como lo son el perfil de clientes, y sus hábitos de compra y venta en torno a otros productos similares. Identificar patrones que hay entre los clientes que tienden a comprar o ignorar productos nuevos es un aspecto importante para crear estrategias comerciales más efectivas, las cuales permitan minimizar el riesgo de fracaso y optimizar la inversión en estos productos para así obtener mejores resultados.

El propósito de este proyecto es abordar los puntos anteriormente mencionados a través del análisis de una base de datos, proporcionada por Arca Continental, que contiene información relevante sobre las características de los clientes, su historial de compras, y su interacción con diferentes productos. Al explorar estos datos, se busca crear un perfil de cada consumidor, con la intención de crear y probar diferentes modelos de Machine Learning para predecir el éxito que tendrían futuros productos de lanzamiento con cada cliente, basado en comportamientos previos. Habiendo hecho esto, se pretende brindar a Arca Continental una lista de los clientes que probablemente comprarían estos productos, la cual no solo aporta valor en términos de conocer a sus clientes, sino que también ofrece una ventaja competitiva en el mercado.

II. Objetivos

A continuación se presentan los objetivos que guiarán el análisis para obtener resultados prácticos y precisos que apoyen la toma de decisiones en el lanzamiento de productos nuevos:

1. Desarrollar modelos predictivos empleando técnicas de Machine Learning e Inteligencia Artificial para identificar clientes con alta probabilidad de compra de productos de lanzamiento.
2. Identificar los segmentos de clientes más dirigidos a comprar productos nuevos, considerando factores como historial de compra y venta, y preferencias de consumo.
3. Optimizar los perfiles de cliente para aumentar las oportunidades de venta de los productos de lanzamiento.
4. Evaluar el impacto de factores demográficos y de infraestructura en el rendimiento de los productos lanzados, identificando patrones que contribuyan a su éxito o fracaso en el mercado.

III. Antecedentes y proyectos relacionados

En el proyecto para Arca Continental, el análisis de datos comenzó con la entrega de tres conjuntos de datos directamente proporcionados por la empresa. Estos datasets contenían información clave sobre productos, características de las tiendas y registros de ventas. Esta base de datos facilitó un punto de partida robusto para realizar el análisis y la evaluación de los productos en su red de tiendas.

El primer paso fue integrar estos tres conjuntos de datos en un único dataset mediante técnicas de unión, lo cual permitió visualizar la información de una manera más ordenada, habilitando un análisis más completo de las relaciones entre productos, características de las tiendas y patrones de compra de los clientes.

Además, se añadieron columnas nuevas con información relevante para la evaluación de algunas características de las tiendas. Uno de los elementos clave en este análisis fue la identificación de los productos de lanzamiento y la evaluación de su éxito en cada tienda. Para cada producto de lanzamiento, se definió un criterio de éxito basado en la frecuencia y continuidad de las compras por parte de los clientes durante los meses posteriores al lanzamiento. De esta forma, aquellos productos que fueron adquiridos de manera sostenida se clasificaron como exitosos, mientras que aquellos sin dicha continuidad no cumplieron con el criterio de éxito.

En proyectos anteriores, se han abordado problemas similares con tecnologías avanzadas de análisis de datos y modelos de machine learning, lo cual demuestra la efectividad de estas herramientas para mejorar la toma de decisiones. La aplicación de técnicas de inteligencia artificial y el análisis predictivo permiten no solo analizar grandes

volúmenes de datos, sino también identificar patrones y segmentaciones que resultan útiles para mejorar la colocación de productos y la predicción de su éxito. Estos avances tecnológicos se han traducido en beneficios estratégicos para empresas al proporcionarles una comprensión más precisa de sus clientes y optimizar su oferta de productos.

IV. Herramientas y recursos a usar

Para abordar el problema de manera efectiva, se emplearon las siguientes herramientas y recursos:

Python
Lenguaje base para el desarrollo del modelo y análisis de datos, elegido por su flexibilidad y amplia gama de bibliotecas orientadas a la ciencia de datos.

Pandas / Numpy
Herramientas esenciales para la manipulación y transformación de datos. Pandas permite trabajar con estructuras de datos complejas como DataFrames, facilitando el filtrado, agrupación, y agregación de datos, mientras que numpy ofrece operaciones eficientes sobre matrices numéricas, optimizando el cálculo a gran escala.

Matplotlib / Seaborn
Librerías de visualización que facilitan la creación de gráficos descriptivos y análisis exploratorio de datos, crucial para identificar patrones y tendencias preliminares en los datos.

