# 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]

Desde la atención a restaurantes y agricultores, se busca generar un ambiente mejor remunerado para cada agente que interviene en el proceso, incluyendo la tecnología escalada que permita tomar decisiones con un mayor volumen de datos y abaratar el 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 solo se usaría sobre los clientes con valor de compra (ticket promedio) significativo, considerando 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.

[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

## 2. IDEACIÓN DEL PROBLEMA DE NEGOCIO

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, el modelo se descartó porque se generó inconvenientes basados en una incorrecta clasificación de los usuarios y de una ventana de tiempo muy alta según el 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 mediante etiquetas de cara a su comportamiento de compra, para encontrar 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.

[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.

## 3. IMPLICACIONES DEL ABORDAJE DEL PROBLEMA DE NEGOCIO

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 micro zonas 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.

Además, de poco más de 645 mil observaciones, se encontraron 269.133 registros válidos, con lo que 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

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

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.

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, se puede ver a través de un gráfico de barras que los días de mayor cantidad de órdenes son martes y miércoles, y que los domingos son días en los que 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 fácil 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.

## 8. PREPARACIÓN DE DATOS

Tras el análisis exploratorio de datos como paso previo a la estructuración del modelo, se generó el modelo de RFM mencionado para generar un esquema de información orientado a la asociación de categorías de clientes según su monto transaccional.

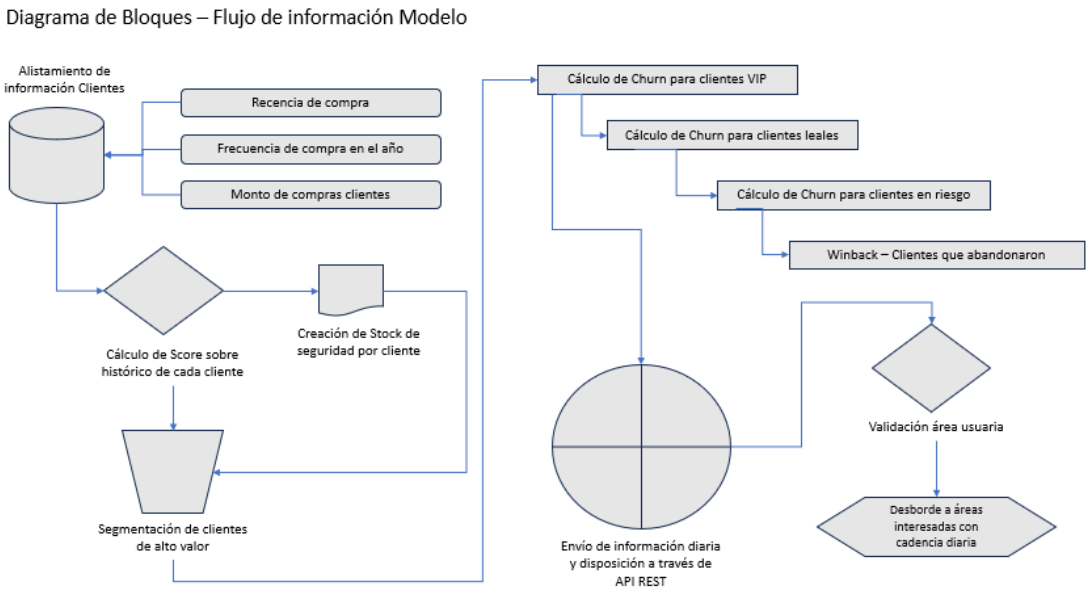



**Figura 1. Diagrama de bloques para flujo de información para el modelo**

Con base en los datos del horizonte temporal proporcionado, se aplicó la metodología observada en la figura 1, en donde, a partir de la garantía de consistencia, coherencia, etc. de los datos, se usó el flujo de información reflejado en la figura 1.

La idea general del flujo es que un proceso automatizado alimente los repositorios del cliente en donde reside la información y que incluya los componentes de recencia, frecuencia y monto monetario de los clientes. Una vez se alimentan sobre una base diaria, se genera un cálculo del score de Churn sobre la información propia de cada usuario, de tal manera que para cada cliente se obtiene su Churn personalizado. Una vez se genera la calificación, se integra a un stock de seguridad y se calcula la clasificación de churneado o no.

De esta manera, el modelo de Churn agrega valor en la medida en que permite clasificar de manera previa los clientes críticos y minimizar la pérdida esperada de ingresos para la compañía, haciendo que el flujo de información esté priorizado en función de la misma métrica.

## 9. ESTRATEGIA DE VALIDACIÓN Y SELECCIÓN DEL MODELO

Como se mencionó en las secciones anteriores, se utilizarán *dos modelos de clasificación* para poder predecir a través de tres modelos de clasificación diferentes: la regresión logística, el Bosque Aleatorio y el modelo SVM.

