PROYECTO FINAL CIENCIA DE DATOS APLICADA

1. DEFINICIÓN DE LA PROBLEMÁTICA Y ENTENDIMIENTO DEL NEGOCIO

A partir del ejercicio de investigación de las empresas que se tuvieron en cuenta al momento de elaborar el proyecto, se determinó abordar un problema particular de la empresa Frubana, empresa de alimentos que utiliza tecnología e innovación en logística para conectar directamente a agricultores y consumidores. La misión de Frubana es hacer la comida de Latinoamérica más barata y ser el servicio "todo en uno" de los restaurantes de la región.

Fundada en 2018 con operación en tres países (Colombia, Brasil, México), Frubana se ha convertido en poco tiempo en una empresa altamente escalable con alto capital de inversión por parte de fondos especializados. Actualmente la compañía cuenta con operación en 10 ciudades de Colombia, México y Brasil, y cuenta con inversión de cerca de 270 millones de dólares por parte de diferentes grupos inversionistas. Atiende a cerca de 35.000 clientes y cuenta con un equipo de cerca de 1.000 personas en oficinas y 4.000 en bodegas (Forbes, 2023)[1]

A partir de la atención a restaurantes y a agricultores, se busca generar un ambiente mejor remunerado para cada agente que interviene en el proceso, incluyendo la utilización de tecnología escalada que permita tomar decisiones con un mayor volumen de datos y abaratamiento del proceso entre productores y consumidores.

A través del método de entrevista directo como método de recolección de datos, se utilizó la información obtenida desde el área de negocio de Frubana para identificar una necesidad puntual de retención de un subgrupo de clientes de la compañía que son una población de interés para que, a través de los datos, se genere un modelo de retención que tenga la capacidad de predecir una posible salida de estos usuarios y de esta manera lograr anticipar su posible perdida y tomar medidas de retención.

El modelo de predicción solamente sería utilizado sobre los clientes en cuyo valor de compra (ticket promedio) sea significativo, teniendo en cuenta información previa que se cotejará en el Análisis Exploratorio de Datos. Como datos adicionales, el negocio manifiesta que el porcentaje de compra y abandono oscila entre el 15% de clientes mensuales. A nivel corporativo, se cuenta con metas de retención mensuales que buscan tener el mayor número de clientes activos en cada periodo. Los porcentajes de abandono varían respecto a cohortes (clasificación interna) pero no a una cifra propia de retención mensual.

2. IDEACIÓN DEL PROBLEMA DE NEGOCIO

[1] Forbes. (2023, 27 de febrero). Frubana logra entrada a la exclusiva red de emprendedores Endeavor, que ya tiene 60 compañías colombianas. Forbes. https://forbes.co/2023/02/27/emprendedores/frubana-logra-entrada-a-la-exclusiva-red-de-emprendedores-endeavor-que-ya-tiene-60-companias-colombianas

A raíz de la información proporcionada por el negocio a través de la entrevista, se confirma la existencia de equipos dedicados a explotar el producto de datos propuesto para desarrollar. A saber, existen los equipos de reactivación y comunicaciones. Para la primera área, existen equipos divididos en abandono y pérdida de costumbre. Con base en la necesidad del negocio, también se propone al grupo investigador el abordaje de los correspondientes a pérdida de costumbre en donde sea posible encontrar esos usuarios que tienen una tendencia decreciente e incentivarlos a través de un mecanismo de Frubonos[2]

El área de comunicaciones a su vez genera esquemas de comunicación directa y autorizada a través de correos electrónicos y/o mensajería de texto a dicha población con comportamiento decreciente y se busca reincluir dentro de la fuerza de ventas que en su momento cerró a los clientes.

Adicionalmente, el producto de datos obtenido involucrará al área interna de Frubana de Churn que se encargará de recibir el entregable y a partir de este generar estrategias de analítica avanzada para refinar cada uno de los impactos o incluirles un mayor volumen de información al modelo. A través del testimonio del negocio, fue posible también saber que anteriormente se hizo un ejercicio similar de Churn con un accuracy del 76%. Sin embargo, dicho modelo fue descartado dado que se generó una serie de inconvenientes basados en una incorrecta clasificación de los usuarios y de una ventana de tiempo muy alta en términos del impacto esperado del modelo elaborado.

De esta manera, los requerimientos del producto de datos están dado por las siguientes características:

- Una fase inicial de catalogación de los usuarios a través de etiquetas de cara a su comportamiento de compra, con el fin de encontrar aquellos usuarios de valor significativo que son la población objetivo de la necesidad del negocio.
- A partir de esta correcta estimación del grupo a estudiar, se buscará aplicar dos metodologías comparativas para el modelo de Churn: Una primera, basada en el algoritmo de Random Forest como técnica de aprendizaje supervisado y el uso de un algoritmo de XG Boost, basado en un sistema de ensamblado secuencial de árboles de decisión que identifiquen la población potencial a salir.

Todo lo anterior con el fin de encontrar aquella muestra en la que será posible predecir su salida de la compañía para que, a través de diferentes acciones posteriores a la entrega de estos resultados, sean contactados a través de diferentes estrategias y devueltos a su etiqueta de compra.

3. IMPLICACIONES DEL ABORDAJE DEL PROBLEMA DE NEGOCIO

[2] Basado en información propia de la compañía, los Frubonos se usan para el pago de domicilios. Dichos bonos corresponden a un sistema de créditos que se obtienen debido a alguna promoción, alguna falla en el domicilio, o a través del programa Frubana Capitals, que es la compañía de financiamiento de Frubana y en donde se utilizan como medio de cambio para posterior amortización.

