Entrega Final Proyecto Final

Juan David Valencia - 201728857
Juan Esteban Cuellar - 202014258

1. Resumen ejecutivo

El proyecto se desarrolla sobre datos de una plataforma de delivery y busca responder una pregunta concreta del equipo de Engagement: ¿cómo priorizar a los usuarios nuevos con mayor potencial de crecimiento en órdenes, para invertir mejor el presupuesto promocional?

Nos enfocamos en usuarios que alcanzan su cuarta orden en la plataforma. En ese momento ya tenemos suficiente historia como para caracterizarlos, pero todavía estamos a tiempo de influir en su comportamiento. A partir de su actividad previa calculamos un indicador de crecimiento futuro en órdenes (delta_orders) y definimos como usuarios de alto crecimiento (high_growth) a quienes quedan aproximadamente en el top 20 % de esa distribución. Sobre esta definición entrenamos un modelo de clasificación supervisado que predice la probabilidad de que cada usuario sea de alto crecimiento.

El modelo se nutre de variables que resumen cinco dimensiones del comportamiento: volumen y actividad, recencia, velocidad de adopción, diversidad de consumo y afinidades (categoría/tienda/marca dominante), además de la segmentación R y el contexto geográfico. Con estas variables logramos un modelo con buen desempeño en validación (medido con AUC-ROC, F1 y precision@top-20 %), que permite discriminar con alta precisión qué usuarios tienen mayor probabilidad de crecer.

Sobre este modelo diseñamos un producto de datos orientado a negocio: un tablero que combina métricas agregadas, un ranking de usuarios con su score de high_growth y filtros por afinidades, país, ciudad y segmento. El objetivo es que el equipo de Engagement pueda decidir a quién contactar, con qué tipo de incentivo y en qué momento, y que este proceso deje de ser una segmentación manual basada solo en reglas estáticas.

Finalmente, discutimos las implicaciones de uso responsable de datos, las principales dificultades del proyecto (integración de fuentes, definición de KPIs y riesgo de sobreajuste) y posibles extensiones, como conectar el modelo con experimentos A/B que midan impacto real en costo por orden incremental.

2. Contexto, problema de negocio y KPIs

2.1. Contexto y alcance

La plataforma cuenta con un equipo de Engagement responsable de que los usuarios nuevos no se queden solo en sus primeras compras, sino que se conviertan en usuarios recurrentes. Tradicionalmente, los incentivos (bonos, descuentos, comunicaciones) se aplican de manera relativamente homogénea sobre una base amplia, sin una priorización sistemática de quién tiene más potencial.

En este proyecto acotamos el problema a un grupo específico: usuarios que alcanzan su cuarta orden dentro de una ventana de observación. A partir de ese hito, medimos qué tan intensivamente siguen usando la plataforma durante los tres meses siguientes y nos preguntamos si es posible anticipar ese comportamiento usando únicamente la información disponible hasta ese punto.

Aunque en los datos trabajamos con un período concreto, la idea es que el pipeline y el modelo puedan ejecutarse de manera periódica sobre nuevas grupos de usuarios que van alcanzando su cuarta orden.

- **Órdenes históricas** (total_orders): número de órdenes realizadas por el usuario hasta la fecha de corte en la que alcanza su cuarta orden.

- **Velocidad a la cuarta orden** (efo_to_four): días transcurridos entre la primera y la cuarta orden. Velocidades bajas indican adopción rápida de la plataforma.

- **Recencia**: número de días desde la última orden hasta una fecha de referencia. A partir de este valor, clasificamos al usuario en categorías de recencia (por ejemplo, “caliente”, “tibio”, “frío”, “perdido”), útiles para interpretación.

- **Crecimiento en órdenes** (delta_orders):
  delta_orders= Ordenes en los 3 meses despues de la cuarta orden - ordenes previas
- **Usuario de alto crecimiento** (high_growth): variable objetivo binaria. Definimos high_growth = 1 para usuarios con delta_orders superior a un umbral que corresponde aproximadamente al **percentil 80** de la distribución (top ~20 %). El resto toma valor 0.
  Esta decisión refleja la restricción de negocio: el equipo no puede impactar a todos, sino a una fracción priorizada.

Además, aunque no lo calculamos directamente en el modelo, el proyecto está enmarcado en la idea de **costo por orden incremental (CPOI)**:

CPOI = inversion en incentivos / ordenes adicionales generadas

Predecir high_growth nos ayuda a reducir el CPOI, dirigiendo los incentivos hacia usuarios con mayor probabilidad de generar órdenes adicionales.

2.3. Preguntas de negocio e hipótesis

Las preguntas que guían el trabajo son:

1. ¿Qué características observables en el momento de la cuarta orden se asocian con un crecimiento alto en órdenes en los meses siguientes?
2. ¿Podemos usar esa información para construir un modelo que priorice qué usuarios deberían recibir incentivos?
3. ¿Cómo se integraría esa predicción en el flujo de trabajo habitual del equipo de Engagement?

