Olist

Proyecto Final
Grupo 3

Cabrera R. - Fogliatti V. - Martínez M. - Ortiz M.

Índice

Introducción....4
Equipo de trabajo....5
Situación actual de Olist....6
Objetivo general....10
Objetivos específicos....11
Alcance....13
Solución propuesta....14
Stack tecnológico....16
Stack tecnológico – Diario....17
Stack tecnológico - Principal....18
Stack tecnológico – Plus....19
Metodología de trabajo....20
Diagrama de Gantt....22
Creación de la API....23
Flujo de trabajo....24
Diagrama Entidad-Relación....26
Diccionario de datos....27

Informes de análisis.....30
Objetivos y KPI's.....32
Análisis.....38
Sugerencias.....51
Machine Learning.....56
Predicción.....62
Conclusión.....64

Introducción

Equipo de trabajo – Roles y responsabilidades

- Matías Martínez :Data Engineer
- Maciela Ortiz: Data Engineer
- Ronal Cabrera : Data Scientist
- Valentín Fogliatti : Data Analyst

Situación actual

Para el siguiente proyecto hemos sido contratados como consultores externos por la empresa Olist, con la tarea de analizar y buscar alternativas a una problemática concreta: la compañía quiere que sus usuarios incrementen sus ventas y alcancen una mayor cantidad de clientes.

Olist es una empresa de origen brasilero, fundada en 2015, y opera en el segmento del comercio electrónico, aunque no es e-commerce en sí. Es una gran tienda por departamentos dentro de los mercados, formada por miles de otras tiendas repartidas por todo Brasil. Esta gran tienda ya está conectada a los principales e-commerces de Brasil y ofrece estos espacios privilegiados para que los participantes puedan anunciar sus productos allí, ya sea que estos participantes cuenten con presencia en línea o no.

Continuando con este acercamiento a nuestro cliente, podemos decir que se trata de una empresa de tecnología SaaS (Software as a Service) que ofrece una solución que se compone de tres frentes:

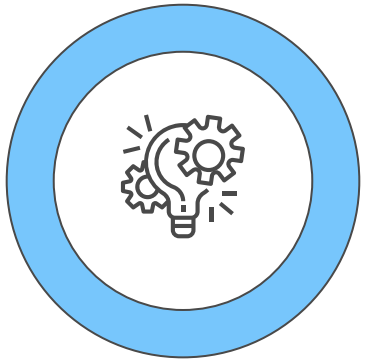
- **Software:** ofrece herramientas para potenciar la competitividad de productos, reportes de desempeño, sistema de gestión ERP, herramientas promocionales, herramientas que facilitan la carga y actualización de productos y precios, de seguimiento de envíos, entre otras tantas.
- **Contratos con los principales Marketplace:** presencia en 15 mercados diferentes, con contratos que les permiten a los vendedores ciertos beneficios a la hora de promocionar sus productos, como ser un mejor posicionamiento en los buscadores y recomendaciones.
- **Compartición de reputación:** los sellers comparten la reputación lograda en las miles de ventas que se realizan diariamente y, como resultado, ocupan las mejores posiciones, muchas de ellas dentro de las 'buy boxes'.

Cabe destacar que los vendedores tienen ciertas restricciones a la hora de vender sus productos, como que deben poder emitir factura electrónica, que solo aceptan tiendas que trabajen con categorías reguladas por canales de venta asociados, que deben poseer una cuenta bancaria vinculada y que los productos deben tener código de barras, entre otras.

También los productos ofrecidos por los sellers tienen ciertas limitantes: no se pueden comercializar productos usados, ni servicios, ni productos pertenecientes a algunas categorías específicas.

Olist a su vez posee 4 diferentes planes de pago que varían de acuerdo al tamaño de operaciones de cada vendedor y que ofrecen diferentes beneficios. Estos planes tienen un costo fijo mensual, semestral o anual, y descuentos de acuerdo al caso. Adicionalmente se paga un costo variable por comisiones por ventas que también va de acuerdo al plan elegido.

Por último, la tarea que se nos encomienda viene acompañada de 11 datasets con datos variados referidos a la empresa, comprendidos entre los períodos de 2016 a 2018, y disponemos de cuatro semanas para darle forma al producto final.



Objetivo general

Realizar un producto mínimo viable que permita a los usuarios (sellers) vender más productos a un mayor número de clientes.

