Benemérita Universidad
Autónoma de Puebla

Facultad de Ciencias de la Computación

Reporte Fase 1: Descubrimiento del proyecto

Ximena Axel Martínez Pelayo
21 de octubre de 2024

Introducción a la ciencia de datos

M.C. Jaime A. Romero Sierra

Objetivo del proyecto

Analizar un conjunto de datos sobre transacciones de ventas para optimizar estrategias comerciales, mejorar la rentabilidad y proporcionar información clave para la toma de decisiones basadas en datos. Identificar patrones de comportamiento de compra, evaluar la rentabilidad de productos y categorías y detectar tendencias de ventas permite diseñar estrategias más efectivas y personalizadas.

Segmentar a los clientes de acuerdo a su perfil demográfico y geográfico facilita la creación de campañas de marketing dirigidas y el enfoque en mercados con mayor potencial de crecimiento. Además de que identificar periodos de alta y baja demanda a través del análisis temporal contribuye a una planificación de inventarios más precisa y estrategias promocionales fundamentadas.

Incorporar modelos predictivos basados en los datos históricos permite anticipar tendencias futuras de ventas y optimizar procesos de planificación, garantizando una mejora continua en la gestión comercial. Este enfoque asegura un crecimiento sostenible y una mayor competitividad en un entorno dinámico

Descripción del proyecto

El análisis de transacciones de ventas es esencial para comprender el desempeño de un negocio, ya que proporciona una visión integral de factores que afectan la rentabilidad y el comportamiento del consumidor. La variedad de datos ofrece la oportunidad de realizar análisis a múltiples niveles, incluyendo segmentación de clientes, evaluación de desempeño por producto o categoría y análisis geográfico y temporal.

Trabajar con este tipo de datos requiere el uso de técnicas avanzadas de análisis para abordar tanto datos numéricos como categóricos, utilizando herramientas de visualización que faciliten la interpretación de resultados. Este proyecto busca descubrir patrones de comportamiento de compra, identificar tendencia de ventas y evaluar la rentabilidad de los productos para informar decisiones estratégicas.

El análisis temporal permite rastrear tendencias por año y mes, identificar ciclos o periodos de alta y baja demanda y con base en ello planificar acciones comerciales que respondan a estas variaciones. Además, el análisis demográfico y geográfico ayuda a entender cómo las preferencias de compra varían entre diferentes segmentos y mercados, lo que permite ajustar estrategias de marketing e identificar áreas con potencial de crecimiento.

Finalmente calcular márgenes de ganancia y comparar costos y precios en diferentes categorías asegura una optimización de recursos y una priorización de productos más rentable. Este enfoque integral busca no solo mejorar el desempeño comercial, sino también fomentar una dinámica de toma de decisiones basadas en datos, impulsando el crecimiento del negocio.

El proyecto se centra en el análisis de un conjunto de datos detallado sobre transacciones de ventas, con el propósito de extraer información clave que permita optimizar las estrategias comerciales y mejorar el desempeño general del negocio. El conjunto de datos incluye información relevante sobre las características demográficas de los clientes, como edad y género, además de detalles geográficos, temporales y financieros relacionados con las ventas. Este análisis es crucial para identificar patrones en el comportamiento de compra de los clientes, comprender las tendencias de ventas a lo largo del tiempo y evaluar la rentabilidad de los productos y categorías ofrecidos.

Recursos Disponibles

Tecnología y herramientas

Para llevar a cabo el análisis del proyecto, se utilizará Google Colaboratory (Google Colab), una herramienta basada en la nube que permite la escritura y ejecución de código en notebooks de Jupyter sin necesidad de instalación local. Google Colab facilita la colaboración en tiempo real, la integración con Google Drive y el uso de otros recursos que facilita manejar grandes volúmenes de datos.

El lenguaje de programación principal que se empleará es Python, popularmente utilizado en proyectos de análisis de datos y aprendizaje automático por su facilidad de uso y gran repositorio de bibliotecas. Entre las bibliotecas utilizadas en el análisis estarán:

- NumPy: Para cálculos numéricos y manipulación de arrays.
- Pandas: Para la manipulación y análisis de datos estructurados, especialmente útil para conjuntos de datos tabulares.
- Matplotlib & Seaborn: Para la creación de visualizaciones que permitan interpretar patrones y tendencias.
- TensorFlow: Para la implementación de modelos avanzados de aprendizaje automático, si es requerido para el análisis.
- Otras bibliotecas como Scikit-learn o Plotly, dependiendo de las necesidades específicas del proyecto.