De estas preguntas surgen tres hipótesis principales:

- H1 – Recencia y velocidad: usuarios más recientes y que llegaron más rápido a la cuarta orden tienen mayor probabilidad de seguir creciendo.
- H2 – Diversidad y afinidades: patrones de diversidad (número de tiendas/categorías) y de afinidad (categoría/tienda dominante) se asocian con distintos niveles de crecimiento.
- H3 – Segmentación existente: la segmentación R definida por el negocio (r_segment) contiene información relevante, que el modelo puede complementar pero no reemplazar.

El modelo y el análisis exploratorio buscan precisamente medir hasta qué punto estas hipótesis se cumplen.

3. Datos y preparación

3.1. Fuentes y construcción del dataset

Trabajamos con varias fuentes internas de la plataforma:

- Tablas transaccionales de pedidos, con una fila por orden.
- Tablas de usuarios, con información demográfica básica y segmentación R.
- Tablas de referencia para categorías, marcas, tipo de tienda y ciudades.

A partir de estas fuentes se construye un dataset consolidado a nivel usuario, que contiene:

- Identificador de usuario y país/ciudad.

- Fechas de primera y cuarta orden, y fecha de última orden observada.
- Número total de órdenes y métricas de actividad.
- Contadores por categoría, tienda y marca (como estructuras tipo diccionario).
- Segmento R asignado por el negocio.
- Cálculo de delta_orders en el horizonte de 3 meses posteriores a la cuarta orden.

Tras limpieza básica (eliminación de duplicados, manejo de nulos y verificación de consistencia en fechas), se obtiene una base de decenas de miles de usuarios con información suficiente para análisis y modelado.

3.2. Análisis exploratorio

El análisis exploratorio se centra en entender:

- La distribución de delta_orders y el efecto de fijar el umbral de high_growth en el top ~20 %.
- Cómo se comportan recencia, efo_to_four y total_orders en usuarios de alto vs bajo crecimiento.
- Qué categorías, tipos de tienda y marcas aparecen con más frecuencia en cada grupo.

En general observamos que:

- Los usuarios de alto crecimiento tienden a tener recencias menores y velocidades a la cuarta orden más rápidas.
- Suelen presentar mayor diversidad de tiendas y categorías, aunque manteniendo una o dos categorías dominantes.
- La segmentación R introduce diferencias de nivel en el comportamiento, pero dentro de cada segmento sigue habiendo variación que el modelo ayuda a capturar.

Estos hallazgos motivan la construcción de variables que resuman volumen, recencia, velocidad y diversidad, en lugar de trabajar solo con contadores crudos.

3.3. Feature engineering y variable objetivo

En el notebook de preparación se implementa un pipeline que transforma el dataset crudo en un conjunto de variables más interpretables:

- Volumen y actividad
    - total_orders y transformaciones logarítmicas para estabilizar distribuciones (log_total_orders). Creamos log_total_orders para “aplanar”

la cola larga de total_orders: la mayoría tiene pocas órdenes y unos pocos muchísimas. El cambio de esta variable se puede ver en el anexo 1.
    - Con el log la distribución queda menos sesgada y el modelo aprende mejor sin que esos valores extremos dominen.
    - orders_per_day como razón entre órdenes y tiempo activo.

- Velocidad y recencia
    - efo_to_four escalada (días entre primera y cuarta orden). Como se puede ver en el anexo 2 la variable efo_to_four antes está en días (promedio ~15), y después del scaling queda centrada en 0 y en unidades "estándar", manteniendo la forma de la distribución pero cambiando la escala.
    - days_since_first_order.
    - categoria_recencia y/o recencia numérica a la fecha de corte.

- Diversidad y lealtad

    - Índices de diversidad tipo Shannon para categorías y tiendas.
    - Número de categorías y tiendas distinto, normalizados por volumen (por ejemplo, categorías por orden).
    - Indicadores binarios como is_multi_category (usa ≥ 3 categorías) e is_multi_shop (compra en ≥ 5 tiendas).
    - dominant_category y dominant_category_ratio para capturar afinidad principal.

- Segmentación y contexto

    - r_segment como categoría codificada.
    - city_token o city_id como proxy geográfico.

La variable objetivo se define como:

- high_growth = 1 si delta_orders supera el umbral elegido (aprox. percentil 80).
- high_growth = 0 en otro caso.

Como se puede evidenciar en el anexo 3 se ve que la mayoría de usuarios no son high_growth y solo una parte más pequeña queda en la clase 1.

Además, cuando el usuario sí es high_growth, su delta_orders es muchísimo más alto y dispersa: el grupo 0 se queda cerca de pocas órdenes extra, mientras que el grupo 1 tiene medianas altas y muchos casos con decenas de órdenes adicionales.

Guardamos también delta_orders como objetivo continuo secundario, útil para análisis complementarios.

Finalmente, se divide la base en tres subconjuntos estratificados:

- Train (60 %)
- Validation (20 %)
- Test (20 %)

Como se puede ver en el anexo 4 nos aseguramos que la proporción de high_growth sea similar en los tres grupos y que las distribuciones básicas de variables clave se mantengan.