Objetivos específicos

01

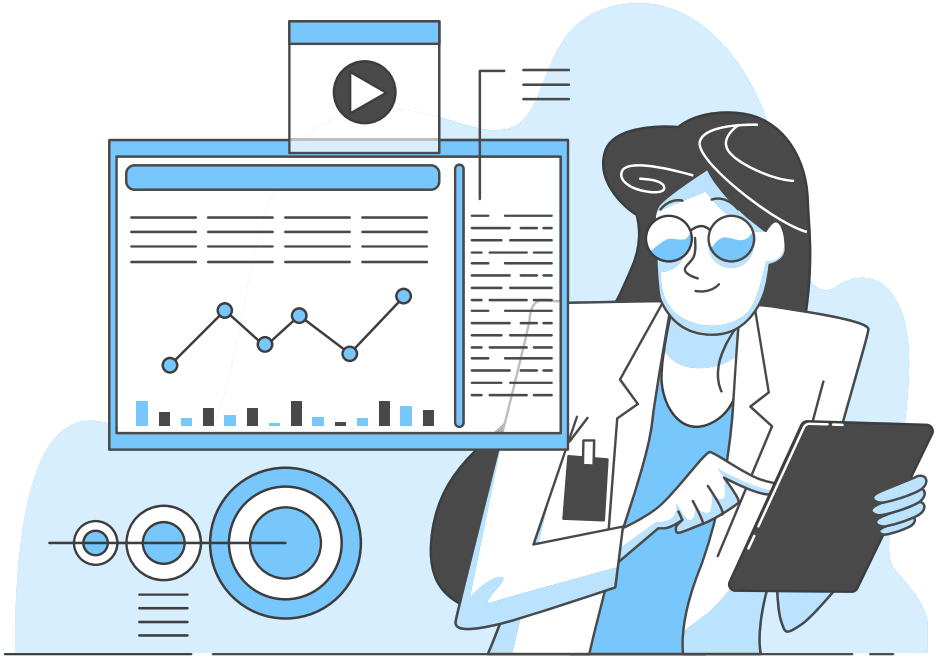
Realizar un análisis exploratorio de los datos y elaborar un reporte de calidad de los datos, acompañado de un diccionario de datos.

02

Crear un pipeline para los procesos de ETL.

03

Crear un DataWarehouse que corra de manera local o en un proveedor de servicios en la nube.



04

Automatizar todo el flujo de trabajo.

05

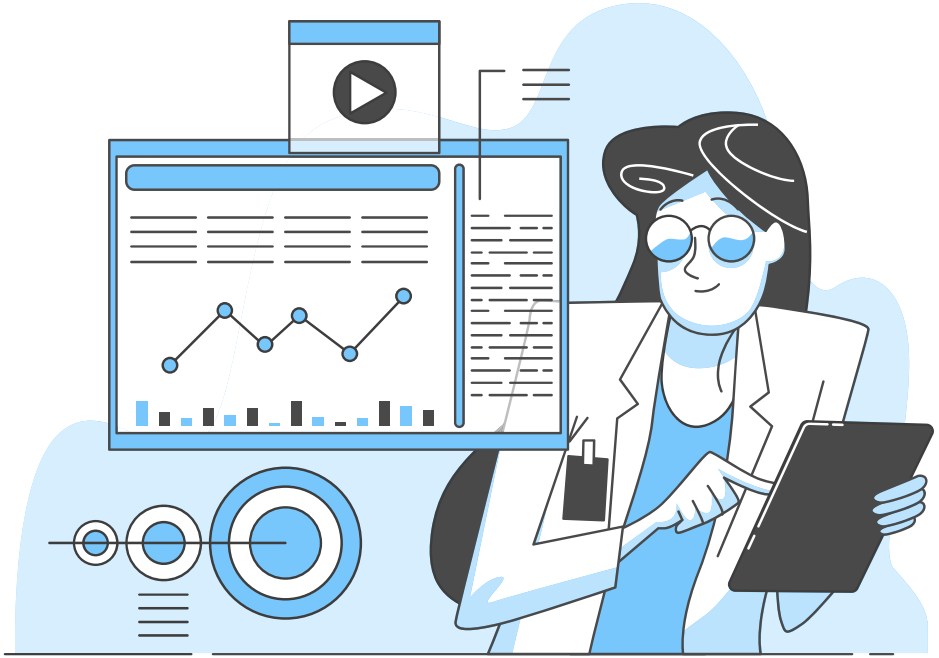
Realizar un análisis acompañado de KPIs para el buen desempeño y mejora del e-commerce.

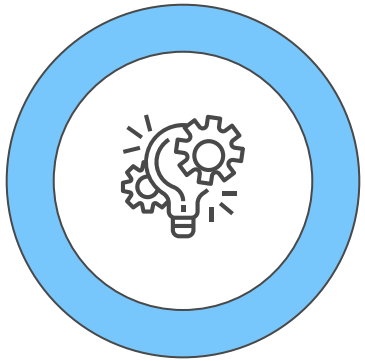
06

Confeccionar un reporte que incluya un dashboard que brinde información pertinente para la toma de decisiones del negocio.

07

Entrenar y poner en producción un modelo de machine learning que sea de utilidad al negocio en lo referente a dar solución a la problemática planteada.



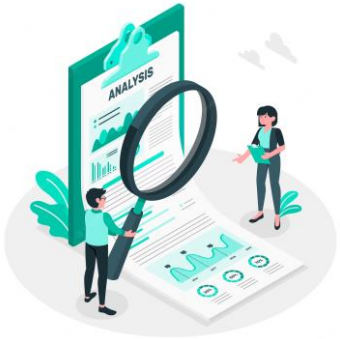


Alcance del proyecto:

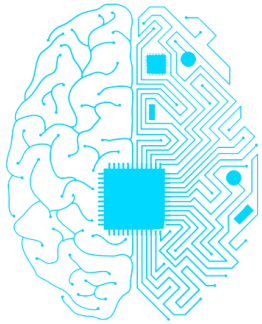
La finalidad de este trabajo es que nuestro equipo pueda generar información, a través de procesos de ingeniería y análisis de datos, que le permita al cliente disponer de fundamentos o de soporte para la toma de decisiones orientados a alcanzar los objetivos últimos de negocios de su empresa.