Para el desarrollo de la actividad se obtuvieron datos de una organización legalmente constituida (Frubana) que garantiza el consentimiento informado de los clientes para recopilar y utilizar sus datos, así como la protección de su información personal para evitar accesos o divulgaciones no autorizadas, dando así cumplimiento con las leyes y regulaciones de protección de datos (Ley Habeas Data).

Así mismo, en comunicación con el negocio, hemos obtenido su autorización para acceder a un dataframe de 645.758 registros que no incluye información sensible de los clientes, como nombres, teléfonos o documentos de identidad. Es importante destacar que esta información será utilizada exclusivamente con fines académicos por parte del grupo de investigacion. Respetamos la confidencialidad y privacidad de los datos de los clientes, asegurando que se mantengan protegidos y no sean utilizados de manera inapropiada.

Las técnicas de IA que se plantean son en pro de generar valor al negocio y de aumentar la satisfacción del usuario, por lo que su fin, no irán en contravía de los derechos de los usuarios ni de las regulaciones vigentes.

4. ENFOQUE ANALÍTICO

La pregunta de negocio que se busca responder en el abordaje de Frubana está basada en encontrar aquella proporción de clientes que tienen la mayor probabilidad de abandonar la compañía, dado un comportamiento esperado que permita obtener primero un rotulamiento correcto de los clientes de alto valor que son críticos para la compañía; y segundo una probabilidad de abandono de dicho segmento para generar acciones de manera anticipada y evitar una fuga de usuarios no deseable.

Pregunta clave No. 1: ¿Qué segmento de clientes son esenciales para que la compañía active campañas de recuperación dada su importancia en facturación?

Usualmente al catalogar datos es posible contar con un enfoque de analítica visual que permita a los grupos de interés mencionados tener información de manera relevante y consolidada sobre los grupos de alto valor. De esta manera una primera necesidad se origina desde la capacidad de saber cuáles son los clientes de alto valor que son indispensables para no abandonar, haciendo uso de un tablero de visualización en donde se pueda tener la información de los clientes para tomar decisiones.

Adicionalmente, se utilizará un modelo basado en clasificación para catalogar a los clientes en diferentes categorías, como "propensos a churn" o "no propensos a churn". Para ello, se entrenaría un modelo con datos históricos de clientes que hayan abandonado la aplicación y clientes que se hayan mantenido activos. El modelo aprenderá a reconocer los patrones y características que diferencian a ambos estados, permitiendo predecir la probabilidad de churn para nuevos clientes.

Pregunta clave No. 2: ¿Qué clientes son propensos a abandonar la compañía?

Usualmente, se abordará la respuesta a la pregunta a partir del enfoque de modelos de aprendizaje automático supervisado, con los que es posible encontrar la siguiente división:

Se utilizará un modelo basado en el algoritmo de regresión para predecir la tasa de churn de los clientes en función de variables relevantes. Por ejemplo, se utilizará en variables para predecir la frecuencia de uso de la aplicación, el tiempo transcurrido desde la última compra, el número de quejas o reclamos realizados, entre otros.

5. RECOLECCIÓN DE DATOS

La información proporcionada por el negocio corresponde a la ubicación de la ciudad de Sao Paulo, en donde, parte del equipo investigador generó una serie de consultas realizadas a un set de datalakes con los que cuenta la compañía. El Query se dirigió a un Redshift en donde fueron almacenados los datos de una bodega de almacenamiento de alimentos que realiza entregas a 19.340 microzonas de la ciudad. A partir del ejercicio, se obtuvieron 645.758 observaciones que corresponden a pedidos realizados por los clientes en el periodo en 2023. (Ver Anexo 1)

6. ENTENDIMIENTO DE LOS DATOS

6.1. Calidad de datos

A partir de la información proporcionada por el negocio, se procedió con el Análisis Exploratorio de Datos para la población objetivo. Particularmente, se obtuvo la siguiente información[3] :

El grupo investigador generó un alistamiento previo de la información basado en la limpieza de datos faltantes en algunas columnas del dataset. Por razones de privacidad, este alistamiento no será divulgado, sino que se mantendrá como información interna de la compañía, toda vez que se solicitó de esta manera.

Adicionalmente, del total de poco más de 645 mil observaciones, se encontraron 269.133 registros válidos, con lo cual se corrigió el problema de duplicidad en la información.

Basado en el análisis exploratorio preliminar, no se cuenta con problemas de lógica asociada a posibles contradicciones de los datos. La información obtenida corresponde a información oficial del área de Churn de Frubana, razón por la cual está enmarcada dentro de los estándares de calidad de información de la compañía, gobierno de la información, privacidad y seguridad del área.

Se concluye que la información es relevante dado que corresponde a datos puntuales de pedidos de la empresa y a características propias del registro de sus activos digitales al momento de recibir la compra. La información es recolectada a partir de alimentos que se recogen en ubicaciones exactas de la ciudad de Sao Paulo, y son almacenadas a través de la infraestructura tecnológica de vanguardia proporcionada por la compañía para sus clientes.

[3] La información detallada del Análisis Exploratorio de Datos se encuentra como adjunto al repositorio de Github como anexo al presente documento.

La información resultante de 269 mil registros fue objeto de múltiples consultas, uniones, combinaciones y algunas transformaciones internas de la compañía, por lo que se puede afirmar que la información es procesable y será útil para el abordaje de los modelos mencionados. Los datos proporcionados son tanto de clientes activos como de clientes que ya se retiraron de la compañía con el propósito de tener información completa para el entrenamiento del modelo de machine learning.