4. Enfoque analítico y modelado

4.1. Tarea de modelado

La tarea principal es una clasificación supervisada:

Dada la información del usuario hasta su cuarta orden, queremos predecir la probabilidad de que sea un usuario de alto crecimiento (high_growth = 1) en los tres meses siguientes.

En términos de negocio:

- Usuarios con alta probabilidad (score alto) son candidatos prioritarios para recibir incentivos, comunicaciones y propuestas de valor adicionales.
- Usuarios con probabilidad baja pueden recibir tácticas más ligeras o simplemente continuar en los flujos estándar.

4.2. Estrategia de validación

Sobre el conjunto de entrenamiento se entrena y valida mediante validación cruzada (5-fold) distintos algoritmos:

- Random Forest
- XGBoost

- LightGBM (siempre que la capacidad computacional lo permitió)

La métrica objetivo del GridSearch es AUC-ROC, por su robustez frente al desbalance (aprox. 20 % de positivos).

Como métricas complementarias se utilizan:

- F1-score global.
- Precision top-20 % (qué proporción de usuarios marcados como high_growth hay dentro del 20 % con mayor score), muy alineada con el caso de uso “tengo presupuesto para impactar solo al top-20 %”.
- Matriz de confusión para analizar errores.

El Test set queda totalmente reservado para la evaluación final del modelo elegido.

4.3. Resultados y selección de modelo

Tras entrenar las tres familias de modelos con un conjunto controlado de hiperparámetros, comparamos su desempeño en el Validation set:

- En general, XGBoost y LightGBM ofrecen mejor AUC-ROC y F1 que el Random Forest, con tiempos de entrenamiento aceptables.
- La Precision@top-20 % es particularmente relevante: los modelos de boosting logran una proporción muy alta de usuarios realmente high_growth dentro de ese 20 % superior de scores, lo que hace muy eficiente el uso de recursos promocionales.

Con base en estas métricas se elige un modelo de gradient boosting como candidato final. Este modelo se reentrena con los mejores hiperparámetros sobre Train+Validation y se evalúa una sola vez en el Test set, donde mantiene niveles altos de AUC-ROC y buena combinación de precisión y recall.

4.4. Importancia de variables e interpretación

Tomamos el modelo final de clasificación (XGBoost) entrenado con todas las variables candidatas y usamos la importancia que entrega el propio modelo (basada en cuánto mejora la calidad de las divisiones cuando usa cada feature). Con esas importancias armamos un ranking y nos quedamos con las 15 variables con mayor peso. Luego revisamos ese top a mano, para quitar columnas casi duplicadas y priorizar las que tienen una lectura clara desde negocio (por ejemplo, is_multi_shop es más fácil de explicar que un conteo crudo). Sobre ese conjunto reducido es que se hace la interpretación que sigue. Viendo los resultados se puede decir que las features más

relevantes no son tanto las de volumen bruto, sino las que describen qué tan “amplia” es la relación del usuario con la plataforma.

Las principales variables son:

1. is_multi_shop – Es la variable más importante (≈ 48 % de la importancia total).

   Indica si el usuario compra en muchas tiendas distintas (por encima de un umbral definido). El modelo aprende que los usuarios que reparten sus órdenes entre varias tiendas tienen mucha mayor probabilidad de ser de alto crecimiento. En términos de negocio, esto sugiere una relación más “profunda” con la plataforma y menos dependencia de un solo restaurante.

2. is_multi_category – Segunda variable en importancia (≈ 16 %).

   Señala si el usuario compra en varias categorías (por ejemplo, comida rápida, cafés, mercado, etc.). Los usuarios multicategoría parecen estar más integrados en distintos momentos de consumo (almuerzo, snacks, mercado del hogar, etc.), lo que aumenta su probabilidad de crecimiento.

3. shop_diversity, shops_per_order y categories_per_order – En conjunto, estas tres variables explican otra buena parte de la importancia del modelo.

   shop_diversity captura cuán repartido está el consumo entre tiendas.

4. shops_per_order y categories_per_order normalizan la diversidad por el volumen, evitando confundir “diversidad” con “muchos pedidos”.

   El modelo refuerza la idea de que no solo importa cuánto ordena el usuario, sino cómo reparte esas órdenes.

5. category_diversity y dominant_category_ratio – Combinan diversidad con “foco”:

   category_diversity alto indica que el usuario explora varias categorías.

6. dominant_category_ratio mide cuánto peso tiene la categoría principal.

   El hecho de que ambas aparezcan como relevantes sugiere que los usuarios high_growth tienden a tener una categoría favorita clara, pero sin dejar de probar otras.

7. Variables temporales y de contexto:
   - days_since_first_order: el tiempo desde la primera orden ayuda a distinguir usuarios que consolidan rápido su relación con la plataforma.