Solución propuesta

Dashboard y Reporte



Machine Learning



Stack tecnológico

Diario

01

Google Meet

Dailies con el Henry Mentor

02

Discord

Reuniones, debates, consultas de código

03

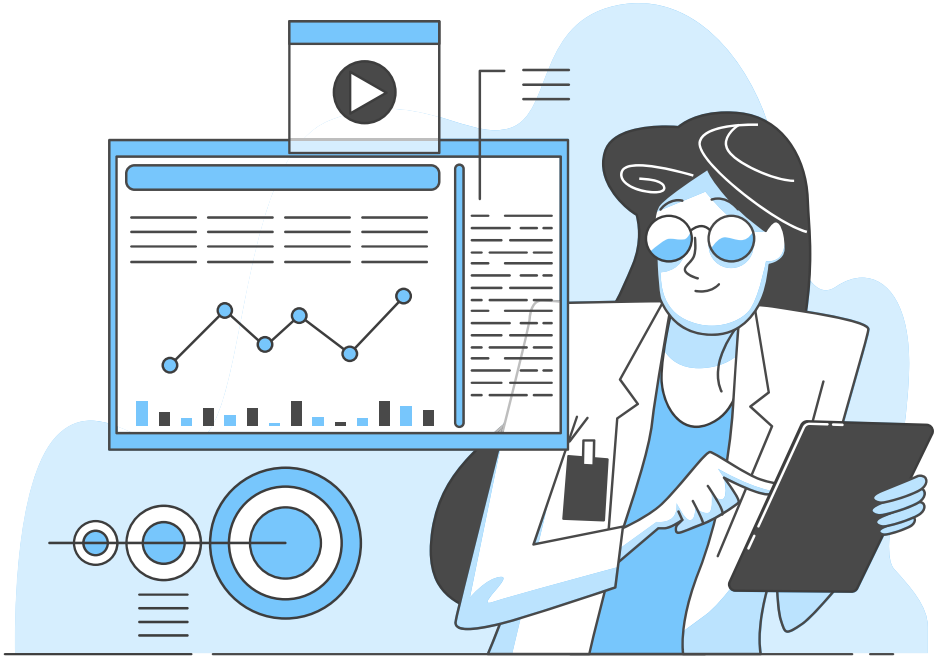
Trello

Organización del proyecto, asignación de tareas

04

GitHub

Administración y control del código del proyecto



Principal

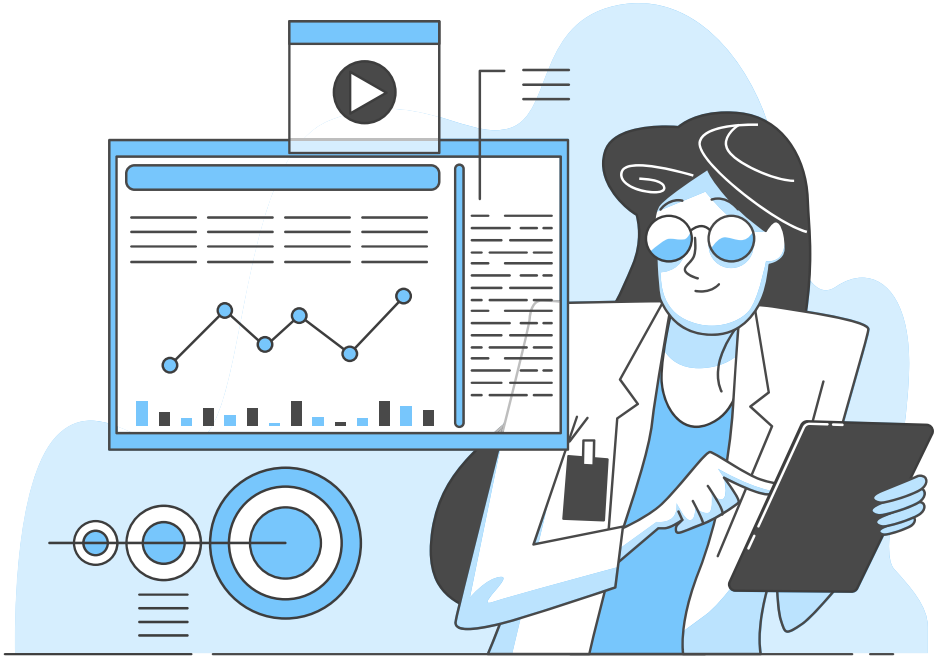
01
Python
Programación, ETL, EDA, Machine Learning