Cabe resaltar que las características de disponibilidad, temporalidad y credibilidad fueron validadas por el área interna de negocio al momento de proporcionar la información al grupo investigador.

6.2. Análisis Exploratorio de los Datos

La estructura de los datos está dada por la tabla 1, contenida en el Anexo 1 de información de variables de interés, mientras que el detalle del ETA se encuentra en el Anexo 2 del documento y en los documentos soporte de Python.

- El conjunto de datos cuenta con 269.133 registros
- La columna *year_order* cuenta con un único valor: 2023, por lo que se infiere que los pedidos son del mismo año, así como la variable *city* en donde su único valor pertenece a la ciudad de Sao Paulo (SPO)
- El GMV tiene un rango que va desde aproximadamente 0 hasta 6,405.27, con una mediana de 70.34
- El número de paradas tiene valores que varían de 1 a 33, con una mediana de 5.
- La columna segmento tiene 6 valores únicos, donde "Restaurante" es el más frecuente.
- Existen 7 métodos de pago únicos, siendo "CASH" el más común.
- El campo birthday tiene 704 fechas únicas, siendo “2020-12-03" la más común
- La variable delivery tiene 207 fechas únicas, siendo “2023-01-27" la más común
- Algunas columnas como active, reactivation_status, city, warehouse y dispatch_warehouse tienen un solo valor único en su registro o están vacías.
- No hay datos faltantes en el conjunto de datos para ninguna columna.
- El segmento "Restaurante" es, con mucho, el más popular, seguido por "Super", "Farmacia" y "Otros". Los segmentos "Express" y "Drinks" tienen menos pedidos en comparación.
- "CASH" y "PIX" son los métodos de pago más populares. Otros métodos de pago, como "PIX_ON_DELIVERY" y "CREDIT_CARD", también son comunes, pero no tanto como los dos primeros.
- Es posible observar algunos picos de registro ocurridos en los meses de diciembre de 2020 y marzo de 2022, seguidos por un comportamiento particularmente uniforme en el resto de los meses con un leve descenso de registros hacia el último trimestre del año 2021. A priori es posible pensar que hay un descenso en la actividad económica propia de Brasil debido a festividades propias de las fechas.
- De manera complementaria se observa que 94% de las entregas son a tiempo mientras que el 6% de las entregas sufren algún tipo de retraso.

- Adicionalmente, es posible ver a través de un gráfico de barras que los días de mayor cantidad de órdenes son los martes y miércoles, y, al mismo tiempo, que los domingos son días en donde no se manejan envíos de pedidos.
- La mediana de la cantidad de órdenes parece ser más alta durante los días laborables en comparación con el fin de semana. La variabilidad (representada por la altura de las cajas) es similar a lo largo de los días, aunque hay algunas diferencias. Se observa una mayor variabilidad en la cantidad de órdenes los viernes y sábados en comparación con otros días.

6.3. Gestión de Etiquetado de Clientes

Con base en la información y necesidades proporcionadas por el negocio, se estableció a partir de los datos un Modelo RFM (Recency, Frequency, Monetary) para generar una primera segmentación de usuarios a partir de la recencia de la última compra, la frecuencia de compras y el monto o importe de las compras para cada usuario.

Haciendo en primer lugar un abordaje de las cifras a partir del Lifetime Value del cliente, se encontraron las siguientes inferencias:

- En promedio, se espera que un cliente aporte US$97,23ua lo largo de su tiempo como cliente:
- El valor promedio de GMV por compra es de aproximadamente US$116,92 .
- En promedio, un cliente realiza aproximadamente 27.67 compras
- La duración promedio de un cliente (tiempo de registro y última orden) es de aproximadamente 175, 36 días.

Con los datos de LTV claros, es posible generar una primera aproximación a la segmentación deseada:

Tipo de Cliente	Recencia	Frecuencia	Monto	Tamaño
VIP	Alto	Alto	Alto	2.048
Leales	Medio-Alto	Alto	Alto	1.989
En Riesgo	Bajo	Medio	Medio	2.126
Posible Churn	Muy bajo	Bajo	Bajo	2.124

Con esto, y con base en la información proporcionada por negocio es posible verificar que a pesar de que existe una población basada en un posible churn, el segmento de alto valor como los VIP y los leales son aquellos segmentos en los que vale la pena que el grupo de actores interesados genere esfuerzos grandes, mientras que pueden generar esfuerzos pequeños en los clientes que son significativamente inferiores en monto. Vale la pena resaltar que los clientes en riesgo y en posible churn son aproximadamente el 51 % de la muestra, lo cual es representativo a priori de cara a que en suma pudieran tener un impacto financiero alto. Con lo anterior, se busca responder a la pregunta No. 1 abordada anteriormente.

7. PRIMERAS CONCLUSIONES – PROXIMAS APROXIMACIONES

- La información proporcionada por Frubana contaba con robustez lo que hizo más facil el alistamiento, procesabilidad y demás virtudes de los datos para poder ejecutar un posterior modelo.
- El mecanismo de recolección de datos a partir del método de entrevista fue crucial para una estimación más estratégica del problema a solucionar y poder responder a preguntas que agreguen valor al negocio.
- Con la integración del modelo RFM es posible ya contestar uno de los dos interrogantes abordados en el documento de cara a la correcta categorización de usuarios.
- El análisis exploratorio de datos arrojó componentes de estacionariedad en las series de tiempo utilizadas, lo que permitirá mejores validaciones teóricas a partir de que se cumple un mayor número de propiedades.
- Es necesario generar un segundo punto de contacto con el negocio para socializar el proceso de etiquetado propuesto y ver si hay algún comentario que mejore o corrija el enfoque generado en este documento.