    - categoria_recencia_Frío (31–90d): la pertenencia a esta banda de recencia aporta señal sobre usuarios que podrían estar “en riesgo” si no se activan, o que ya empezaron a enfriarse.
    - city_token_city002 y city_token_city003: capturan diferencias de comportamiento por ciudad (penetración, oferta, hábitos locales).
8. Variables de afinidad específica:
- Dummies como dominant_category o main_category indican que ciertas categorías dominantes están asociadas de forma sistemática con mayor probabilidad de high_growth.

En la práctica, esto ayuda a identificar “perfiles tipo”: por ejemplo, usuarios cuya categoría dominante es X tienden a ser más (o menos) valiosos a futuro.

En conjunto, la lectura de importancia de variables respalda una idea central:

Los usuarios de alto crecimiento no son solo los que más ordenan hoy, sino los que construyen una relación amplia con la plataforma (multitienda, multicategoría, con cierta diversidad) y lo hacen en un plazo razonable desde su primera orden.

Esta interpretación es directamente accionable: el equipo de Engagement puede pensar en estos usuarios como “candidatos naturales” para campañas que consoliden esa relación (beneficios para nuevas tiendas, cross-selling entre categorías, etc.).

5. Producto de datos y uso en el negocio

La idea es que este modelo sea funcional y facil de usar para los usuarios operativos que envían los incentivos. El producto final combina el motor de scoring con una capa de visualización y una lógica sencilla de segmentación accionable.

El producto se apoya en un motor de scoring que, para los usuarios que llegan a su cuarta orden en un periodo dado, calcula las variables de actividad, recencia, velocidad, diversidad de consumo y afinidades, y las pasa al modelo que estima la probabilidad de que cada uno sea high_growth. De este proceso sale una tabla con un identificador interno, el score, banderas como is_multi_shop e is_multi_category, medidas de diversidad y datos de contexto (categoría dominante, ciudad, segmento). Con eso no solo se puede ordenar a los usuarios por potencial de crecimiento, sino entender rápidamente qué tipo de relación tienen hoy con la plataforma (monotienda vs multitienda, monocategoría vs multicategoría).

Sobre esa tabla se construye la aplicación Growth Predictor. En el Dashboard Ejecutivo (Anexo 8) se ve el tamaño de la base, el porcentaje de high_growth, las órdenes adicionales promedio y las variables que más pesan en el modelo. El Explorador de

Segmentos (Anexo 9) permite filtrar por recencia y segmento y comparar la distribución de las features entre grupos. La vista de Predicción (Anexo 10) muestra, para un usuario puntual, su probabilidad de high_growth, la clasificación final y una recomendación de prioridad. Finalmente, el módulo de Análisis de Afinidades (Anexo 11) resume las categorías en las que más consumen los usuarios y cómo cambia la diversidad entre perfiles.

Recomendación de acciones según perfil

A partir de las variables que el modelo considera más importantes y de las vistas anteriores, se propone una lógica sencilla de acciones:

- Multitienda + multicategoría + score alto: usuarios “core” de la plataforma. Priorizar con beneficios de retención (programas de lealtad, beneficios por frecuencia, acceso anticipado a nuevas funcionalidades).
- Multitienda pero monocategoría: alta dependencia de un tipo de consumo. Buenos candidatos para campañas de cross-selling hacia categorías complementarias.
- Monotienda con score alto: relación intensa con una sola tienda. Tienen sentido alianzas o campañas co-brandeadas con esa tienda, para capturar valor sin forzar un cambio brusco de hábito.
- Usuarios en recencia “Frío (31–90 días)” con score moderado o alto: segmento ideal para campañas de reactivación antes de que migren a estados de pérdida definitiva.

Estas reglas no sustituyen el juicio del equipo, pero entregan una base estructurada para traducir los insights del modelo y del dashboard en decisiones concretas de comunicación e incentivos.

5.2. Flujo operativo propuesto

A un nivel conceptual, el flujo operativo del producto sería:

1. De forma periódica (por ejemplo, semanalmente) se identifica a los usuarios que acaban de alcanzar su cuarta orden.
2. Se ejecuta el pipeline de construcción de features y se aplica el modelo para generar los scores de high_growth y los atributos de perfil (multitienda, multicategoría, afinidades, recencia).
3. Los resultados se almacenan en una tabla de scores accesible desde el dashboard corporativo.
4. El equipo de Engagement ingresa al dashboard, explora los perfiles y selecciona segmentos, por ejemplo:

    a. top-20 % de score en ciudad A,
    b. dentro de ese grupo, usuarios multitienda con recencia tibia o fría,
    c. o bien usuarios de alto score pero monocategoría, para campañas específicas.
5. Estos segmentos se exportan o se integran con las herramientas de marketing para ejecutar campañas.
6. En una fase posterior, las respuestas a esas campañas (órdenes adicionales, uso de cupones, etc.) se pueden registrar y utilizar para refinar el modelo y ajustar las reglas de negocio.

Este flujo está pensado para que el modelo deje de ser un ejercicio puramente académico y se conecte con decisiones cotidianas del equipo de Engagement.