02
MySQL
Gestión de base de datos

03
Docker
Containerizacion de la app

04
FastAPI
Gestión de flujos de trabajo

05
PowerBI
Análisis de datos, informes, dashboards



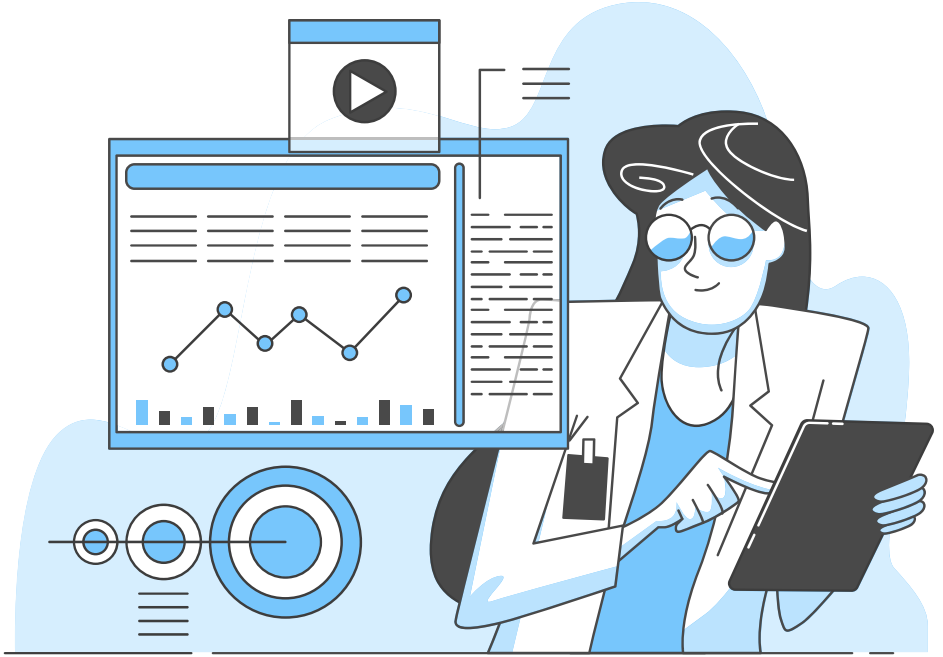
Google Cloud Platform

Plus

01 Cloud Run

02 Cloud SQL

03 Cloud Scheduler



Metodología de trabajo

El proyecto se desarrollará bajo los lineamientos de la metodología Scrum. Cada día se realizará una reunión de unos 20 minutos de duración, donde cada miembro del equipo comentará su avance y/o dificultades que se le presentaron el día anterior y luego se establecerán las tareas a realizar en esa jornada.

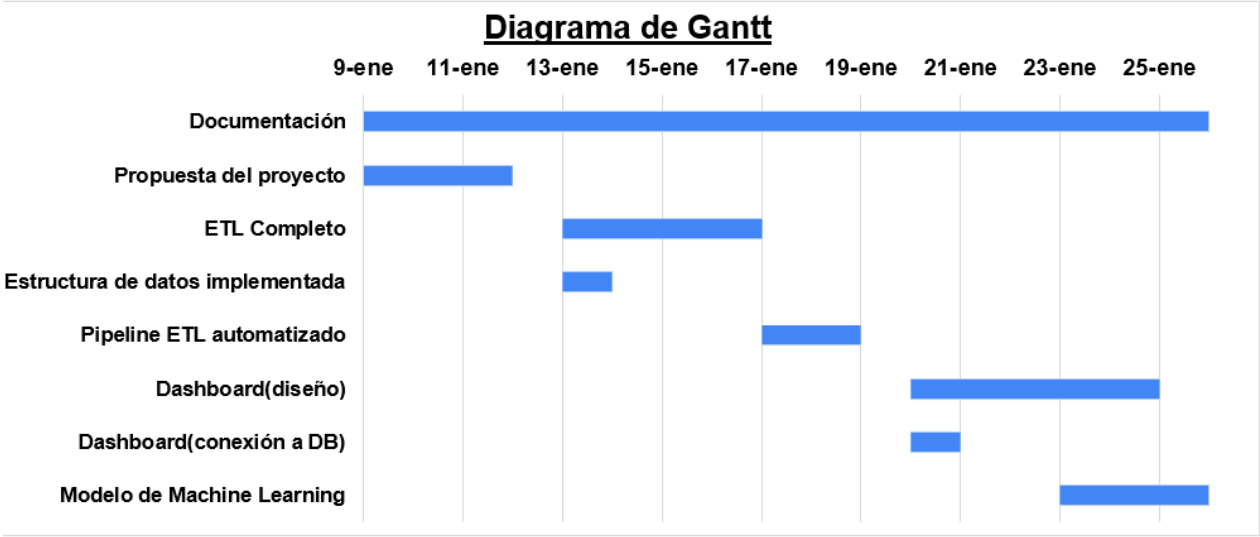
Para poder llevar adelante las dailies se realizará al finalizar cada semana una planning, la misma consiste en el armado y distribución de las tareas y subtareas, teniendo en cuenta que cada sprint tiene una duración de una semana y cada viernes se deben presentar los avances del proyecto al cliente .

Luego de finalizar cada sprint reviews (demo), el equipo hará una análisis retrospectivo, para estudiar la manera en que se trabajó esa semana ,qué podría mejorar y qué cosas continuar haciendo, teniendo en cuenta también el feedback recibido por parte del cliente.