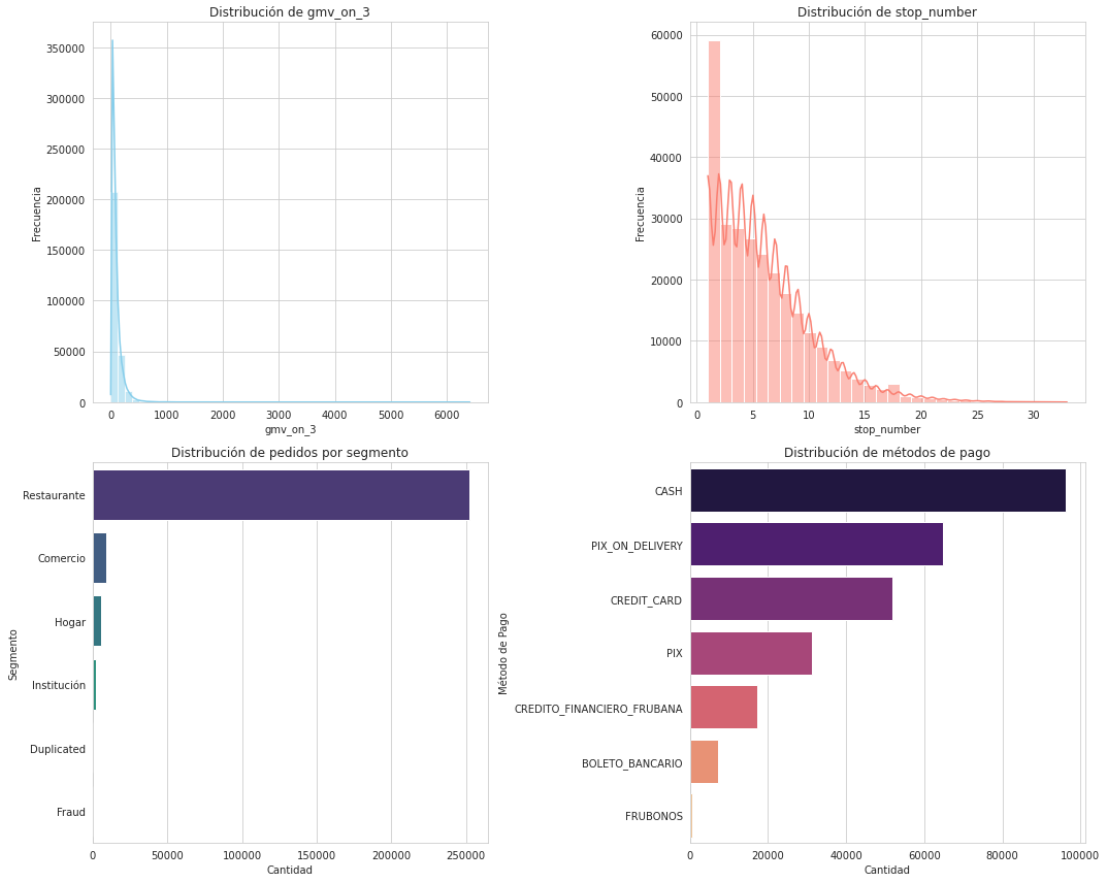
Anexo 1. Descripción de variables utilizadas en el modelo.

Variable	Descripción
source_id (integer)	Identificador del cliente
Birthday (date)	Fecha de registro del cliente
month_order (date)	Mes de la orden
year_order (date)	Año de la orden
City (Geo)	Ciudad en donde se realizó el pedido
Segmento (cat)	Tipología de clientes (Restaurante, Comercio, Hogar, Institución)
gmv_on_3 (float)	Gross Merchandising Volume
Active (boolean)	Estado activo de cliente (Activo o Inactivo)
is_graduated (boolean)	Nivel de antigüedad del cliente
is_kam (boolean)	Indica si el cliente tiene asignado un Key Account Manager
reactivation_status (cat)	Si el cliente ha estado en algún momento en estado reactivo
microzone_source_id (float)	Identificador de la microzona
deliver_date (date)	Fecha de entrega del pedido
trip_id (object)	Identificador del viaje
ontime_num (boolean)	Indica si el viaje llegó a tiempo o no
payment_metod_code (cat)	Método de Pago
dispatch_warehouse (cat)	Nombre de la bodega en dónde se despacha
stop_number (int)	Número de la parada a la que corresponde el pedido dentro del viaje

Anexo 2. Análisis Exploratorio de Datos Complementario

Visualmente, se generaron algunos gráficos para entender mejor el panorama de datos proporcionado.

Figura 1. Análisis de distribución de variables clave de Frubana



Fuente: Elaboración Propia.

Distribución de gmv_on_3:

- La mayoría de los valores se concentran en el rango inferior, con un pico pronunciado cerca de 0. Hay algunos valores extremos, pero son infrecuentes.
- Se debe ver la distribución quitando outliers

Distribución de stop_number:

- La mayoría de las paradas están en el rango de 1 a 10, con un pico en el número de paradas 1 y 2. El número de paradas tiende a disminuir a medida que aumenta el número de paradas.