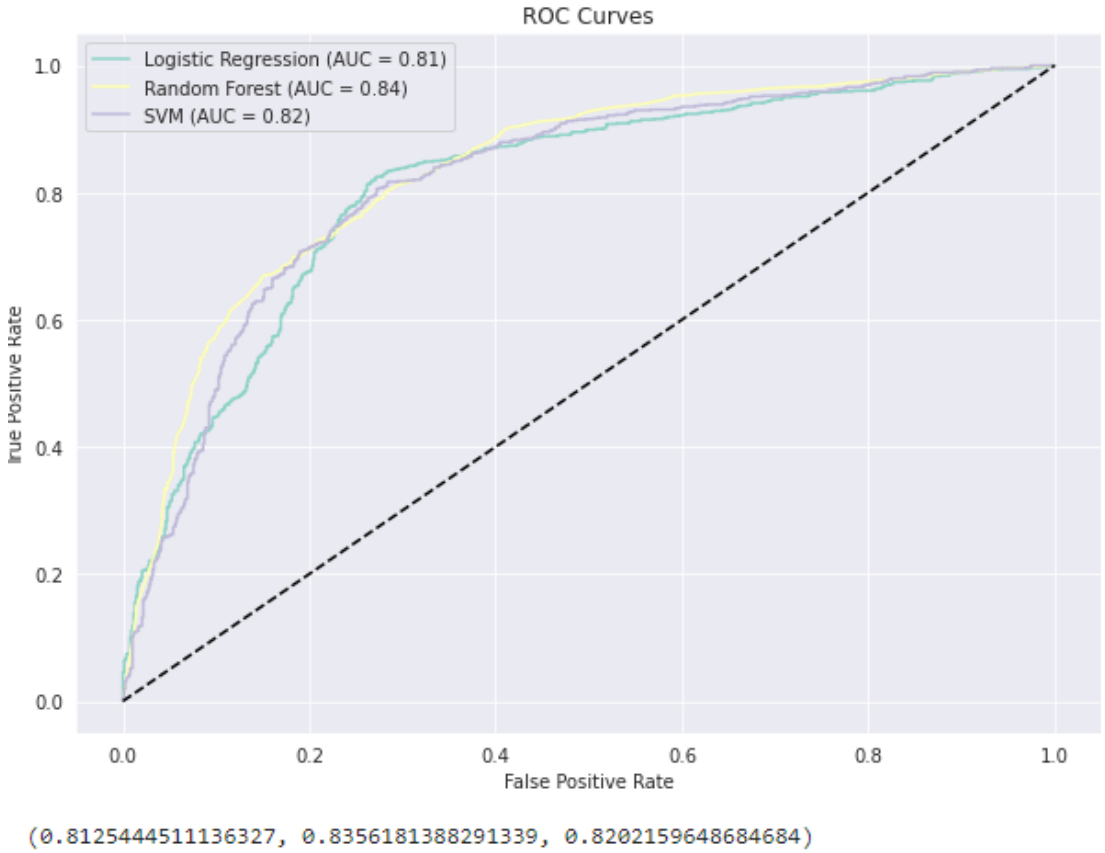



**Figura 2. Curvas ROC para los modelos de datos escogidos.**

Como se observa, los algoritmos de clasificación en sus curvas ROC muestran que hay una correcta tarea se clasificación, con el Random Forest ligeramente por delante del modelo de Regresión Logística. El valor del área bajo la curva es de 0,81 para la Regresión Logística y de 0,84 para el Bosque Aleatorio. Es de conocimiento común que el valor del área bajo la curva ROC superior a 0,5 corresponde a un 81 u 84 por ciento de probabilidad de que el modelo clasifique un caso positivo aleatorio más alto que un caso negativo, esto es, que en más del 80% de los casos.

## 9.1. Análisis del modelo de Regresión Logístico

A continuación, se observa la matriz de confusión asociada al modelo.

### 9.1.1. Matriz de Confusión

**-Verdaderos Negativos (TN): 563**

El modelo predijo correctamente la clase negativa 563 veces.

**-Falsos Positivos (FP): 216**

El modelo predijo incorrectamente la clase positiva cuando en realidad era negativa 216 veces.

**-Falsos Negativos (FN):** 153

El modelo predijo incorrectamente la clase negativa cuando en realidad era positiva 153 veces.

**Verdaderos Positivos (TP): 726**

El modelo predijo correctamente la clase positiva 726 veces.

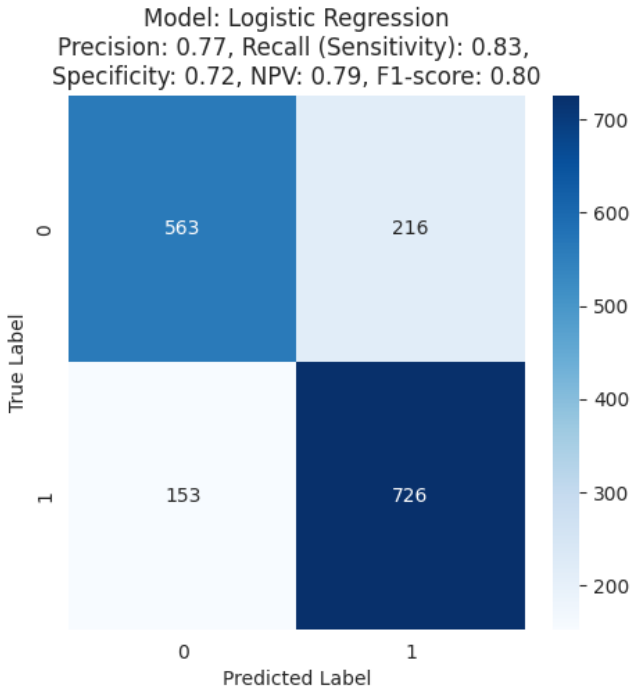



**Figura 3. Matriz de confusión para modelo de regresión logística.**

### 9.1.2. Métricas de rendimiento

**Precisión: 0.77:** De todas las predicciones positivas que hizo el modelo (TP+FP) el 77% eran correctas.

**Recall (Sensibilidad): 0.83**: Del total de casos positivos reales (TP+FN), el modelo pudo identificar correctamente el 83%.

**Especificidad: 0.72**: Del total de casos negativos reales (TN+FP), el modelo identificó correctamente el 72% como negativos.

**Valor Predictivo Negativo (NPV): 0.79:** De todas las predicciones negativas que hizo el modelo (TN+FN), el 79% eran correctas.

**F1-score: 0.80**: Esta es la media armónica de la precisión y la sensibilidad, proporcionando un balance entre estas dos métricas. Un F1-score de 0.80 es relativamente alto, indicando un buen balance entre precisión y sensibilidad.