Hay que destacar que los datos utilizados en este análisis provienen del conjunto de datos titulado "Sales Data for Economic Data Analysis", disponible en la plataforma de colaboración en línea Kaggle. Kaggle es un recurso ampliamente reconocido en la comunidad de ciencia de datos, ya que facilita el acceso a conjuntos de datos variados, la publicación de proyectos y la colaboración entre usuarios con intereses y objetivos similares.

El conjunto de datos proporciona información estructurada y de calidad que permite explorar diversas métricas relacionadas con las transacciones de ventas, como características demográficas, detalles geográficos, entre otras. Kaggle, además de ser una fuente confiable de datos, fomenta el aprendizaje automático y el análisis de datos.

La combinación de herramientas tecnológicas avanzadas como Google Colab y bibliotecas especializadas de Python, junto con un conjunto de datos estructurado y accesible desde Kaggle, asegura una base sólida para la ejecución exitosa del proyecto. Lo anterior, garantiza un entorno de análisis eficiente y adaptable a los requerimientos del proyecto.

Datos

El análisis de la base datos depende de variables clave que permiten explorar tendencias, patrones de comportamiento del consumidor y métricas del rendimiento financiero. A continuación, se describen las columnas que representan características específicas en las transacciones y se explica su utilidad en el análisis.

- Year (Año): Indica el año en que ocurrió la transacción. Esta columna es crucial para identificar tendencias a largo plazo y analizar variaciones en las ventas por periodos anuales.
- Month (Mes): Representa el mes de la transacción, lo que permite un análisis más detallado de la estacionalidad y de los picos o caídas de la demanda.
- Customer Age (Edad del Cliente): Define la edad del cliente, proporcionando información valiosa para segmentar el mercado por grupos etarios y analizar sus comportamientos de compra.
- Customer Gender (Género del Cliente): señala el género del cliente, permitiendo explorar diferencias en las preferencias de compra según el género.
- Country (País): Indica el país donde se realizó la transacción, ofreciendo una perspectiva geográfica de las ventas y la oportunidad de comparar mercados internacionales.
- State (Estado): Especifica el estado dentro del país, permitiendo un análisis más detallado a nivel regional.
- Product Category (Categoría del Producto): Clasifica los productos en categorías generales, facilitando la comparación del rendimiento entre diferentes líneas de productos.
- Sub Category (Subcategoría): Detalla la categoría específica del producto vendido, permitiendo un análisis granular de las ventas.
- Quantity (Cantidad): Define el volumen de unidades vendidas, proporcionando datos esenciales para medir la demanda y calcular ingresos.
- Unit Cost (Costo Unitario): Indica el precio al que se vendió una unidad del producto, ayudando a analizar estrategias de precios.
- Cost (Costo Total): Calculado como el producto de la cantidad vendida y el costo unitario, refleja los gastos asociados a las ventas.
- Revenue (Ingresos): Calculando como el producto de la cantidad vendida y el precio unitario, indica los ingresos totales generados por las ventas.

Esta organizada estructura de datos proporciona la base necesaria para realizar un análisis profundo que apoye la toma de decisiones informadas y la implementación de estrategias comerciales efectivas.

Hipótesis Iniciales

1) **Relación entre la categoría del producto, el ingreso generado y el costo unitario**

   Hipótesis: Los productos de la categoría “Bikes” generan mayores ingresos promedio debido a su alto costo unitario, en comparación con las categorías “Clothing” y “Accessories”

2) **Impacto de la edad del cliente en las compras considerando subcategorías y costo unitario**

   Hipótesis: Los clientes más jóvenes (menores de 35 años) prefieren productos de subcategorías con menor costo unitario, como “Socks” y “Caps” en la categoría de accesorios, mientras que los clientes mayores al rango de edad establecido optan por subcategorías con mayor costo unitario, como “Mountain Bikes” y “Road Bikes” en la categoría de bicicletas

3) **Diferencias geográficas en la preferencia de productos**

   Hipótesis: Los clientes en estados europeos tienden a una mayor preferencia por productos de ropa y accesorios en comparación con clientes en Estados Unidos, que prefieren bicicletas.