Distribución de pedidos por segmento:

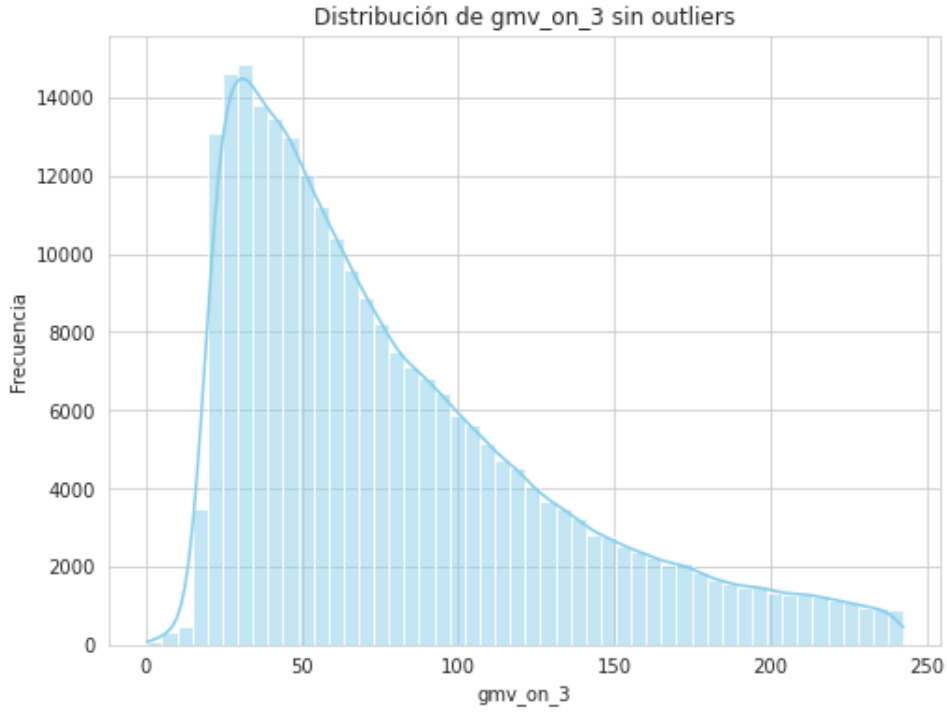
- El segmento "Restaurante" es, con mucho, el más popular, seguido por "Super", "Farmacia" y "Otros". Los segmentos "Express" y "Drinks" tienen menos pedidos en comparación.

Distribución de métodos de pago en payment_method_code:

- "CASH" y "PIX" son los métodos de pago más populares. Otros métodos de pago, como "PIX_ON_DELIVERY" y "CREDIT_CARD", también son comunes, pero no tanto como los dos primeros.

Para asegurar una correcta interpretación de los datos, se utilizó el método de rango intercuartílico para suprimir aquellos datos atípicos de la base y generar un comportamiento más uniforme:

Figura 2. Distribución del Gross Merchandising Volume ajustado sin datos atípicos

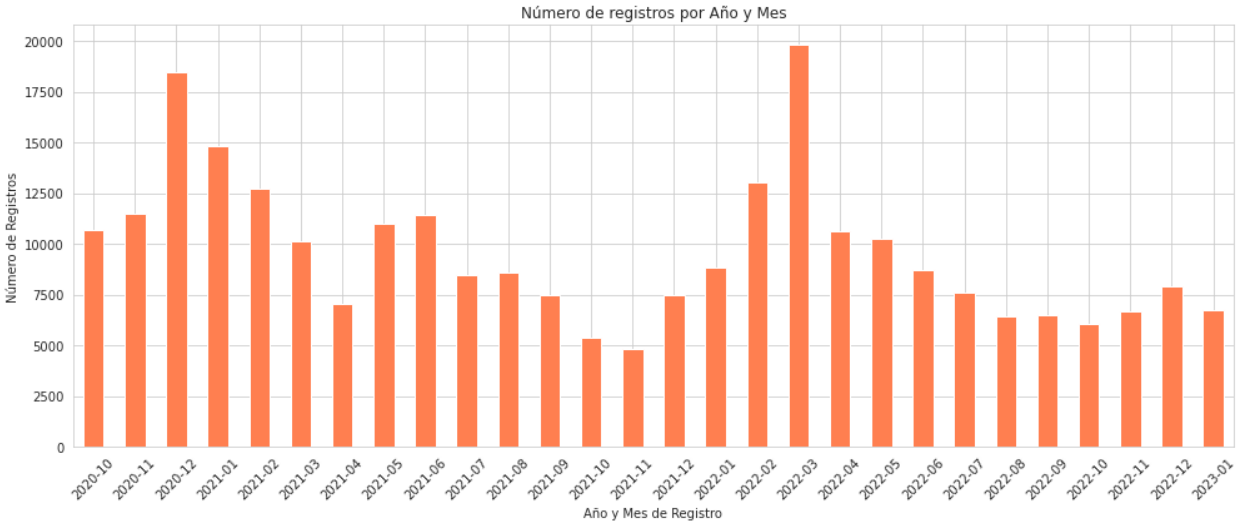


Fuente: Elaboración propia

La distribución de gmv_on_3 sin outliers muestra un comportamiento más claro y conciso. Aunque todavía está sesgada hacia la derecha (distribución asimétrica positiva), la eliminación de los valores extremos permite apreciar mejor la concentración de los datos en el rango inferior.

Adicionalmente, se generó un gráfico de barras para observar el número de registros por año y por mes.

Figura 3. Número de registros de clientes por año y mes



Fuente: Elaboración propia.

