Primera Entrega Proyecto Final

Juan David Valencia - 201728857
Juan Esteban Cuellar - 202014258

1. Definición de la problemática y entendimiento del negocio

La empresa analizada es una plataforma de **delivery de comida** que busca aumentar la cantidad de órdenes. Dentro del negocio existen dos equipos encargados de los usuarios: **Growth**, encargado de acompañar al usuario desde su primera hasta su cuarta orden, y **Engagement**, que toma el relevo una vez los usuarios alcanzan esa cuarta compra.

El análisis se enfoca en los **usuarios nuevos de Engagement**, es decir, aquellos que completaron su cuarta orden entre el 29 de marzo y el 29 de septiembre y que no hicieron parte de la segmentación inicial del año. Se consideraron únicamente los usuarios clasificados con **r_segment,** que es una clasificación proveniente de otra línea de negocio, y que históricamente muestra que estos son usuarios con mejor comportamiento y mayor potencial futuro en la vertical de comida, se separa los usuarios en 4 grupos dependiendo de qué tan valiosos son en esa vertical.

El problema principal es que el equipo de Engagement **no cuenta con un esquema claro para priorizar recursos** y definir qué usuarios recientes tienen mayor probabilidad de seguir creciendo en órdenes. Esto limita la efectividad de las estrategias de retención e incrementa el costo por adquisición.

El objetivo del proyecto es **caracterizar y segmentar** a estos nuevos usuarios para **i**dentificar perfiles de **alto potencial**, entendiendo su comportamiento en los tres meses posteriores a la cuarta orden. Los resultados permitirán orientar de forma más eficiente las campañas e incentivos.

KPIs principales:

- Delta de órdenes entre periodos (Δ órdenes)
- Tasa de actividad por recencia (≤7d, 8–14d, 15–30d, 31–90d)
- Retención posterior a la cuarta orden
- Costo por orden incremental (CPOI)

2. Ideación del producto de datos

El producto propuesto busca apoyar al equipo de Engagement en la toma de decisiones sobre **a quién dirigir incentivos y comunicaciones**, optimizando el uso del presupuesto promocional. La idea es construir una herramienta analítica que combine visualización y modelado para identificar usuarios dentro del grupo de nuevos usuarios que alcanzaron su cuarta orden.

Usuarios internos:

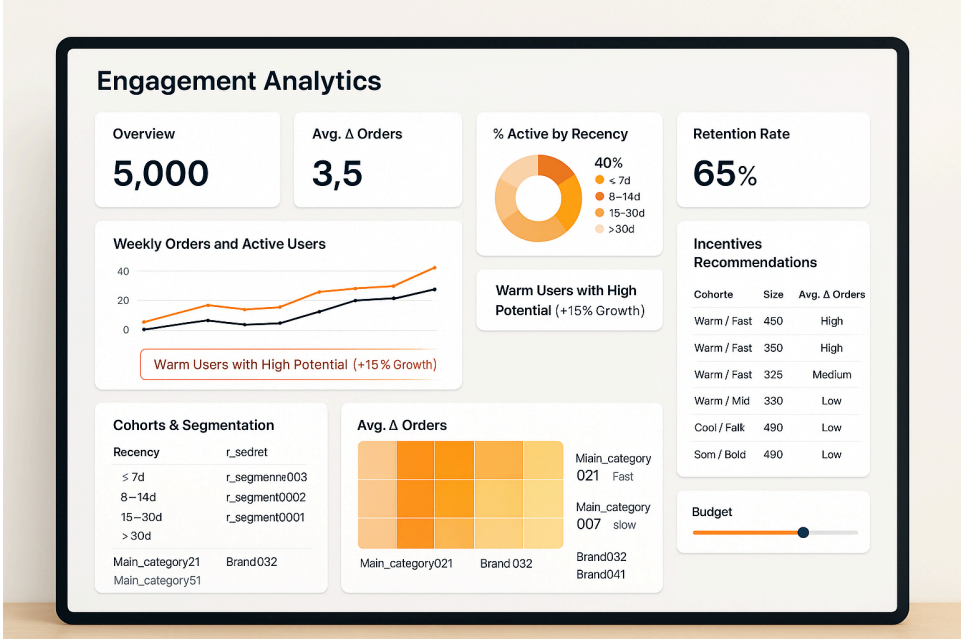
- **Equipo de Engagement:** para definir estrategias de retención y priorización de campañas.
- **Equipo de Operaciones:** para ejecutar envíos segmentados según tipo de usuario.
- **Equipo de Data:** para ajustar modelos de propensión y validar resultados.