Para organizar y distribuir tareas se utilizará la plataforma Trello.

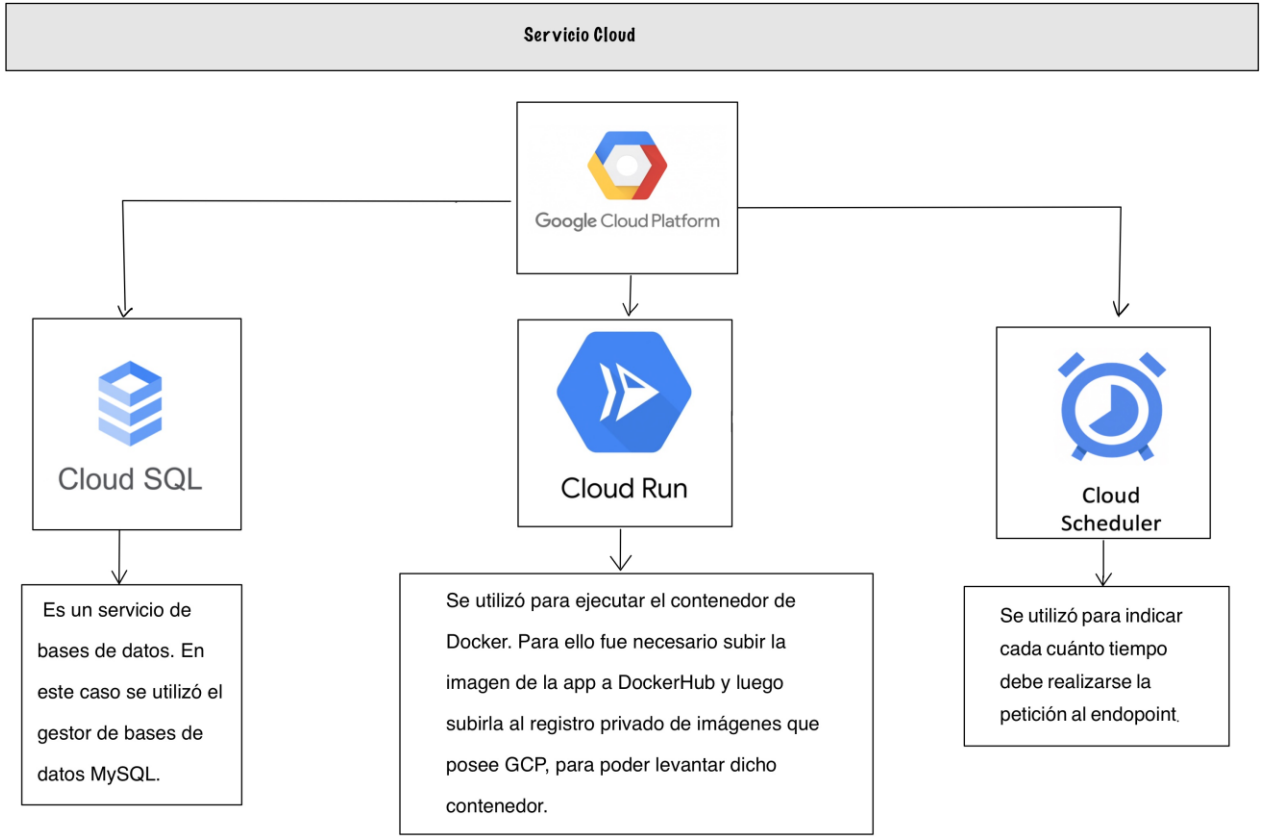
Diagrama de Gantt

	Inicio	Duración	Fin	% Porcentaje Completado	Dias completados
Documentación	9-ene	17	26-ene	100%	17
Propuesta del proyecto	9-ene	3	12-ene	100%	3
ETL Completo	13-ene	4	17-ene	100%	4
Estructura de datos implementada	13-ene	1	14-ene	100%	1
Pipeline ETL automatizado	17-ene	2	19-ene	100%	2
Dashboard(diseño)	20-ene	5	25-ene	100%	5
Dashboard(conexión a DB)	20-ene	1	21-ene	100%	1
Modelo de Machine Learning	23-ene	4	26-ene	100%	4