Stakeholders clave

Los stakeholders clave son factores con una relación directa con el proyecto, cuya participación es esencial para la toma de decisiones estratégicas, la interpretación de resultados y la implementación de mejoras basadas en el análisis de datos. A continuación, se describen los principales para este proyecto:

- Proveedores o Socios Comerciales: Se refiere a las empresas que producen o distribuyen los productos analizados. Su rol es ajustar su oferta y logística con base a las tendencias descubiertas en los datos. Además, colaboran estrechamente con los equipos de ventas y marketing para alinear estrategias comerciales, maximizando el impacto en el mercado y mejorando la experiencia del cliente final.
- Directivos o Inversionistas: Son tomadores de decisiones de alto nivel y gran importancia estratégica, interesados en los resultados financieros, tomando decisiones sobre asignación de recursos hacia áreas con mayor potencial de crecimiento o mejora. Su rol es crucial para garantizar que los insights se traduzcan en acciones concretas y alineadas con los objetivos organizacionales.
- Clientes finales: Involucra a todas las personas que adquieren los productos analizados. Representan la base del análisis, ya que sus patrones de comportamiento, preferencias y tendencias de compra determinan las decisiones estratégicas.

Preguntas clave

¿Qué ajustes en la oferta de productos pueden realizarse para maximizar los ingresos por categoría?

¿Cómo afectan las decisiones de precios a las ventas generales por categoría?

¿Qué proporción de los ingresos proviene de cada categoría y cómo esto impacta la rentabilidad general?

¿Prefieren los clientes productos de menor costo unitario, pero con mayor frecuencia de compra?

¿Qué subcategorías muestran mayor crecimiento en ventas entre distintos segmentos de edad?

¿Qué grupos de edad representan la mayor proporción de ingresos para cada subcategoría?

¿Cómo puede influir el comportamiento por edades en las proyecciones de crecimiento?

¿Qué características buscan los clientes jóvenes en contraste con los clientes mayores?

¿Cómo varía la demanda de ropa y accesorios entre estados europeos y Estados Unidos?

¿Qué regiones geográficas generan mayores ingresos para cada categoría de producto?

¿Cómo influyen los costos de distribución entre regiones en la rentabilidad?

¿Sería importante considerar factores culturales en las preferencias regionales?

¿Cuáles son los productos más rentables para cada segmento de clientes?

Fuentes de datos

La base de datos probablemente fue construida a partir de una combinación de fuentes comunes utilizadas en análisis comerciales y de mercado. Es esencial que los datos hayan sido ajustados y procesados para garantizar su relevancia, precisión y aplicabilidad en el contexto del proyecto. A continuación, se describen posibles fuentes que pudieron contribuir a su elaboración y seguir asegurando su confiabilidad:

- Registros de Ventas: Sistemas de Punto de Venta o plataformas de comercio electrónico donde pudieron considerar datos como:
    - Productos vendidos, categoría, subcategoría y costo unitario
    - Información geográfica (ubicación del cliente o tienda)
    - Fecha y hora de compra
    - Total de ingresos generados por cada venta

- Perfiles de clientes: Base de datos de clientes o programas de fidelidad, donde pudieron considerar datos como:
    - Edad, género y ubicación
    - Historial de compras
    - Preferencias de productos y patrones de compra
    - Segmentación por demografía o comportamiento

- Datos de Mercado y Competencia: Informes de investigación de mercado, estudios sectoriales o benchmarking, donde pudieron considerar datos como:
    - Tendencias generales de consumo en categorías como bicicletas, ropa y accesorios.
    - Preferencias geográficas o culturales en el consumo
    - Comparaciones entre marcas competidoras

- Inventarios y Producción: Sistemas de gestión de inventarios o Enterprise Resource Planning (ERP), donde pudieron considerar datos como:
    - Costos unitarios asociados a la fabricación o distribución
    - Disponibilidad de productos por región o tienda

  - Rotación de inventarios

- Encuestas y opiniones de clientes: Encuestas directas, opiniones en línea o redes sociales, donde pudieron considerar datos como:
  - Preferencias y comportamientos según categorías de producto
  - Satisfacción con las subcategorías ofrecidas
  - Factores decisivos de compra, como precio o diseño