Scikit-Learn
Biblioteca robusta para el modelado predictivo. Incluye una variedad de algoritmos de machine learning, preprocesamiento, y herramientas para evaluar el rendimiento del modelo.

TensorFlow
Framework avanzado para construir y entrenar redes neuronales. Su capacidad para manejar datos complejos y optimizar el entrenamiento de modelos profundos lo hace ideal para abordar problemas predictivos de alta dimensionalidad.

Google Colab / Jupyter Notebooks
Entornos interactivos en la nube que permiten la ejecución y documentación de código en tiempo real. Facilitan la colaboración entre los miembros del equipo, ya que soportan la integración con bibliotecas externas y comparten resultados fácilmente.

Estas herramientas se seleccionaron por su capacidad de manejar grandes volúmenes de datos, proporcionar análisis detallados y construir modelos predictivos robustos, alineándose con los objetivos de optimización de lanzamientos de productos.

V. Metodología

Para el análisis de éxito de productos en la red de tiendas de Arca Continental, se implementó un modelo de aprendizaje profundo en Python utilizando librerías como numpy, pandas, y tensorflow, siguiendo los siguientes pasos:

1. Carga e Integración de datos

Se utilizaron dos datasets (df_entrenamiento y df_prueba). Estos dataframes se obtuvieron a partir de los productos exitosos y fueron divididos de forma que en el dataframe de entrenamiento se encontraban todos los datos de 2020 hasta junio del 2022, y en el dataframe de prueba se encontraban todos los datos de julio del 2022 en adelante. Ambos dataframes se combinaron en un solo dataframe (df_combinado) para tener una vista completa de los datos.

2. Preprocesamiento de Datos

- **Definición de Variables Independientes y Dependientes:** La variable objetivo fue definida como "successful", que indica si un producto fue exitoso. El conjunto de caracteristicas fue creado eliminando columnas irrelevantes como "CustomerID", "Material" y "Producto_Por_Empaque".

- **Codificación de Variables Categóricas:** Para trabajar con modelos de machine learning se utilizaron variables dummy para las columnas categóricas ("sub_canal_comercial", "ProductType" y "categoria_instalaciones"), aplicando codificación One-Hot Encoding.

3. **Estandarización de Datos**

Se empleó StandardScaler para escalar las características de x, lo cual mejora la estabilidad y eficiencia del entrenamiento de redes neuronales al normalizar los datos en un rango estándar.

4. **Construcción del Modelo de Red Neuronal**

Se diseñó un modelo secuencial en TensorFlow con las siguientes capas:

- **Capa de Entrada:** Una capa densa con 64 unidades y función de activación relu, seguida de BatchNormalization y Dropout (0.5) para regularización y prevención de sobreajuste.

- **Capas Ocultas:** Dos capas ocultas adicionales con 32 y 16 unidades respectivamente, ambas utilizando BatchNormalization y Dropout (0.5) para mantener la estabilidad y mejorar la generalización del modelo.

- **Capa de Salida:** Una única unidad con activación sigmoid, adecuada para la clasificación binaria de los productos en "exitoso" o "no exitoso"

5. **Compilación y Entrenamiento del Modelo**

El modelo fue compilado usando la función de pérdida binary_crossentropy y el optimizador adam, evaluando la métrica de precisión. El entrenamiento se realizó con 50 épocas y un tamaño de lote de 32, utilizando el conjunto de validación para monitorear el desempeño de cada época.

6. **Visualización de Resultados**

La evolución de la precisión y la pérdida durante el entrenamiento y la validación se graficó usando matplotlib. Esto permitió observar el comportamiento del modelo y detectar posibles problemas de sobreajuste.

7. **Evaluación del Modelo**

- **Predicciones:** Se generaron predicciones en el conjunto de prueba, ajustando el umbral de decisión a 0.5 para clasificar los productos como exitosos o no exitosos.

- **Métricas de Rendimiento:** Se calcularon la precisión en el conjunto de prueba y el reporte de clasificación como accuracy_score, classification_report, y confusion_matrix de sklearn, proporcionando un análisis detallado del desempeño del modelo.

8. Predicciones en Nuevos Datos Filtrados

El modelo permite realizar predicciones de éxito para un tipo específico de producto (ProductType). Tras seleccionar el tipo de producto, se procesan y escalan los datos correspondientes, generando una tabla con el CustomerId y la probabilidad de éxito para cada cliente. Esta funcionalidad permite ordenar y priorizar a los clientes con mayor probabilidad de adoptar un nuevo producto.

VI. Resultados
VII. Conclusiones y trabajo a futuro
VIII. Referencias