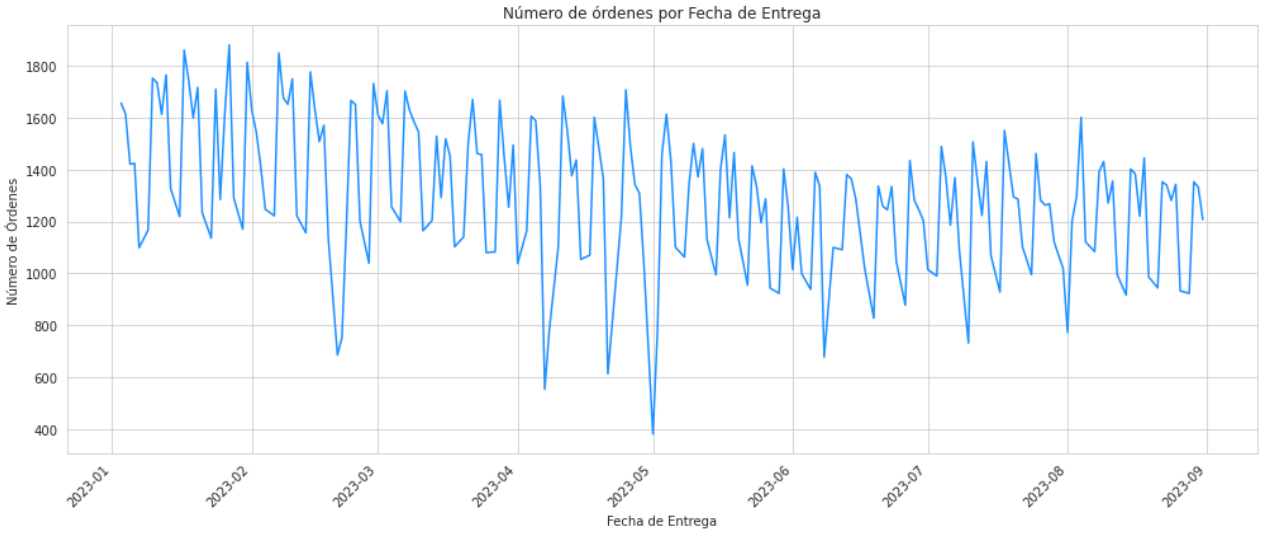
A partir del análisis visual, es posible observar algunos picos de registro ocurridos en los meses de diciembre de 2020 y marzo de 2022, seguidos por un comportamiento particularmente uniforme en el resto de los meses con un leve descenso de registros hacia el último trimestre del año 2021. A priori es posible pensar que hay un descenso en la actividad económica propia de Brasil debido a festividades propias de las fechas

Figura 4. Proporción de órdenes entregadas a tiempo vs con retraso

Fuente: Elaboración propia

De manera complementaria se observa que 94% de las entregas son a tiempo mientras que el 6% de las entregas sufren algún tipo de retraso.

Figura 5. Número de órdenes por fecha de entrega



Fuente: Elaboración propia

Cuando se observa el número de órdenes por fecha de entrega es posible ver algunos rasgos de estacionariedad de la serie de tiempo, dada por un comportamiento uniforme con esbozos de varianza constante a lo largo del tiempo.

Para analizar de una manera más robusta esta serie temporal, se procede a descomponerla para analizar la tendencia y la estacionalidad, así como el componente residual.

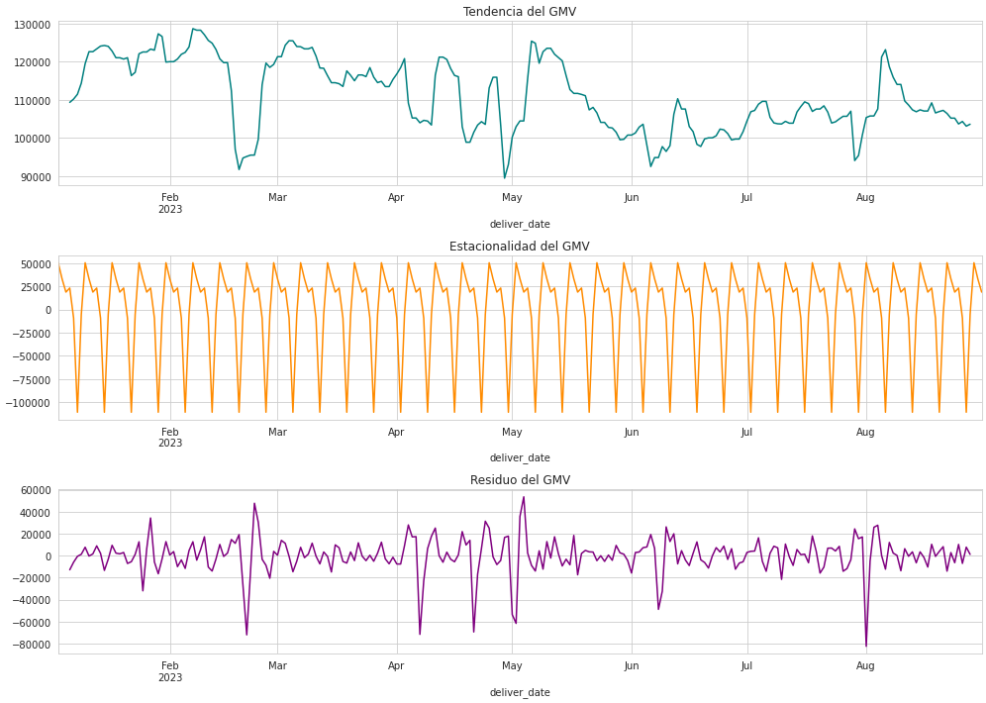
Las visualizaciones muestran la descomposición de la serie temporal del GMV en sus componentes:

Tendencia: Muestra la tendencia subyacente en los datos. Se observa una tendencia creciente en el GMV a lo largo del tiempo, aunque con algunas fluctuaciones.

Estacionalidad: Revela patrones repetitivos en la serie temporal a intervalos regulares. En este caso, no se observa una estacionalidad claramente definida en el GMV.