- Datos Geográficos y Demográficos: Fuentes externas como censos, datos gubernamentales o bases de datos comerciales, donde pudieron considerar datos como:
  - Distribución poblacional y segmentos por región
  - Niveles de ingreso y gasto promedio por área geográfica
  - Cultura que influya en las preferencias de productos

- Análisis de Marketing Digital: Herramientas de análisis web (Google Analytics, redes sociales, entre otras.), donde pudieron considerar datos como:
  - Comportamiento del usuario en sitios web o aplicaciones móviles
  - Popularidad de productos específicos según campañas publicitarias
  - Localización de los clientes que interactúan con las plataformas digitales

Justificación

En busca de una respuesta o alternativa estratégica a la creciente necesidad de comprender y anticipar los patrones de compra de los clientes en un mercado global competitivo y dinámico, en un entorno donde la variedad de productos, el acceso a información y las preferencias de los consumidores cambian rápidamente, las empresas deben apoyarse en análisis profundos y fundamentados para tomar decisiones informadas y adaptativas.

El análisis de la base de datos tomando como base las hipótesis propuestas, permiten:

1) Maximizar la rentabilidad: Al identificar categorías y subcategorías de productos que generan mayores ingresos y analizar su relación con otras variables se facilita la toma de decisiones estratégicas que optimizan la inversión y priorizan las áreas con mayor retorno potencial.
2) Segmentación de mercado avanzada: Comprender las preferencias de los clientes en función de factores clave, permite desarrollar estrategias específicas para cada segmento de variabilidad por cliente, asegurando una experiencia de compra más personalizada.
3) Eficiencia en la gestión de recursos: Al alinear la oferta de productos y las estrategias logísticas con las tendencias detectadas, los socios comerciales y proveedores pueden reducir costos operativos, mejorar la distribución y asegurar que los productos más demandados estén disponibles en los mercados adecuados.
4) Aseguramiento de la toma de decisiones estratégicas: Los directivos e inversionistas necesitan información clara, confiable y accionable para orientar el crecimiento del negocio. Resultarán insights valiosos que impactan la planeación a corto plazo y también establecen una visión estratégica a largo plazo basada en datos reales.
5) Fortalecimiento de relaciones comerciales y competitividad: Ajustar la oferta a las demandas específicas del mercado, fomentando relaciones sólidas con clientes y socios comerciales. Asimismo, asegurar que la organización pueda responder con agilidad a las tendencias emergentes, garantizando su sostenibilidad y diferenciación en el mercado.
6) Adaptabilidad a diferencias geográficas: Las conclusiones sobre las preferencias según la ubicación ayudan a atender mercados específicos con mayor precisión, permitiendo estrategias diferenciadas en regiones clave como Estados Unidos y Europa.

Al fomentar una cultura empresarial orientada al análisis y la mejora continua, asegurando que la organización no solo se adapte a las necesidades actuales del mercado, sino que también lidere su evolución. Se garantiza que la empresa está preparada para enfrentar los desafíos del presente y del futuro, posicionándose como un referente de innovación, eficiencia y orientación al cliente en su industria.

Tipo de datos

Con un total de 44,611 registros y 16 columnas, la información detalle por columna es la siguiente:

Columna	Tipo de Dato	Notas de análisis
index	float64	Este es un tipo de dato numérico.
Date	object	Es una columna de tipo texto. Probablemente debiera convertida a tipo fecha.
Year	float64	Este es un tipo de dato numérico.
Month	float64	Tiene solo valores nulos. Debería ser eliminada o revisada.
Customer Age	float64	Este es un tipo numérico, pero tiene valores nulos que deberían ser tratados.
Customer Gender	object	Tipo de dato texto (categoría). Tiene valores nulos.
Country	object	Tipo de dato texto (categoría). Sin valores nulos.
State	object	Tipo de dato texto (categoría). Sin valores nulos.
Product Category	object	Tipo de dato texto (categoría). Sin valores nulos.
Sub Category	object	Tipo de dato texto (categoría). Sin valores nulos.
Quantity	float64	Tipo numérico. Sin valores nulos.
Unit Cost	object	Aunque es texto, debería ser numérico. Necesita conversión.
Unit Price	object	Aunque es texto, debería ser numérico. Necesita conversión.
Cost	float64	Tipo numérico. Sin valores nulos.
Revenue	float64	Tipo numérico. Sin valores nulos.
Column1	float64	La mayoría de los valores son nulos. Debería ser revisada o eliminada.