6. Uso responsable de los datos

En este proyecto usamos datos de comportamiento (qué piden, cuándo, dónde y qué tipo de restaurantes), siempre anonimizados y alineados con políticas de privacidad y normativa de protección de datos. El objetivo del modelo es mejorar la experiencia y asignar mejor los beneficios, no tomar decisiones sensibles sobre personas. Reconocemos el riesgo de sesgos (favorecer a quienes ya piden más o a ciertas ciudades), por lo que el score debe ser una señal adicional combinada con reglas de negocio que garanticen cobertura a segmentos estratégicos. Además, el modelo se entrenó con un grupo y periodo específicos, así que su uso debe acompañarse de reentrenamientos periódicos, validaciones con nuevas grupos y, idealmente, pruebas A/B para asegurar que siga siendo útil y no se convierta en una “caja negra” incuestionable.

7. Retroalimentación, dificultades y trabajo futuro

A lo largo del curso el proyecto fue cambiando con la retroalimentación. Al principio la formulación era más amplia: se hablaba de clasificación, regresión, posibles clusters y varios KPIs que no estaban tan bien aterrizados. Con los comentarios del profesor quedó claro que había que ordenar la historia: definir mejor qué es recencia, cómo se mide el crecimiento en órdenes y escoger una tarea principal. A partir de eso, el foco pasó a ser predecir qué usuarios son de alto crecimiento (high_growth).

Algo parecido pasó con las variables. Las primeras versiones usaban muchos conteos crudos (cuántas tiendas distintas, cuántas categorías distintas, etc.), que en el fondo se movían casi igual que el número de órdenes. Con el tiempo fuimos reemplazando esos conteos por índices de diversidad, variables normalizadas por orden y flags como is_multi_shop e is_multi_category. Cuando miramos la importancia de variables del modelo final, justamente estas nuevas features quedaron arriba del todo. Eso ayudó a

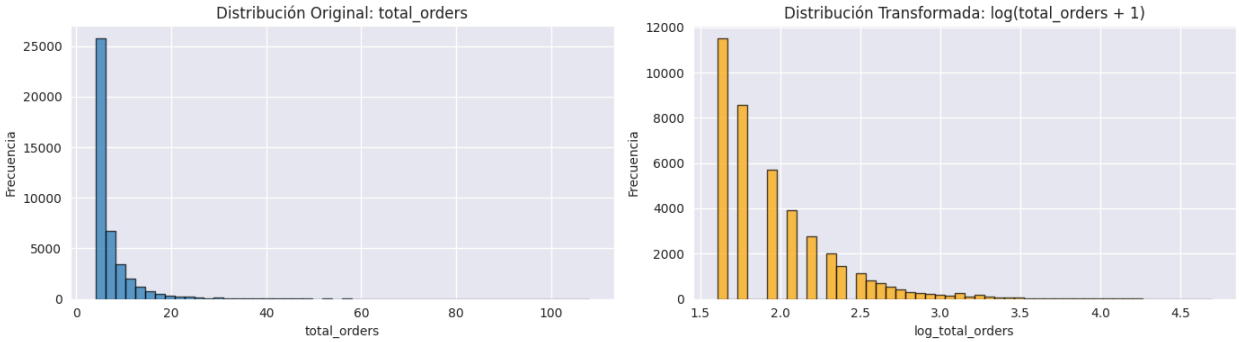
darle una lectura más clara: los usuarios valiosos no son solo los que piden mucho, sino los que usan la plataforma en más contextos (más tiendas, más tipos de comida).

El proyecto deja varias cosas útiles. Por un lado, un pipeline completo que va desde los datos crudos hasta un modelo que ayuda a priorizar usuarios para campañas. Por otro, un insight claro: los usuarios de alto crecimiento se parecen más a “usuarios multitienda y multicategoría” que simplemente a “heavy users”. Esto abre ideas de acción: por ejemplo, diseñar campañas que acompañen a usuarios monotienda a probar nuevas tiendas, o que empujen a usuarios monocategoría a explorar otros momentos de consumo.