Componentes del producto:

1. **Dashboard interactivo** que muestre métricas clave (órdenes totales, delta de órdenes, recencia, y segmentación por afinidades).
2. **Modelo analítico** que calcule la probabilidad de que un usuario vuelva a ordenar en los próximos 30–90 días.
3. **Recomendador de incentivos**, que priorice los usuarios según su potencial y afinidades (categorías, marcas o tipo de tienda).

Mockup conceptual:

- Vista principal con KPIs y evolución de cohortes.
- Segmentación dinámica por frecuencia y velocidad de adopción (EFO-to-Four).
- Panel de afinidades que muestra las categorías más frecuentes por grupo.



3. Responsible

El proyecto se basa en datos internos de usuarios transaccionales, por lo que se deben considerar aspectos de privacidad, confidencialidad y transparencia. Todos los identificadores fueron anonimizados y tokenizados, evitando la exposición de información personal o sensible.

Desde el punto de vista ético, el análisis debe garantizar que las recomendaciones no generen discriminación ni sesgos hacia grupos de usuarios específicos. También se busca mantener la transparencia en el uso de los datos, comunicando que el propósito del proyecto es mejorar la experiencia del usuario y optimizar las estrategias de retención.

En cuanto a aspectos regulatorios, el uso de la información se ajusta a la Ley 1581 de 2012 y al Decreto 1377 de 2013 sobre protección de datos personales en Colombia, así como a los principios de

tratamiento legítimo, proporcionalidad y finalidad definidos por la Superintendencia de Industria y Comercio (SIC, 2024). (Superintendencia de Industria y Comercio)

4. Enfoque analítico

El análisis busca entender **qué factores explican el crecimiento en órdenes** de los nuevos usuarios de Engagement durante los tres meses posteriores a su cuarta compra.

Se parte de tres hipótesis principales:

1. Los usuarios con **menor tiempo entre su primera y cuarta orden (EFO-to-Four)** tienden a mostrar mayor crecimiento posterior.
2. La **frecuencia de actividad** es un buen predictor de retención y volumen de órdenes.
3. Las **afinidades de consumo** (categorías, marcas y tipo de tienda) pueden orientar estrategias personalizadas de incentivo.

El dataset contiene un número elevado de variables derivadas de las **órdenes históricas**, muchas de ellas representadas como variables tipo one-hot encoder (por ejemplo, conteos de categorías, marcas y tiendas). Esto genera alta dimensionalidad, lo que puede dificultar el modelado y aumentar el riesgo de sobreajuste.

Para mitigar esto, se propone:

- **Agrupar variables** similares (por categoría o tipo de tienda) para reducir el número de columnas.
- Aplicar **técnicas de reducción de dimensionalidad** como PCA o selección de características basada en varianza o importancia del modelo.

Las métricas de evaluación serán la variación de órdenes (Δ órdenes), la tasa de reactivación, y en etapas posteriores, la precisión del modelo (AUC o F1-score).

5. Recolección de datos:

El dataset se construyó a partir de múltiples tablas internas del sistema, que almacenan información con diferentes estructuras y frecuencias de actualización. Las principales fuentes fueron tablas de órdenes, usuarios, tiendas y segmentaciones.

Uno de los mayores retos fue integrar datos de varias tablas con diferentes estilos, se identificaron tres tipos de tabla:

- **Tablas incrementales** (como dwm_finance_order_d_increment) que agregan información día a día y pueden tener millones de registros, requiriendo filtros de rango de fechas y condiciones específicas de negocio.
- **Tablas de versión diaria** (como dwm_shop_wide_d_whole o dwm_user_order_info_label_d_whole) que se sobrescriben cada día, lo que obliga a usar snapshots recientes para mantener consistencia entre fuentes.
- **Tablas estáticas o de referencia** (como dim_city), usadas para relacionar identificadores con variables descriptivas.

Además, fue necesario unir el r_segment proveniente de otra línea del negocio y alinear fechas para evitar duplicidad de registros. Esto implicó manejar múltiples uniones, condiciones de negocio y validaciones de consistencia antes de generar la base final.

El resultado fue un dataset limpio y consolidado, con un registro por usuario que resume su comportamiento, afinidades y nivel de actividad, listo para el análisis exploratorio.