En síntesis, el modelo de regresión logístico hizo una clasificación de los churneados buena y con porcentajes por encima de 0,5 que indican una bondad de ajuste del modelo adecuado para la situación particular. Vale la pena resaltar que, a pesar de esto, existen análisis y variables complementarios para poder mejorar en un margen más grande la capacidad de clasificación del modelo.

## 9.2. Análisis del modelo de Random Forest

En la siguiente figura se observa la matriz de confusión del correspondiente modelo

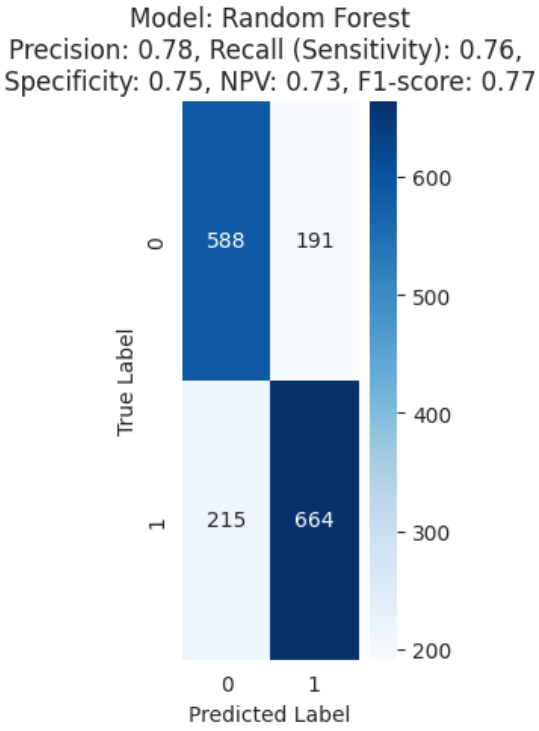



**Figura 4. Matriz de confusión para modelo de Random Forest.**

### 9.2.1. Matriz de Confusión

**Verdaderos Negativos (TN):** 588 - El modelo predijo correctamente que 588 instancias no pertenecían a la clase positiva.

**Falsos Positivos (FP): 191** - El modelo incorrectamente predijo que 191 instancias pertenecían a la clase positiva.

**Falsos Negativos (FN): 215** - El modelo incorrectamente predijo que 215 instancias pertenecían a la clase negativa cuando en realidad eran positivas.

**Verdaderos Positivos (TP): 664** - El modelo predijo correctamente que 664 instancias pertenecían a la clase positiva.

### 9.2.2. Métricas de rendimiento

**Precisión: 0.78** - De todas las instancias que el modelo predijo como positivas, el 78% de esas predicciones fueron correctas.

**Recall (Sensibilidad): 0.76** - De todas las instancias que son realmente positivas, el modelo fue capaz de identificar correctamente el 76% de ellas.

**Especificidad: 0.75** - De todas las instancias que son realmente negativas, el modelo identificó correctamente el 75% de ellas como negativas.

**Valor Predictivo Negativo (NPV): 0.73** - De todas las instancias que el modelo predijo como negativas, el 73% de esas predicciones fueron correctas.

**F1-score: 0.77** – El modelo sugiere un equilibrio razonable entre la precisión y la sensibilidad, indicando que el modelo es relativamente robusto.

El modelo Random Forest tiene un desempeño equilibrado con respecto a la precisión y el recall, con una ligera disminución en la sensibilidad en comparación con la regresión logística (basándonos en la matriz de confusión previa). La precisión es ligeramente mayor, pero esto se compensa con un menor recall. La especificidad y el NPV son comparables a la regresión logística, lo que indica que el modelo es relativamente bueno para identificar tanto los verdaderos positivos como los verdaderos negativos. La puntuación F1-score ligeramente más baja que en la regresión logística sugiere que hay un balance un poco óptimo entre precisión y recall.

## 9.3. Análisis del modelo de Support Vector Machine

A continuación, se describe la matriz de confusión y las métricas de rendimiento para el modelo de Support Vector Machine

### 9.3.1. Matriz de Confusión

**Verdaderos Negativos (TN):** 541 - El modelo SVM predijo correctamente que 541 instancias no pertenecían a la clase de interés.

**Falsos Positivos (FP):** 238 - El modelo predijo incorrectamente que 238 instancias pertenecían a la clase de interés.

**Falsos Negativos (FN):** 159 - El modelo predijo incorrectamente que 159 instancias no pertenecían a la clase de interés cuando en realidad sí lo hacían.

**Verdaderos Positivos (TP):** 720 - El modelo predijo correctamente que 720 instancias pertenecían a la clase de interés.

### 9.3.2. Métricas de Rendimiento

**Precisión:** 0.75 - De todas las instancias clasificadas por el modelo como positivas, el 75% fueron clasificaciones correctas.

**Recall (Sensibilidad):** 0.82 - El modelo fue capaz de identificar el 82% de todas las instancias que son realmente positivas.

**Especificidad:** 0.69 - El modelo identificó correctamente el 69% de todas las instancias que son realmente negativas.

**Valor Predictivo Negativo (NPV):** 0.77 - De todas las instancias clasificadas por el modelo como negativas, el 77% fueron clasificaciones correctas.

**F1-score:** 0.78.

El modelo SVM demuestra ser bastante efectivo, con un recall alto (82%), lo que indica una fuerte habilidad para detectar la clase positiva. Sin embargo, la precisión y la especificidad son algo menores en comparación con la regresión logística y el Random Forest, lo que significa que hay más falsos positivos relativos al total de casos negativos predichos. A pesar de esto, el F1-score es competitivo, indicando que el equilibrio general entre la precisión y el recall sigue siendo bueno.