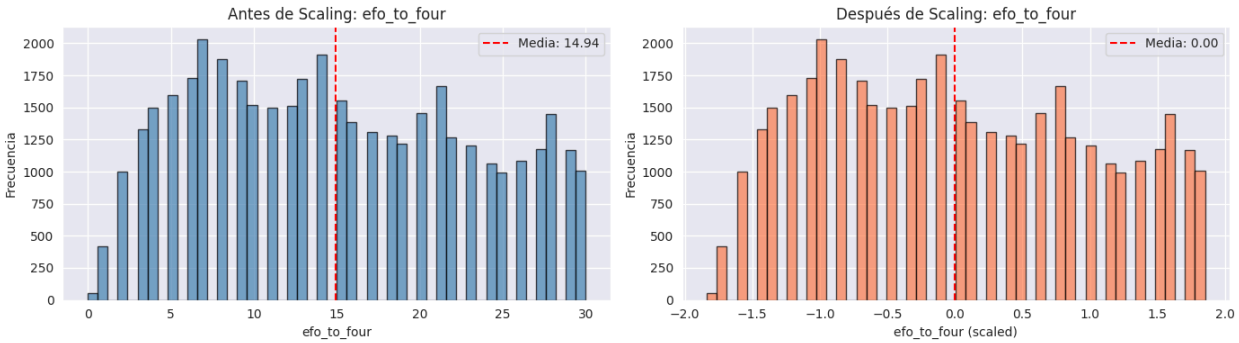
Como pasos siguientes, tendría sentido probar el modelo con un grupo más recientes y, si es posible, hacer splits temporales más estrictos para simular mejor un uso en producción. También sería interesante conectar el score con campañas reales y medir su impacto en órdenes adicionales y costo por orden incremental. Y, si el tiempo lo permite, explorar modelos complementarios (regresión sobre delta_orders, modelos de supervivencia, segmentaciones no supervisadas) usando las mismas variables de diversidad y afinidad. Eso convertiría este trabajo en un buen primer MVP analítico, con camino para crecer más adelante.