Creación de la API

Flujo de trabajo



Workflow

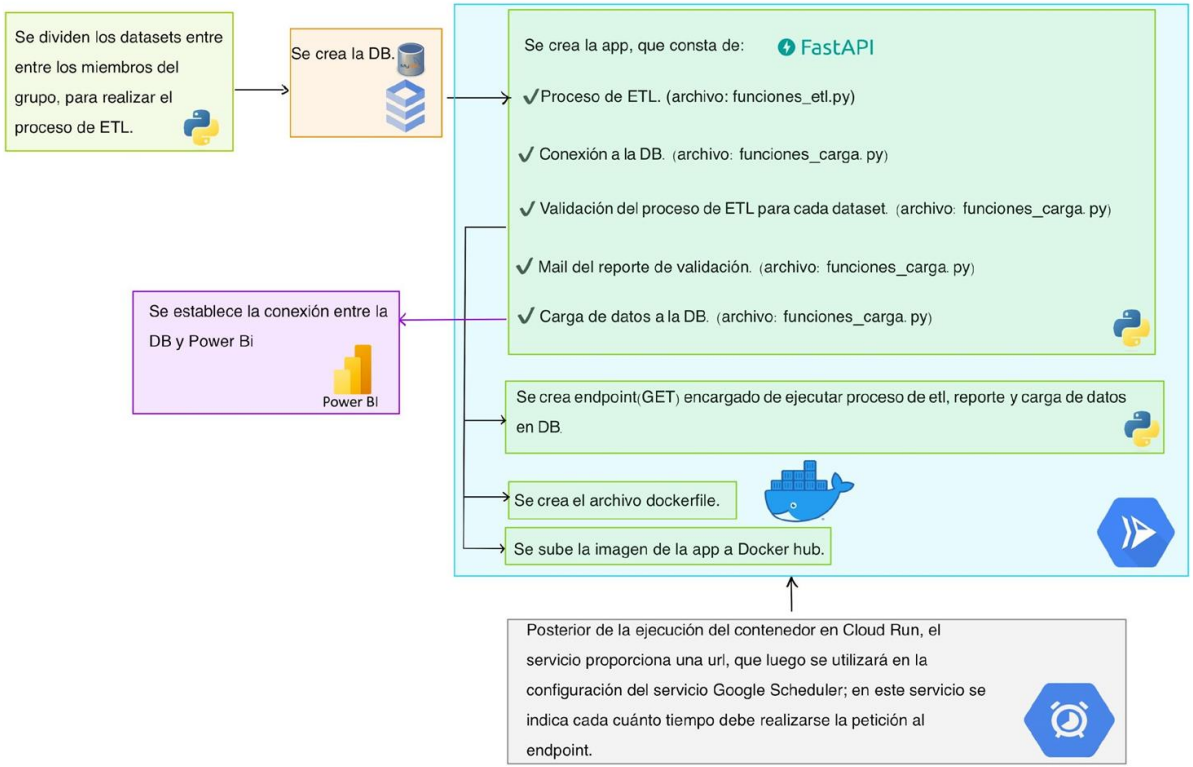
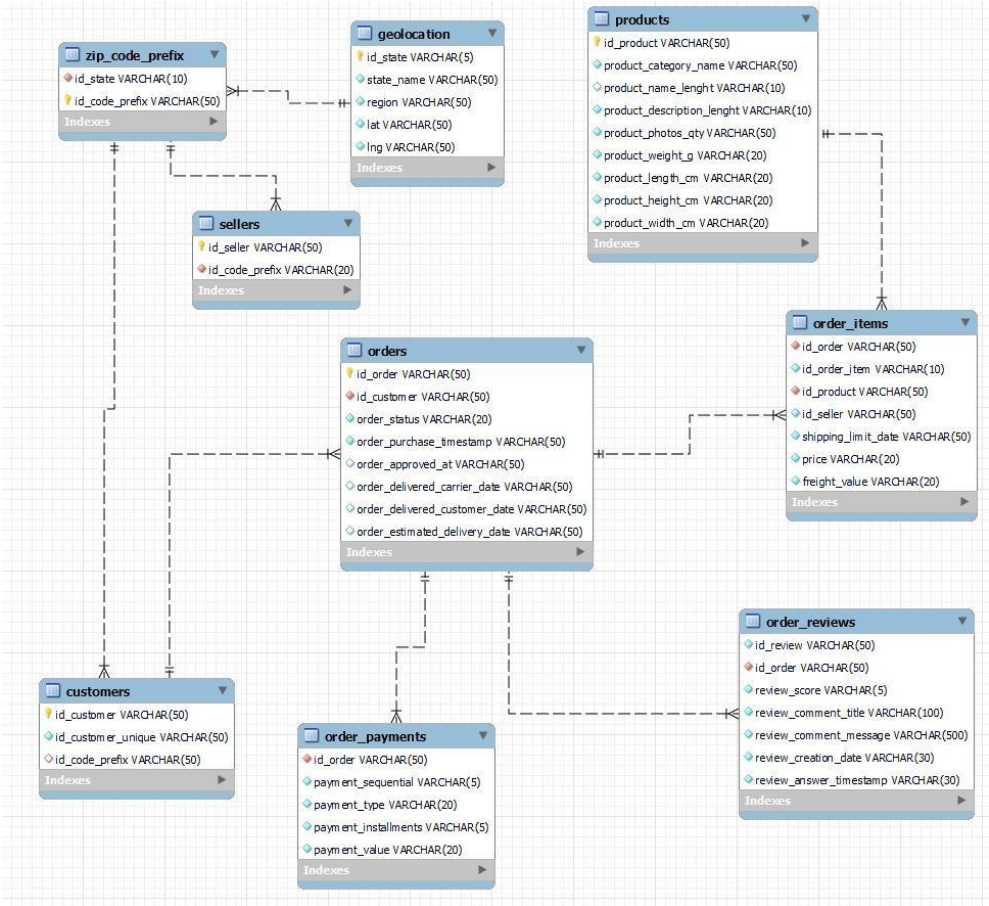