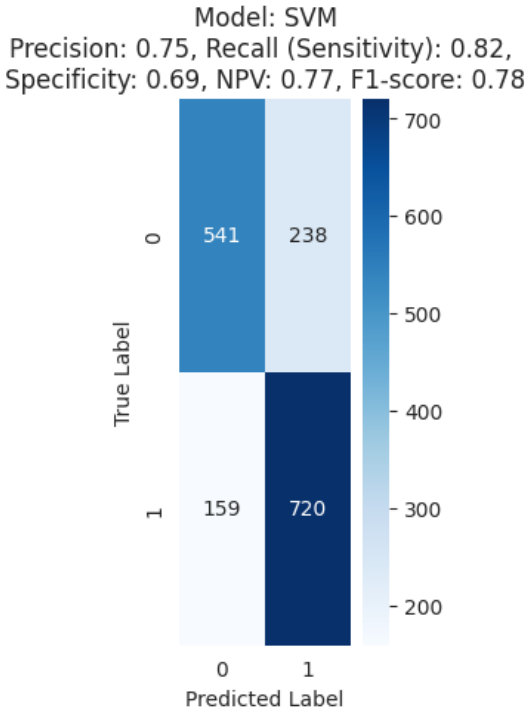



**Figura 5. Matriz de confusión para modelo de Support Vector Machines.**

## 10. CONCLUSIONES

Como alternativa complementaria al análisis inicial se utilizó el modelo de Support Vector Machine para observar cuál era su capacidad explicativa. Se obtiene que el modelo de *Random Forest* es el más apropiado para los datos en relación a que tiene un valor de área bajo la curva más alto que los demás, razón por la cual se optaría la publicación de los resultados del API REST bajo esta acepción.

- ***¿Cuáles son las mayores dificultades que se han tenido en el proyecto?***

Hasta el momento no ha habido un mayor volumen de dificultades diferente a la correcta e inicial metodología de comparación de clientes, lo que, a partir de reuniones con el cliente, ha dado lugar a varios escenarios de socialización y consenso, así como la escogencia del *stock de seguridad* que nace de la escogencia arbitraria de un valor en diálogo con el negocio.

- ***¿Qué estrategias se plantean para mitigarlas?***

Se establecieron 4 puntos de control con el negocio para definir la capacidad de ajuste del modelo en las variables imputadas por el usuario final, tales como el *stock de seguridad* y la temporalidad del Churn para los clientes. A partir del movimiento de tales variables se buscó ser lo más objetivo con el negocio de ahí la identificación individual de Churn guardando proporciones con valores por debajo de la curva del modelo superiores a 0,8 y omitiendo resultados vistos en algunas simulaciones realizadas por el equipo de 0,99.

Se debe hacer la claridad de que podría ser aconsejable verificar la calidad y diversidad del conjunto de datos, así como realizar validación cruzada para asegurarse de que los modelos no estén sobreajustados o que el conjunto de datos no sea demasiado fácil.

- ***¿Qué condiciones considera debería tener los datos para obtener mejores resultados?***

La ventana de análisis descrita en el documento corresponde a datos de un año obtenidos del negocio. Como siempre, si el negocio incorpora un conjunto de variables más robusto, los resultados pueden verse mejor representados en los modelos. De manera conjunta, un análisis posterior de canasta sobre las características de los usuarios puede ser una opción adicional que con una mayor robustez de información sea posible incorporar, así como una ventana de tiempo más alta en los datos ya entregados.

- ***¿El mejor modelo obtenido hasta el momento es suficiente para dar solución al problema u oportunidad de negocio abordado?***

La evidencia en los análisis numéricos indica un modelo de ajuste aceptable sobre los datos obtenidos. La metodología o la escogencia de los modelos puede ser suficiente en la medida en que se haga un trabajo paralelo con una ventana de datos de mayor magnitud que evalúe si los modelos de clasificación se comportan igual. En la literatura, sin embargo, es posible

encontrar amplios usos del modelo de Random Forest como típico en el abordaje del modelo de Churn, razón por la cual es evidente su utilidad y dependencia de un set de datos integral.