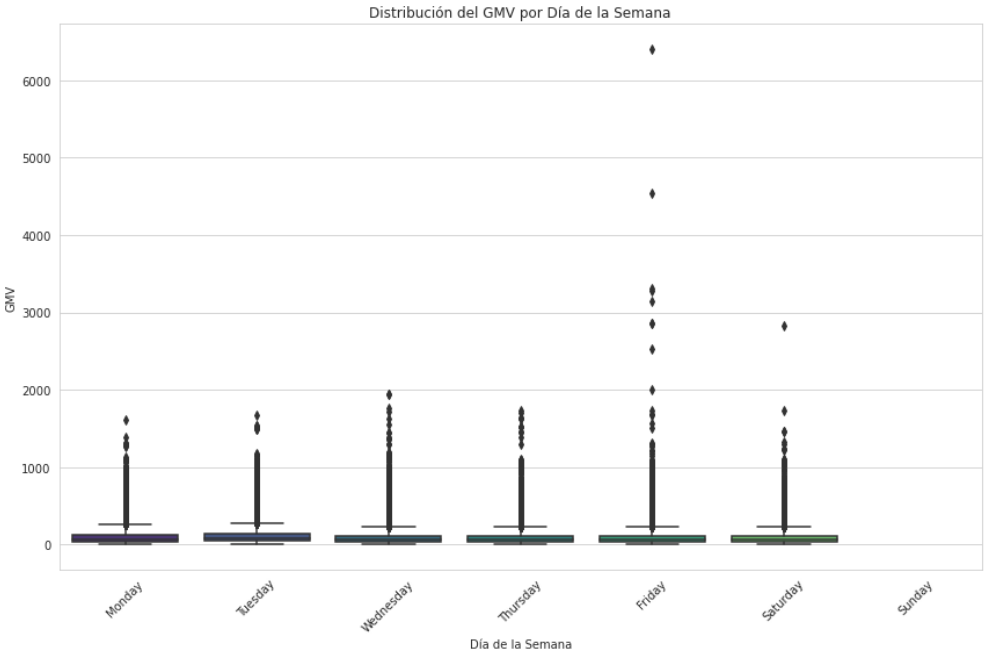
Residuo: Es la diferencia entre la serie temporal original y la reconstruida a partir de la tendencia y la estacionalidad. Los residuos parecen centrarse alrededor de cero, aunque con algunas variaciones.

Figura 6. Descomposición de la serie de número de órdenes por fecha en Tendencia, Estacionalidad y Residuos



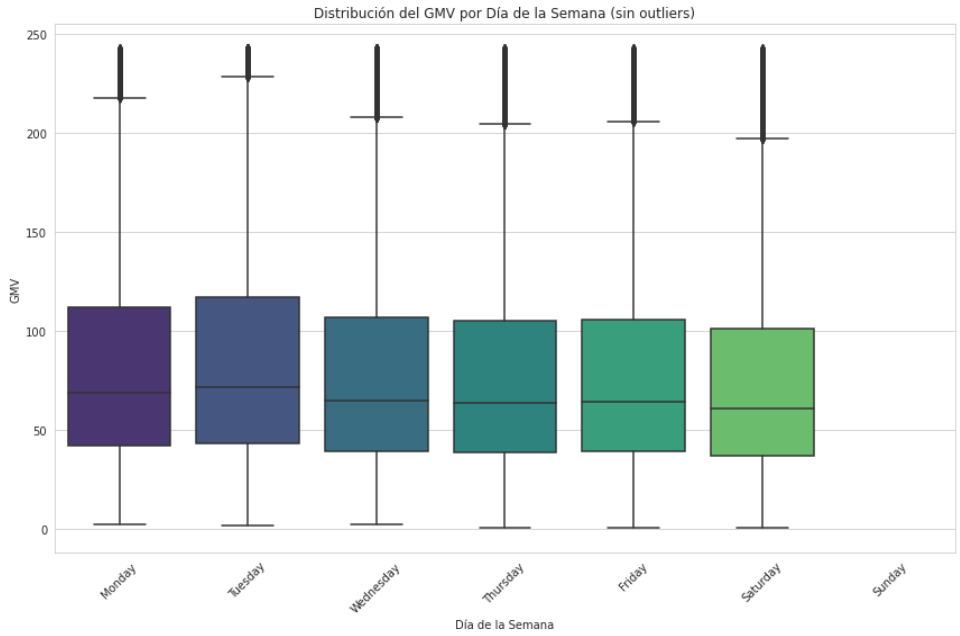
Fuente: Elaboración propia

Figura 7. Distribución del GMV por Día de la semana



Fuente: Elaboración propia

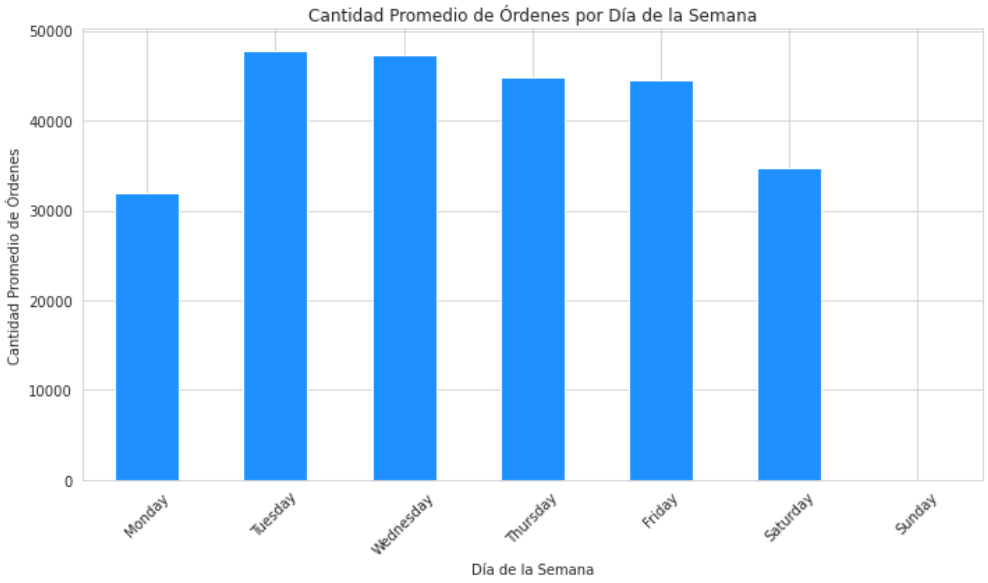
Figura 7. Distribución del GMV por Día de la semana sin Outliers