Se dividen los datasets entre entre los miembros del grupo, para realizar el proceso de ETL.
Se crea la DB.
Se crea la app, que consta de:
FastAPI
✓Proceso de ETL. (archivo: funciones_etl.py)
✓ Conexión a la DB. (archivo: funciones_carga. py)
✓ Validación del proceso de ETL para cada dataset. (archivo: funciones_carga. py)
✓ Mail del reporte de validación. (archivo: funciones_carga. py)
✓ Carga de datos a la DB. (archivo: funciones_carga. py)
Se establece la conexión entre la DB y Power Bi
Power BI
Se crea endpoint(GET) encargado de ejecutar proceso de etl, reporte y carga de datos en DB.
Se crea el archivo dockerfile.
Se sube la imagen de la app a Docker hub.
Posterior de la ejecución del contenedor en Cloud Run, el servicio proporciona una url, que luego se utilizará en la configuración del servicio Google Scheduler; en este servicio se indica cada cuánto tiempo debe realizarse la petición al endpoint.

Diagrama Entidad–Relación



Diccionario de datos

PRODUCTS

Llave	Campo	Tipo	Descripción
PK	id_product	string	Identificador del producto
	product_category_name	string	Categoría del producto
	product_name_lenght	int	Longitud del nombre del producto
	product_description_lenght	float	Longitud de la descripción del producto
	product_photos_qty	int	Cantidad de fotos del producto
	product_weight_g	float	Peso del producto
	product_length_cm	float	Longitud del producto
	product_height_cm	float	Altura del producto
	product_width_cm	float	Ancho del producto

SELLERS

Llave	Campo	Tipo	Descripción
PK	id_seller	string	Identificador del vendedor
FK	id_code_prefix	int	Identificador del prefijo de código postal del vendedor

CUSTOMERS

Llave	Campo	Tipo	Descripción
PK	id_customer	string	Clave para el conjunto de datos Pedidos. Cada id_order tiene un ID de cliente único.
	id_customer_unique	string	Identificador único del cliente
	id_code_prefix	int	Prefijo de código postal del cliente

ORDERS

Llave	Campo	Tipo	Descripción
PK	id_order	string	Identificador del pedido
FK	id_customer	string	Identificador del cliente
	order_status	string	Situación del pedido
	order_purchase_timestamp	datetime	Fecha y hora de compra
	order_approved_at	datetime	Fecha de aprobación del pedido
	order_delivered_carrier_date	datetime	Fecha y hora de cuando se envió el pedido al socio logístico.
	order_delivered_customer_date	datetime	Fecha y hora de la entrega real del pedido al cliente.
	order_estimated_delivery_date	datetime	Fecha y hora estimada de entrega que se informó al cliente en el punto de compra.

ORDER_REVIEWS

Llave	Campo	Tipo	Descripción
PK	id_review	string	Identificador de la reseña de un pedido
FK	id_order	string	Identificador del pedido al que pertenece la reseña
	review_score	int	Puntaje de la reseña
	review_comment_title	string	Título del comentario de la reseña dejada por el cliente, en portugués
	review_comment_message	string	Mensaje de comentario de la reseña dejada por el cliente, en portugués
	review_creation_date	datetime	Muestra la marca de tiempo cuando se envió la encuesta de satisfacción al cliente
	review_answer_timestamp	datetime	Muestra la marca de tiempo de la respuesta de la encuesta de satisfacción

ORDER_PAYMENTS

Llave	Campo	Tipo	Descripción
PK	id_order	string	Identificador del pago de un pedido
	payment_sequential	int	Un cliente puede pagar un pedido con más de un método de pago. En tales casos, se creará una secuencia.
	payment_type	string	Tipo de pago
	payment_installments	int	Cantidad de cuotas del pago
	payment_value	float	Valor del pago

ORDER_ITEMS

Llave	Campo	Tipo	Descripción
PK	id_order	string	Identificador del pedido
FK	id_order_item	string	Identificador del ítem de un pedido
FK	id_product	string	Identificador del producto de un pedido
FK	id_seller	string	Identificador del vendedor de un pedido
	shipping_limit_date	datetime	Fecha límite de envío
	price	float	Precio del ítem
	freight_value	float	Valor del flete

ZIP_CODE_PREFIX

Llave	Campo	Tipo	Descripción
PK	id_state	string	Identificador del estado
	id_code_prefix	int	Identificador del código postal del estado

GEOLOCATION

Llave	Campo	Tipo	Descripción
PK	id_state	string	Identificador del estado
	state_name	string	Nombre del estado
	region	string	Region del estado
	lat	float	Latitud del estado
	lng	float	Longitud del estado

Informe de Análisis

Como primera medida, debemos enfatizar que el objetivo general de Olist es vender más productos a un mayor número de clientes, o lo que es lo mismo, incrementar sus ventas e incrementar el número de compradores.
Para poder llevar a cabo esta tarea, se plantean objetivos específicos que serán medidos mediante indicadores claves que nos permitirán realizar un seguimiento de manera de poder evaluar en qué medida se están alcanzando estos objetivos.