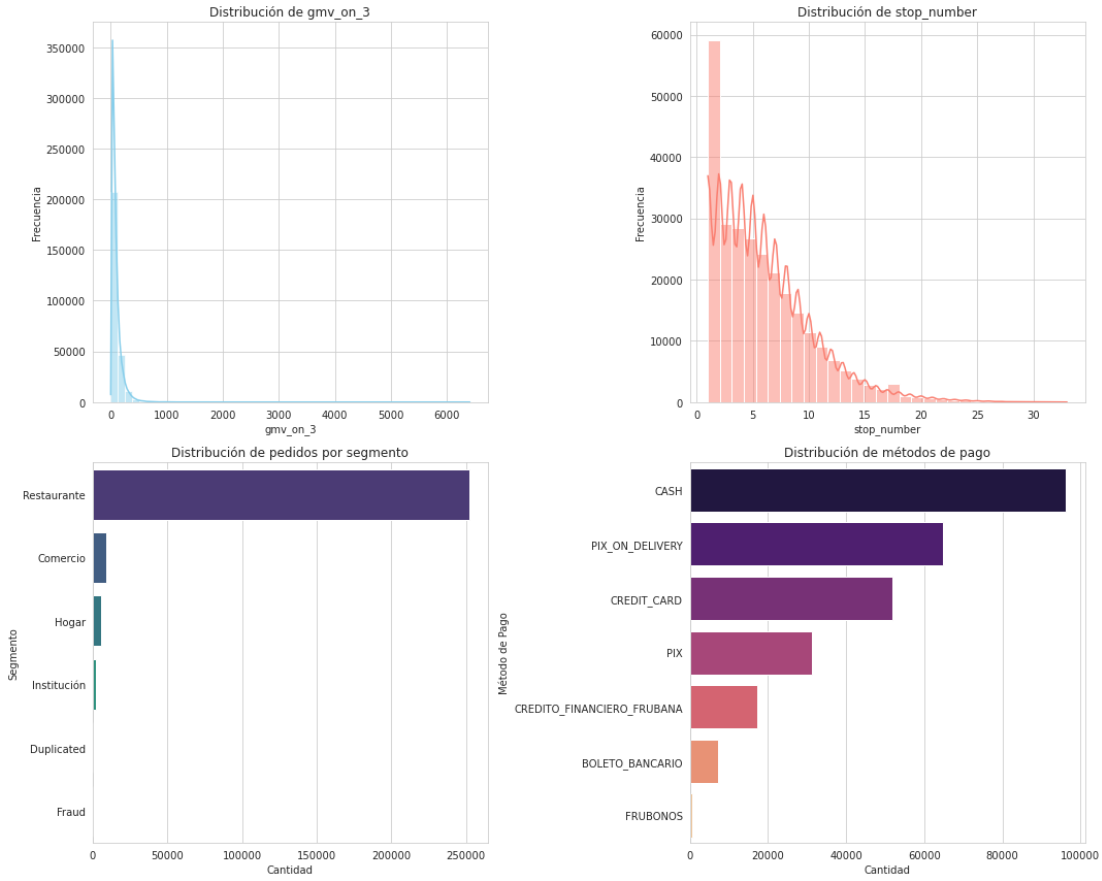
## Anexo 1. Descripción de variables utilizadas en el modelo.

| 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 micro zona |
| 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 6. 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:

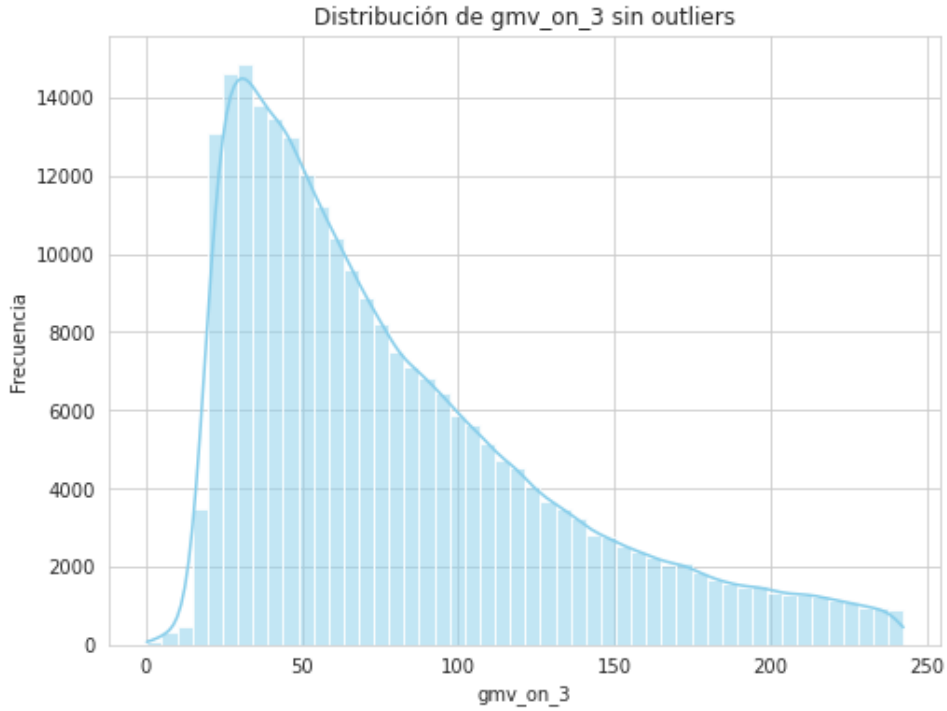
- El segmento "Restaurante" es, con mucho, el más popular, seguido por "Super", "Farmacia" y "Otros". Los segmentos "Express" y "Drinks" tienen menos pedidos en comparación.

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 7. Distribución del Gross Merchandising Volumen ajustado sin datos atípicos**

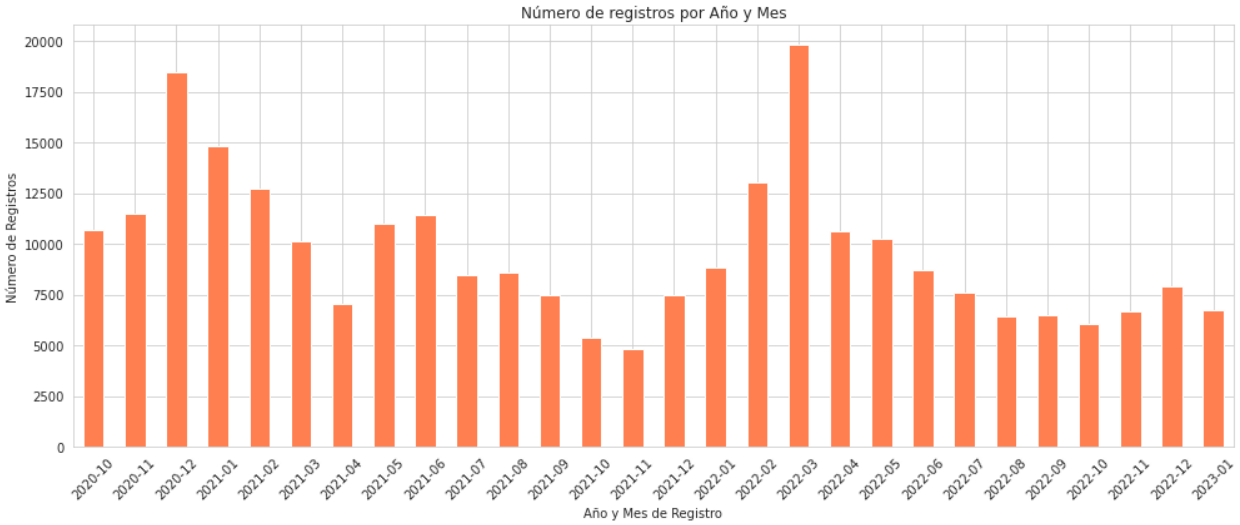



**Fuente: Elaboración propia**

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.