Fuente: Elaboración propia

Para verificar si hay una diferencia significativa en las medias del GMV entre los días de la semana, utilizaremos el análisis de varianza (ANOVA).

Figura 8. Cantidad promedio de órdenes por día de la semana



Fuente: Elaboración Propia

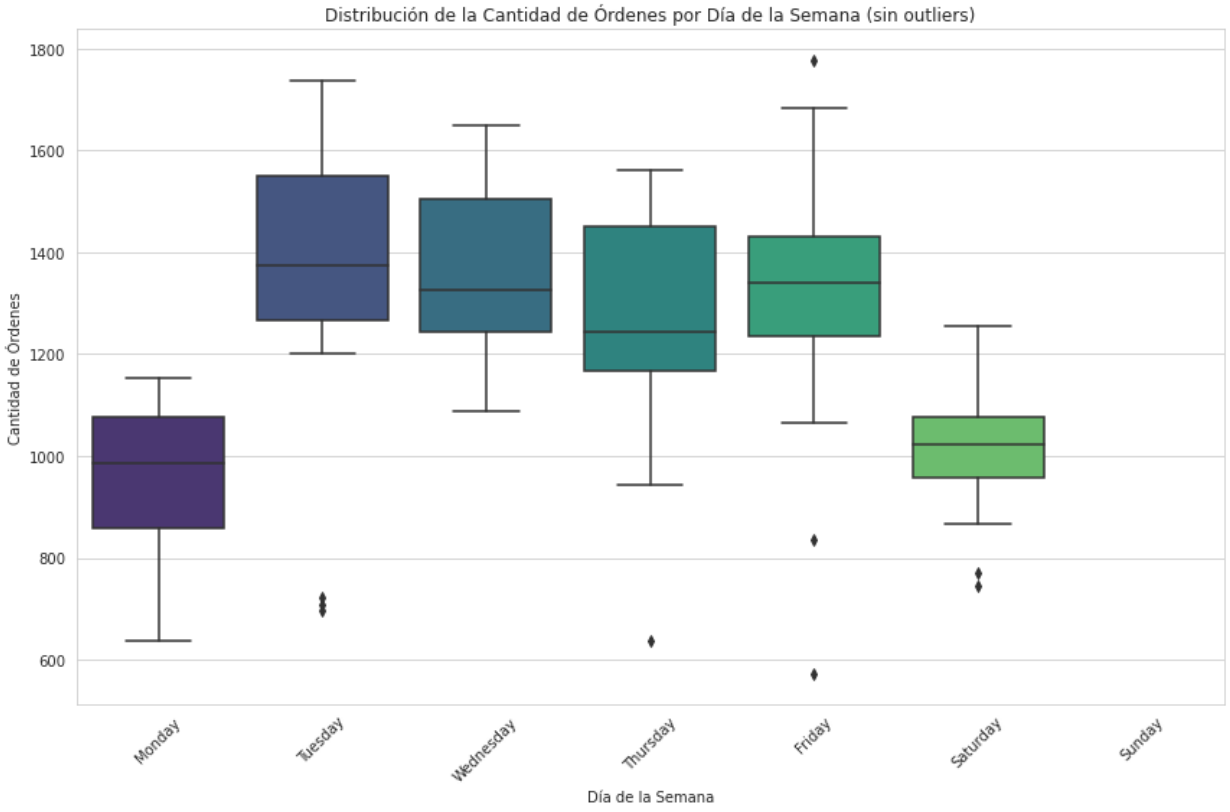
El resultado del ANOVA es el siguiente:

Valor-F:	214.07	214.07
Valor-p: 1.08 × 1 0 − 228 1.08×10 −228		

Dado que el valor-p es extremadamente pequeño (mucho menor que el nivel de significancia típico de 0.05), rechazamos la hipótesis nula (H0). Esto sugiere que hay una diferencia significativa en las medias del GMV entre al menos dos días de la semana.

Adicionalmente, es posible ver a través de un gráfico de barras que los días de mayor cantidad de órdenes son los martes y miércoles, y, al mismo tiempo, que los domingos son días en donde no se manejan envíos de pedidos.

Figura 9. Cantidad promedio de órdenes por día de la semana



Fuente: Elaboración propia

El boxplot muestra la distribución de la cantidad de órdenes por día de la semana después de eliminar los outliers. Aquí hay algunas observaciones:

La mediana de la cantidad de órdenes parece ser más alta durante los días laborables en comparación con el fin de semana. La variabilidad (representada por la altura de las cajas) es similar a lo largo de los días, aunque hay algunas diferencias. Se observa una mayor variabilidad en la cantidad de órdenes los viernes y sábados en comparación con otros días.