Anexos

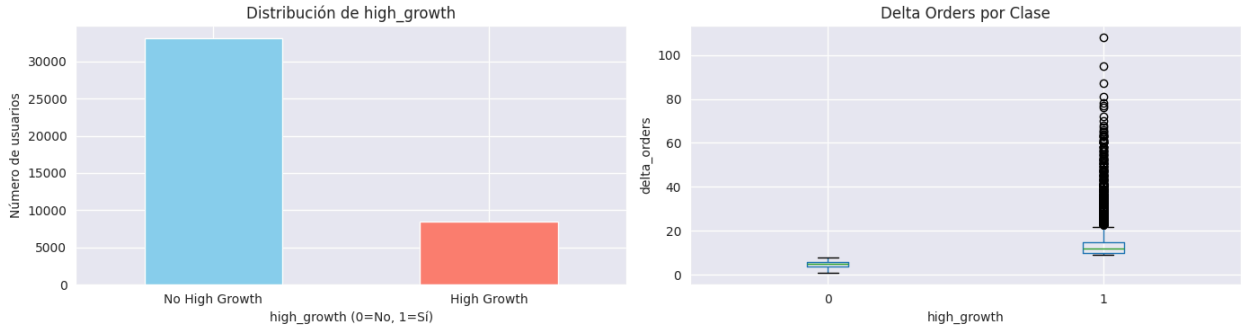
Anexo 1



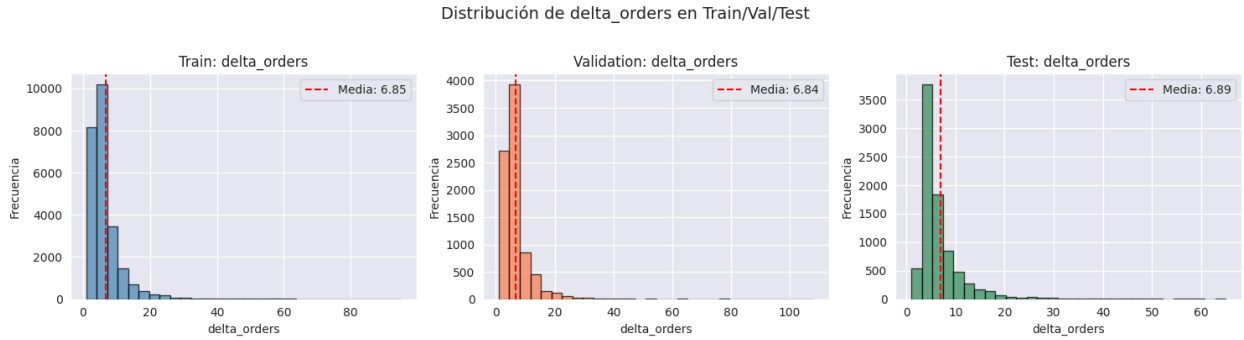
Anexo 2



Anexo 3



Anexo 4



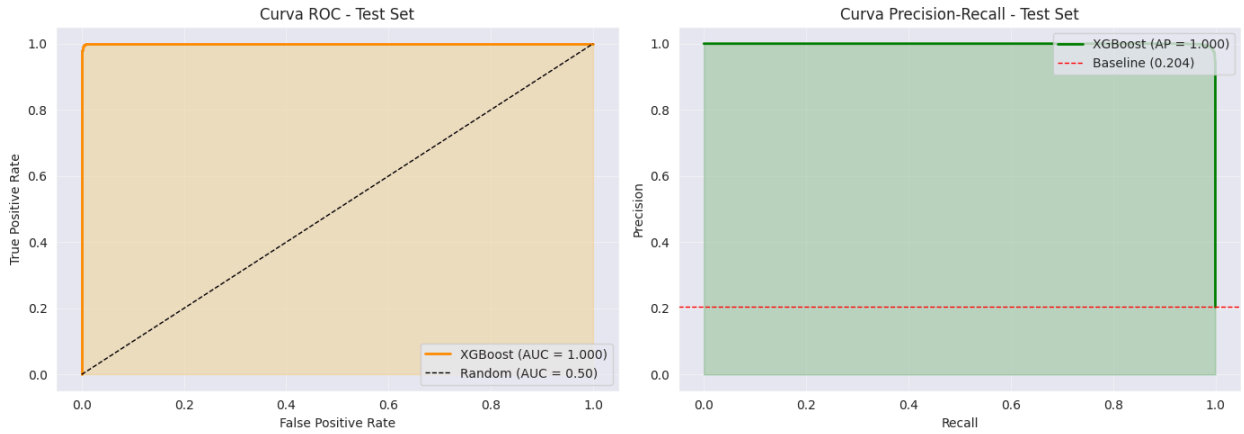
Anexo 5

📊 TABLA COMPARATIVA:

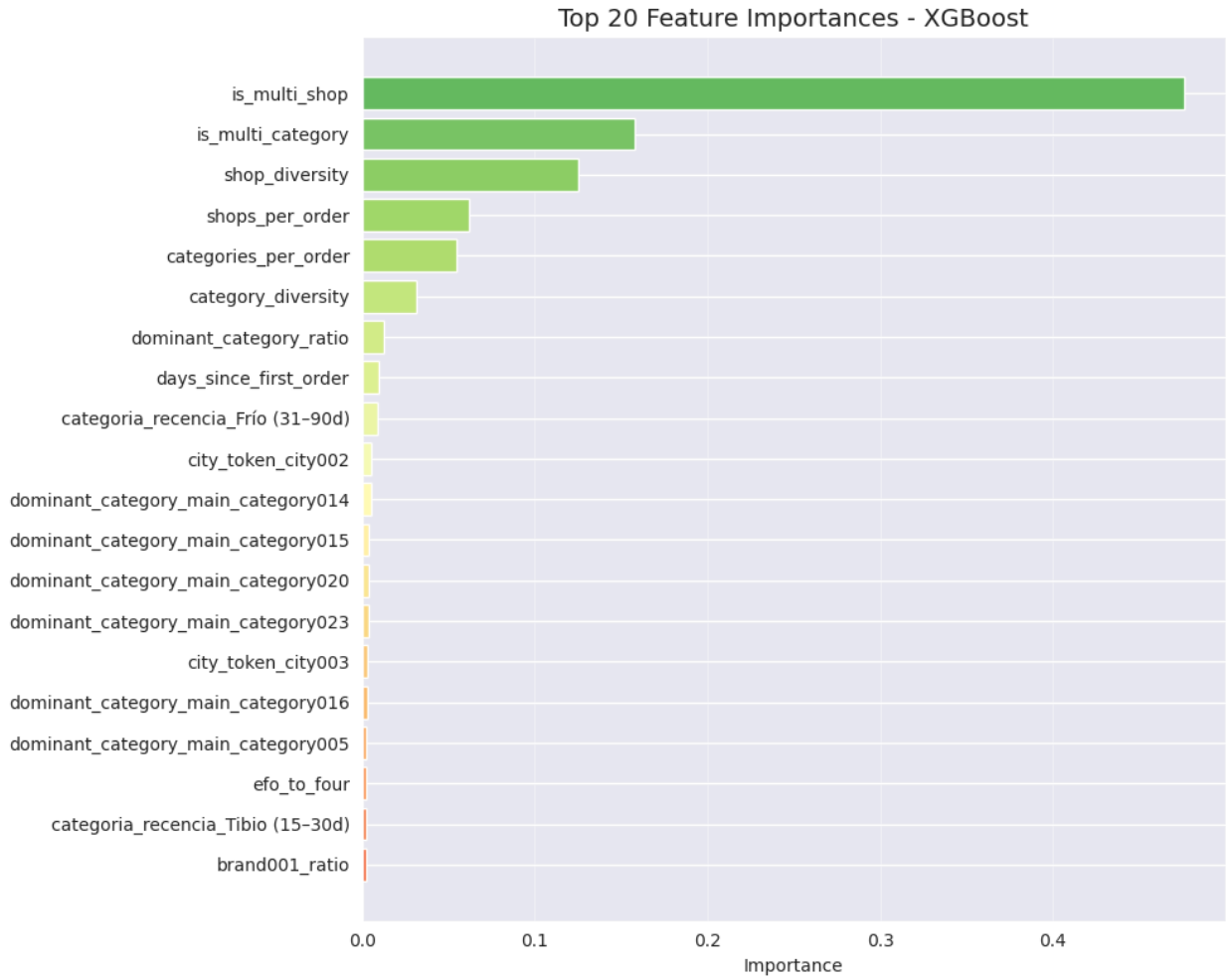
Métrica	Random Forest	XGBoost	Mejor
AUC-ROC	0.9982	0.9998	XGB
Precision@20%	0.9670	0.9934	XGB
F1-Score	0.9614	0.9868	XGB
Precision	0.9324	0.9808	XGB
Recall	0.9923	0.9929	XGB
Accuracy	0.9838	0.9946	XGB