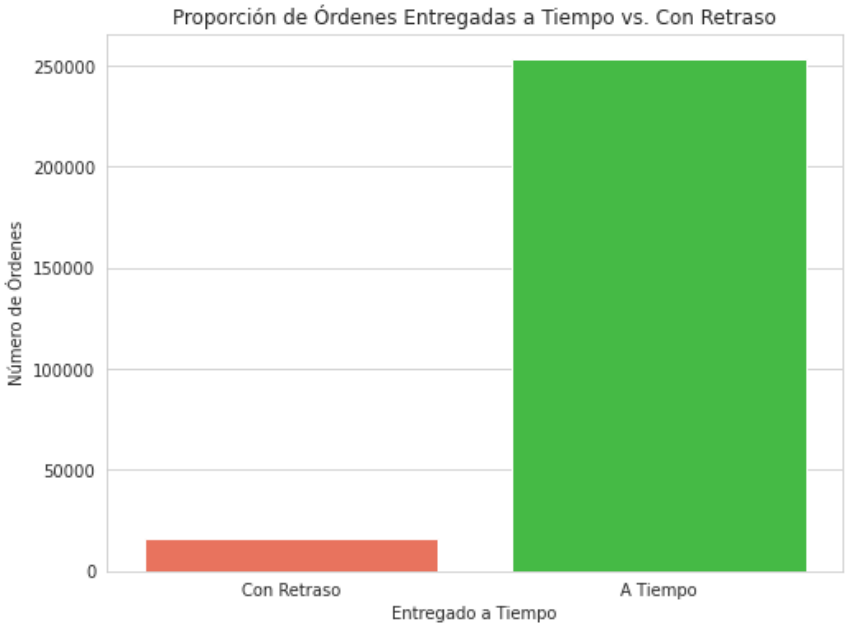
**Figura 8. Número de registros de clientes por año y mes**




**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

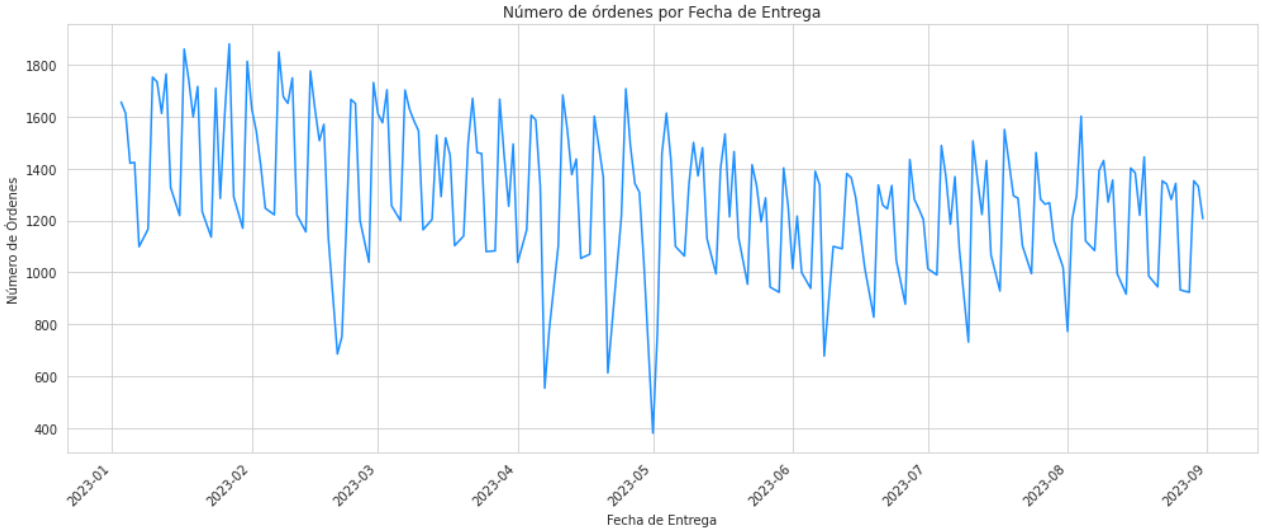
**Figura 9. Proporción de órdenes entregadas a tiempo vs con retraso**




