UNIVERSIDAD PRIVADA FRANZ TAMAYO

FACULTAD DE INGENIERIA DE SISTEMAS

CARRERA DE INGENIERIA DE SISTEMAS

OBJETIVO DEL ANÁLISIS Y EXTRACCIÓN DE DATOS

Estudiante:

Cesar Leonel Aliaga Bacarreza

Materia:

BIG DATA

Docente:

Ing. Enrique Laurel Cossio

La Paz – Bolivia

Objetivo del análisis

El análisis se orienta a explicar y optimizar el rendimiento de un e-commerce de pastelería que integra herramientas de CRM y marketing digital para impulsar la visibilidad y las ventas de pastelerías emergentes en la ciudad de La Paz. Se busca descubrir patrones de navegación, adición al carrito, abandono y compra a lo largo del embudo digital, así como caracterizar segmentos de clientes de alto valor mediante enfoques como RFM y cohortes. En paralelo, se pretende predecir la probabilidad de conversión por canal y campaña, el ticket de compra, la recurrencia y el Customer Lifetime Value (CLV), además de pronosticar la demanda por categoría y SKU para apoyar la planificación de producción y abastecimiento. El análisis explica también el efecto de las acciones de CRM correos, WhatsApp y cupones y de la pauta pagada Meta y Google Ads sobre ingresos, margen y retorno de inversión publicitaria.

La relación con los objetivos específicos planteados previamente es directa: el incremento de la conversión digital se aborda con métricas de embudo y modelos de propensión; el aumento de recurrencia y ticket promedio se trata mediante segmentación RFM/CLV y pruebas controladas en CRM; la optimización del gasto publicitario se fundamenta en métricas de atribución y evaluación de ROI a nivel de campaña; y la planificación de producción se sustenta en pronósticos de demanda por semana y producto.

Metodológicamente, el trabajo combina análisis exploratorio para comprender distribuciones, estacionalidades y cohortes; análisis descriptivo para cuantificar indicadores clave como CR, AOV, CAC y churn; y análisis predictivo para conversión, recurrencia, CLV y demanda. Se anticipa la construcción de tableros de funnel y cohortes, mapas de calor por día y hora, curvas ROC/PR para modelos de clasificación, gráficas de importancia de variables y series de tiempo por producto. En cuanto a modelado, se prevén segmentaciones no supervisadas (RFM/K-Means), modelos supervisados para propensión y CLV (regresión logística, random forest o XGBoost) y enfoques de series temporales (ARIMA/Prophet o LSTM, según la granularidad de datos disponible).

Descripción del dataset

El ecosistema de datos se compone de varias fuentes consolidadas en un data lake. La base transaccional del e-commerce (Laravel/MySQL) aporta órdenes, ítems, clientes y cupones; el CRM (p. ej., Mailchimp, Brevo o solución propia) aporta campañas y eventos de engagement; la analítica web (GA4 o Matomo) entrega sesiones y eventos de navegación con sus UTM; las plataformas publicitarias (Meta Ads y Google Ads) proveen gasto, impresiones y clics por campaña, conjunto y anuncio; el procesador de pagos (Stripe o Mercado Pago) entrega estados de pago y comisiones; y el módulo de catálogo/inventario agrega productos, categorías, costos y stock.

A modo de referencia, los conjuntos de datos pueden denominarse ecom_orders, ecom_order_items, ecom_customers, ecom_products, crm_campaigns, crm_events, web_analytics_sessions, web_analytics_events, ads_meta, ads_google y payments_transactions. El tamaño exacto dependerá del histórico disponible; como guía, es habitual superar holgadamente el umbral de 30 000 registros al integrar navegación, campañas y transacciones. Los tipos de datos abarcan variables numéricas (precio, cantidad, costo, margen, ROAS, CTR, CPC), categóricas (canal, campaña, categoría de producto, método de pago, estado del pedido), texto (nombres de productos, términos de búsqueda, contenido UTM) y marcas de tiempo (sesión, pedido, envío de campaña, aprobación de pago). Las variables objetivo varían según el caso de uso: conversión binaria, ticket promedio, recompra, CLV continuo y demanda por SKU/semana.

Como guía para la documentación, puedes incluir un diccionario por fuente con campos, descripción y tipo. Un extracto de ejemplo:

Fuente	Nombre variable	Descripción	Tipo de dato
ecom_orders	order_id	Identificador de la orden	Entero
ecom_orders	order_datetime	Fecha y hora de compra	Fecha/Hora
ecom_orders	order_total	Monto total de la orden	Numérico
ecom_order_items	sku	Código del producto	Texto
ecom_order_items	qty	Cantidad del ítem	Entero
ecom_products	category	Categoría (torta, cupcake, etc.)	Categórico
crm_events	event_type	send, open, click, unsubscribe	Categórico
web_analytics_events	medium	Medio (organic, cpc, email)	Categórico
ads_meta	spend	Gasto publicitario	Numérico
payments_transactions	fee	Comisión del procesador	Numérico

Proceso de extracción de datos

Debido a que el sistema de e-commerce se encuentra en etapa de desarrollo, los datos utilizados en el análisis serán generados mediante simulación controlada, replicando las condiciones reales de operación de una plataforma de venta de productos de pastelería. El objetivo es construir un conjunto de datos representativo de la futura interacción entre usuarios, campañas de marketing digital y gestión de pedidos, para permitir el diseño, prueba y validación de los modelos analíticos propuestos.