📈 Resumen de victorias: RF=0 | XGB=6

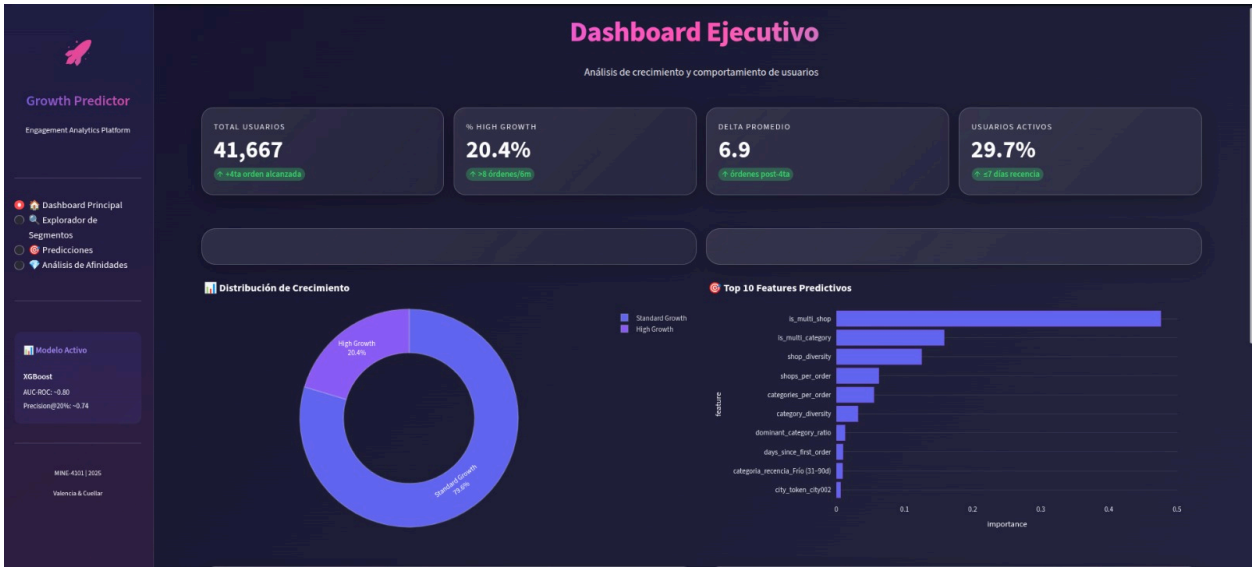
Anexo 6



Anexo 7



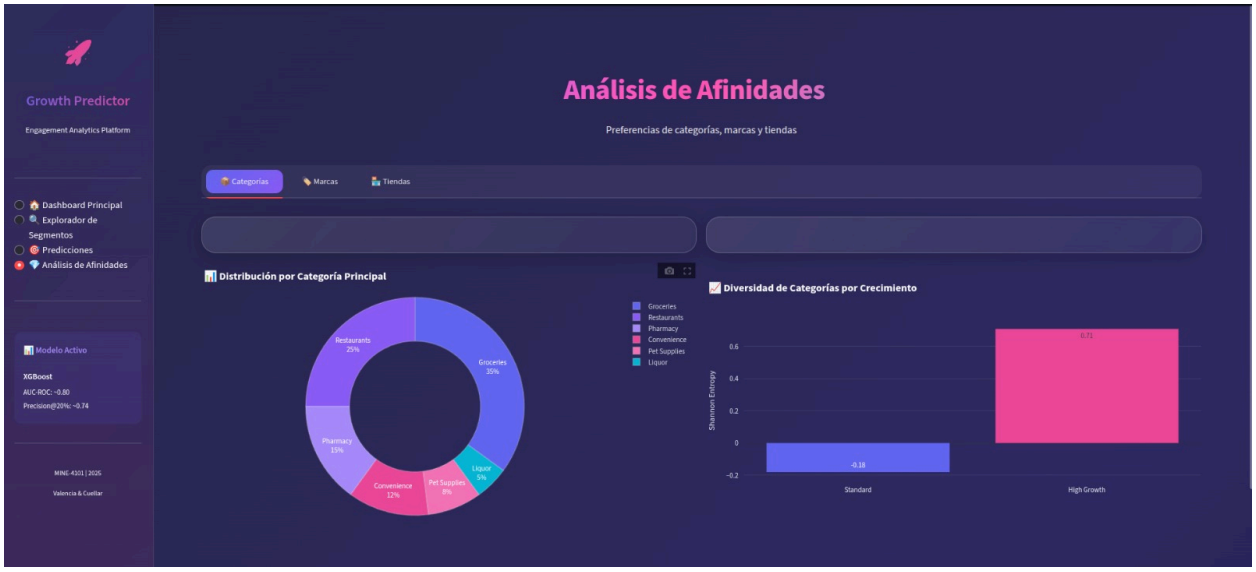
Anexo 8



Anexo 9

Anexo 10

Anexo 11



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.