**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 10. 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.

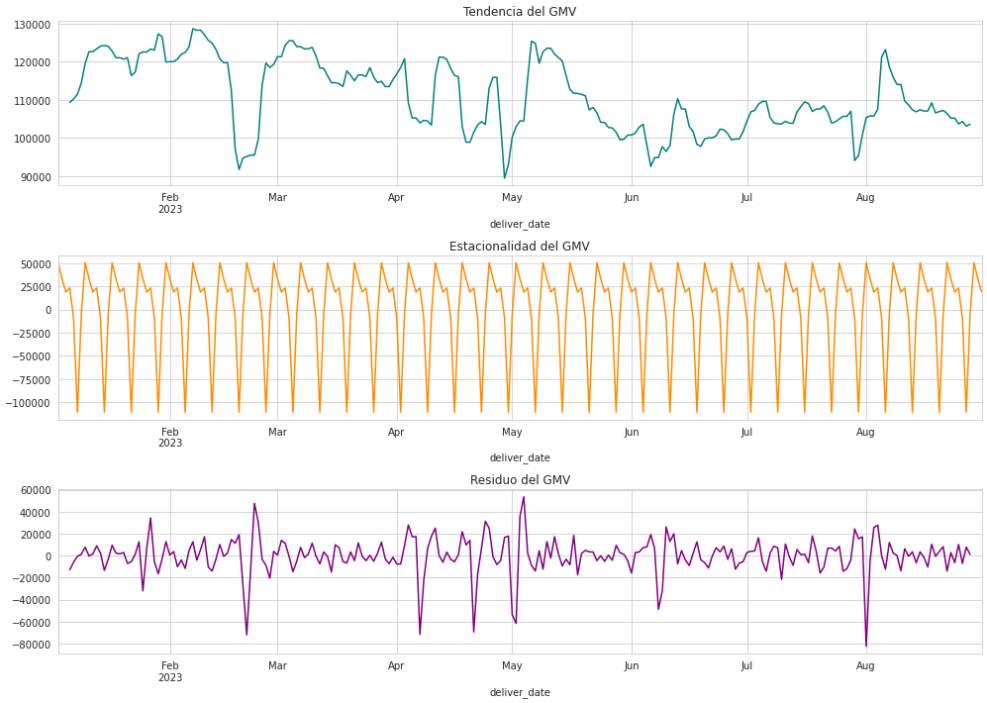
Las visualizaciones muestran la descomposición de la serie temporal del GMV en sus componentes:

**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.

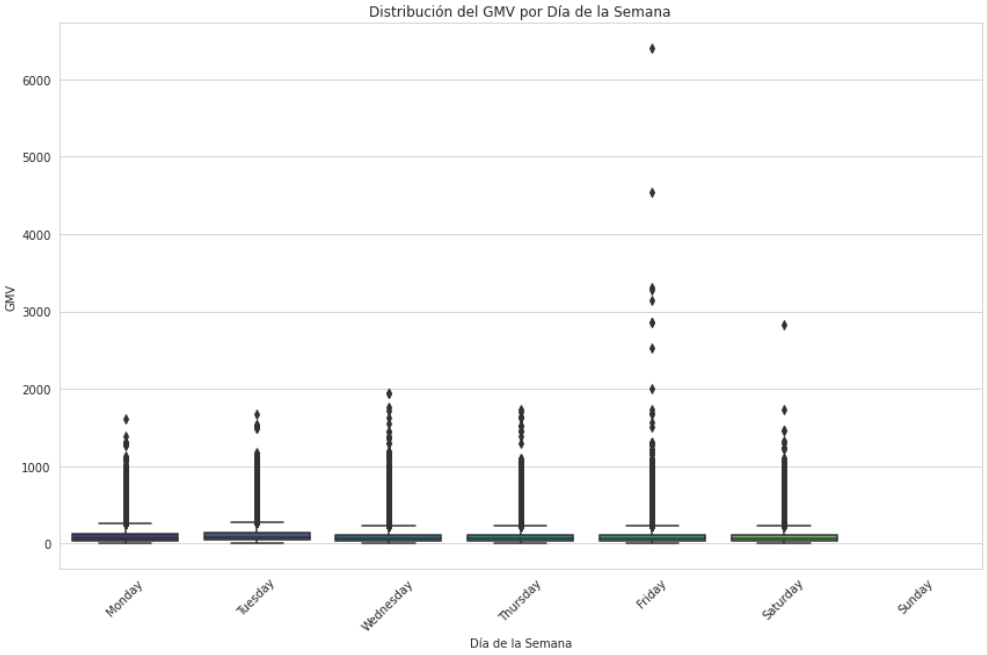
**Residuo:** Es la diferencia entre la serie temporal original y la reconstruida a partir de la tendencia y la estacionalidad. Los residuos parecen centrarse alrededor de cero, aunque con algunas variaciones.