La simulación contempla la creación de registros que imitan la estructura de bases de datos reales: órdenes, productos, clientes, eventos de CRM, tráfico web, campañas publicitarias y transacciones de pago. Cada conjunto se genera con volúmenes y distribuciones plausibles basadas en estadísticas del sector y promedios observados en e-commerces similares. Por ejemplo, se generarán datos históricos de pedidos con variación estacional en fechas clave (Día de la Madre, San Juan, Navidad) y con diferencias de comportamiento entre clientes nuevos y recurrentes. En el caso de las campañas, se simularán interacciones típicas (envíos, aperturas, clics y conversiones), mientras que en el módulo de navegación se incluirán sesiones y eventos web (page_view, add_to_cart, checkout, purchase) con su respectiva fuente o medio (orgánico, social, pago, email).

El flujo de trabajo se organizará en tres fases. En primer lugar, la generación sintética de datos se realizará con scripts en Python utilizando librerías como faker, pandas y numpy, las cuales permiten crear nombres, fechas, montos, categorías y comportamientos realistas. En segundo lugar, se almacenarán los datasets en formato CSV o Parquet dentro de un entorno estructurado en carpetas (“bronze”, “silver” y “gold”), imitando un flujo de datos empresarial. En la fase final se aplicará una limpieza y transformación para garantizar la coherencia referencial entre tablas, corregir valores faltantes, ajustar formatos de fechas y normalizar tipos de datos. Este proceso permitirá reproducir una base de datos lista para análisis exploratorio, descriptivo y predictivo.

El proceso de simulación se documentará de forma reproducible, incluyendo fragmentos de código y registros de ejecución. A modo de ejemplo:

```
from faker import Faker
import pandas as pd, numpy as np
from datetime import datetime, timedelta
fake = Faker('es_ES')


orders = []
for _ in range(5000):
    orders.append({
        "order_id": fake.uuid4(),
        "customer_id": np.random.randint(1000, 2000),
        "order_datetime": fake.date_time_between(start_date='-180d', end_date='now'),
        "order_total": round(np.random.uniform(25, 250), 2),
        "payment_status":        np.random.choice(["Aprobado", "Pendiente", "Rechazado"], p=[0.85, 0.10, 0.05])
    })
df_orders = pd.DataFrame(orders)
df_orders.to_csv("bronze/orders.csv", index=False)
```

Además, se elaborarán scripts similares para generar registros de campañas de CRM, clics publicitarios, eventos de navegación y catálogos de productos, garantizando consistencia en las llaves primarias y foráneas (por ejemplo, customer_id, product_id, campaign_id). Estos datos simulados permitirán construir

modelos analíticos bajo condiciones realistas sin comprometer información sensible ni depender de plataformas externas.

En cuanto a las consideraciones éticas, el enfoque de simulación elimina cualquier riesgo asociado al uso de información personal, al no involucrar datos reales de usuarios. No obstante, el diseño respeta los principios de privacidad, anonimización y seguridad, de modo que cuando el sistema entre en producción, las mismas prácticas puedan aplicarse con datos reales. Los procedimientos de almacenamiento, estructura de carpetas y formatos de archivo reproducen el flujo que se empleará en la etapa operativa, asegurando la escalabilidad del pipeline hacia un entorno productivo futuro.

Formatos y fuentes

Los formatos predominantes en la adquisición son CSV/TSV para extracciones rápidas y JSON para APIs, mientras que Parquet se adopta en las capas analíticas por su eficiencia columnar. El origen de los datos combina plataformas web (GA4/Matomo), publicitarias (Meta, Google Ads), de pagos (Stripe o Mercado Pago), el CRM seleccionado y la base institucional interna del e-commerce. Cuando corresponde, se transforman archivos JSON o CSV a Parquet y se unifican esquemas, se armonizan zonas horarias y se enriquecen dimensiones con taxonomías de campañas, categorías y feriados locales.

Evidencia de volumen y variedad (½ página)

La integración de navegación, campañas, CRM, transacciones y pagos asegura un volumen superior al mínimo de 30 000 registros. Como referencia ilustrativa (sustituible por datos reales): los eventos de navegación en 90 días pueden superar las 100 000 filas; los eventos de CRM (envíos, aperturas y clics) rondan decenas de miles; los datos publicitarios de varias campañas aportan decenas de miles de impresiones y clics; y el histórico de órdenes con sus ítems acumula varios miles de registros adicionales. La variedad se acredita por la diversidad de fuentes —analítica, CRM, publicidad, pagos y transacciones—, por la heterogeneidad de tipos de datos —numéricos, categóricos, texto y marcas de tiempo— y por la coexistencia de

formatos —CSV, JSON, Parquet y dumps SQL—. Como respaldo, es conveniente anexar capturas de las carpetas “bronze/silver/gold” con conteos y pesos, fragmentos de cabeceras de archivos y estadísticas descriptivas básicas (conteos por fuente, proporción de nulos y cardinalidad de variables clave).