Diccionario de datos:

Variable	Descripción	Tipo	Fuente
uid	Identificador único del usuario (anonimizado)	Numérico	Transaccional
country_code	País del usuario	Categórica	Transaccional
city_token	Ciudad del usuario (tokenizada)	Categórica	dim_city
total_orders	Total de órdenes completadas	Numérica	dwm_finance_order_d_increment
total_orders_tmenos1	Total de órdenes en el corte anterior	Numérica	dwm_user_order_accumulate_by_bizline_d_whole
delta_orders	Diferencia entre órdenes actuales y anteriores	Numérica	Derivada
categoria_recencia	Nivel de recencia según última orden (≤7d, 8–14d, etc.)	Categórica	Derivada
efo_to_four	Días entre la primera y cuarta orden	Numérica	dwm_finance_order_d_increment
r_segment	segment_r (Loyal, Casual, Rare)	Categórica	ssl_freq_rider_segmentation
main_category_counts	Conteo de órdenes por categoría	JSON/dict	orders_enriched
ka_type_counts	Conteo de órdenes por tipo de tienda	JSON/dict	orders_enriched
shop_name_counts	Conteo de órdenes por tienda	JSON/dict	orders_enriched
brand_name_counts	Conteo de órdenes por marca	JSON/dict	orders_enriched

7. Conclusiones Iniciales

7.1 Logros de la Primera Entrega

Se completó exitosamente el entendimiento del negocio y los datos, cumpliendo los objetivos propuestos.

Se definió claramente la problemática (falta de esquema de priorización en usuarios nuevos de Engagement) y se diseñó el producto de datos compuesto por Dashboard, Modelo Predictivo y Recomendador.

Los datos fueron recolectados, validados y analizados con métodos univariados y multivariados, considerando además los aspectos éticos y regulatorios requeridos por la normativa colombiana.

7.2 Principales Hallazgos

El análisis confirmó tres factores claves de crecimiento:

Recencia: es el predictor más fuerte; los usuarios activos crecen 7 veces más que los inactivos.

Velocidad de adopción: correlación negativa con crecimiento (r=-0.201); los usuarios que llegan a su cuarta orden en menos de 14 días crecen 2.3 veces más.

Segmento R: el grupo r_segment002 presenta consistentemente mejor desempeño.

En cuanto al comportamiento, los usuarios muestran alta exploración y baja lealtad (96.9% compra en múltiples tiendas) y una fuerte concentración en pocas categorías (6 categorías = 80% de las órdenes). Además, el fin de semana concentra 35.8% de la actividad, sugiriendo oportunidades para campañas temporales.

7.3 Validación de Hipótesis y Suficiencia de Datos

Las tres hipótesis fueron validadas estadísticamente, confirmando la viabilidad del enfoque analítico.El dataset cuenta con variables suficientes (recencia, efo_to_four, segmento, afinidades) y robustez para desarrollar los tres componentes del producto: dashboard interactivo, modelo predictivo y sistema de recomendación.

7.4 Próximas Acciones

Las siguientes etapas se centrarán en:

Preparación de datos: creación de nuevas variables (afinidades, interacciones), tratamiento de outliers y codificación categórica.

Modelado predictivo: clasificación para usuarios de alto crecimiento y regresión para estimar delta_orders.

Construcción del producto: dashboard interactivo y recomendador basado en afinidades y predicciones.

Evaluación y retroalimentación: validación cruzada, comparación de modelos y revisión con stakeholders.

7.5 Riesgos y Métricas de Éxito

Riesgos como la alta dimensionalidad y el desbalance de la variable objetivo se mitigarán mediante selección de características, SMOTE y regularización. Las métricas objetivo son: AUC-ROC > 0.75, RMSE < 3.5 órdenes y que el top 20% predicho capture >40% del crecimiento.

7.6 Resumen Ejecutivo

La primera entrega permitió caracterizar a los nuevos usuarios de Engagement y validar los factores que predicen su crecimiento.

Se confirma que la recencia y la velocidad de adopción son determinantes del desempeño futuro y que los usuarios presentan poca lealtad pero patrones predecibles por afinidad y tiempo. El dataset tiene calidad y variabilidad suficientes para construir modelos robustos, desarrollar el sistema de recomendación y crear un dashboard interactivo que oriente decisiones basadas en datos. El siguiente

paso será la preparación de datos y modelado predictivo, priorizando la identificación de usuarios de alto potencial.

Referencias

1. (Superintendencia de Industria y Comercio) - Superintendencia de Industria y Comercio. “Guía oficial de protección de datos personales.” *Superintendencia de Industria y Comercio – Protección de Datos Personales*, 10 10 2023, https://habeasdata.todoenuno.net.co/wp-content/uploads/2023/10/SuperIndustria-publico-la-Guia-oficial-de-proteccion-de-datos-personales_compressed.pdf?utm_source=chatgpt.com. Accessed 16 10 2025.