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



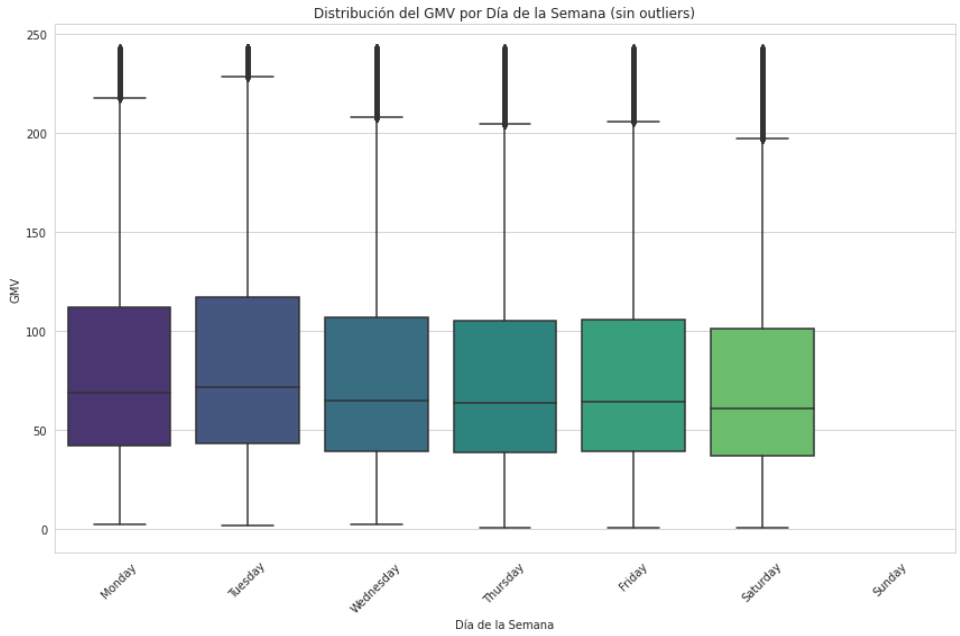

**Fuente: Elaboración propia**

**Figura 11. Distribución del GMV por Día de la semana**




**Fuente: Elaboración propia**

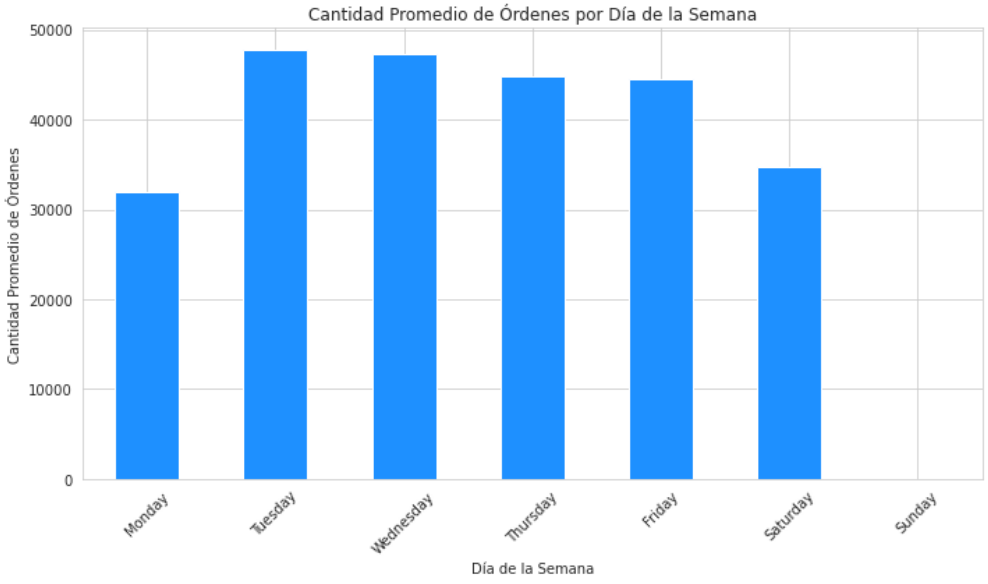
**Figura 12. Distribución del GMV por Día de la semana sin Outliers**




**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 13. Cantidad promedio de órdenes por día de la semana**




**Fuente: Elaboración Propia**

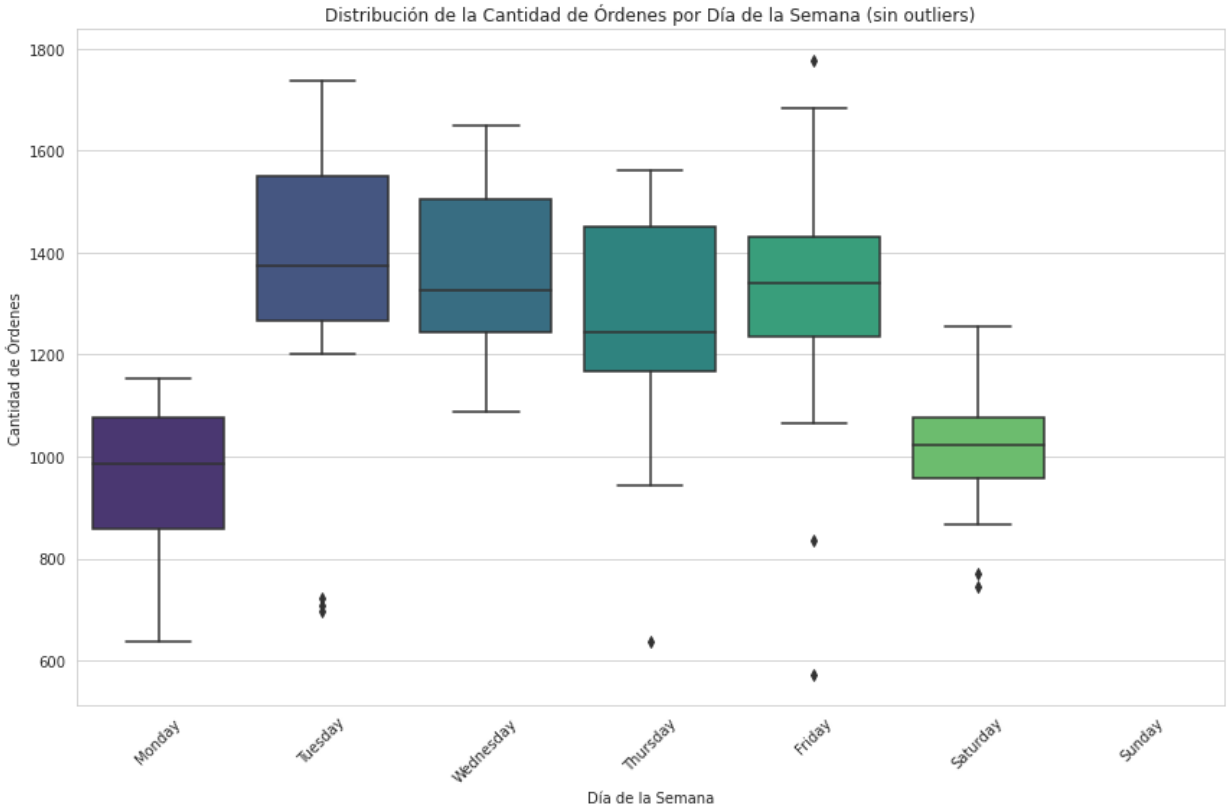
El resultado del ANOVA es el siguiente:

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 14. 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.