Objetivos y KPI's

Implementación de Indicadores Claves de Rendimiento

Implementaremos KPIs que nos darán una medida cuantificable del desempeño de la empresa en relación a objetivos específicos, de manera de poder evaluar si estamos alcanzando lo que nos hemos propuesto:

• **Crecimiento de ventas**

Este KPI está asociado al objetivo específico de la empresa de **incrementar la cantidad de ventas en un 35% para el semestre siguiente**.

La métrica utilizada responde a la siguiente fórmula:

$$Crecimiento\ porcentual\ de\ ventas = \left(\frac{Ventas\ periodo\ 1}{Ventas\ periodo\ 0} - 1\right) * 100$$

Cabe aclarar que el crecimiento de las ventas se evaluará tanto en cantidades como en montos. Los valores porcentuales que obtengamos al analizar esta métrica nos dirán si se está cumpliendo el objetivo o no, y en qué medida. También incluiremos gráficos que nos permitirán monitorear el desempeño no solo de forma semestral sino también bajo otros periodos de tiempo.

• **Crecimiento de nuevos clientes**

Este KPI está asociado al objetivo específico de la empresa de **incrementar la cantidad de nuevos clientes en un 25% para el semestre siguiente**.

La métrica utilizada responde a la siguiente fórmula:

$$Crecimiento\ porcentual\ de\ nuevos\ clientes = \left(\frac{Nuevos\ Clientes\ periodo\ 1}{Nuevos\ clientes\ periodo\ 0} - 1\right) * 100$$

Este indicador sigue la misma lógica que el anterior pero aplicado a clientes en vez de ventas. De su monitoreo podremos evaluar si se está alcanzando el objetivo o no. Para apoyar al análisis también se incluyen gráficos que permiten visualizar el desempeño en distintos rangos de tiempo.

• **Ventas Promedio por Cliente**

Este KPI está asociado al objetivo específico de la empresa de **incrementar el promedio de ventas por cliente en un 25% para el semestre siguiente**.

La métrica utilizada responde a la siguiente fórmula:

$$Ventas\ promedio\ por\ cliente = \frac{Total\ Ventas}{Total\ Clientes}$$

Para este caso, las ventas promedio también se harán sobre cantidades y sobre montos. También incluimos una métrica para evaluar la evolución porcentual, como así también gráficos que permiten monitorear el desempeño en diferentes rangos de tiempo.

• **Tasa de Clientes Inactivos**

Este KPI está asociado al objetivo específico de la empresa de **mantener los clientes inactivos por debajo del 50% para el semestre siguiente**. Por clientes inactivos tomamos a aquellos clientes que no han realizado compras en los últimos 6 meses.
La métrica utilizada responde a la siguiente fórmula:

$$Tasa\ de\ clientes\ inactivos = \frac{Total\ Clientes\ Inactivos}{Total\ Clientes} * 100$$

Este KPI es muy similar al CCR (Customer Churn Rate) con la sutil diferencia que aquí los clientes inactivos no han abandonado su suscripción, sino que no han comprado en los últimos 6 meses.
Mientras más alta sea la tasa de clientes inactivos más clientes activos e ingresos estará perdiendo la empresa. Como soporte al monitoreo de este indicador se dispondrá de gráficos y métricas que expondrán la cantidad de clientes activos, de clientes inactivos, sus variaciones porcentuales, y todo esto podrá ser visible en diferentes rangos de tiempo.
Hemos decidido monitorear este KPI porque a la empresa no le interesa aumentar solamente la cantidad de suscriptores o compradores nuevos, sino que también mantenerlos. El objetivo final siempre es incrementar los beneficios, y mientras más clientes activos tengamos mayor será la probabilidad de alcanzarlo. Evitar que nuestros clientes activos se transformen en inactivos es de suma importancia para potenciar el negocio.

• Tiempo promedio de aprobación de compra

Este KPI está asociado al objetivo específico de la empresa de **mantener el tiempo de aprobación de la compra dentro de las 12 horas desde generada la orden, para el semestre siguiente**.

La métrica utilizada responde a la siguiente fórmula:

$$Tiempo\ promedio\ de\ aprobación\ de\ compra = \sum \frac{(Hs\ de\ aprobación - Hs\ de\ compra)}{Total\ de\ compras\ aprobadas}$$

Este KPI nos permite monitorear el tiempo en que tarda la empresa en aprobar la compra. Esta parte del proceso de logística está controlada enteramente por la empresa y puede tomar acciones concretas para mejorarla. Si bien cada operación debería hacer sonar una alarma si su aprobación demora más de 12 horas, llevar un indicador que nos permita evaluar el promedio de tiempo para todas las operaciones es de gran utilidad. Apoyamos este indicador con otros gráficos que nos permiten visualizar la distribución de los tiempos de aprobación.

• **Tiempo promedio de despacho de compra**

Este KPI está asociado al objetivo específico de la empresa de **lograr que las compras se despachen dentro del plazo de 3 días desde recibida la aprobación de la orden, para el semestre siguiente**.

La métrica utilizada responde a la siguiente fórmula:

$$Tiempo\ promedio\ de\ despacho\ de\ compra = \sum \frac{(Fecha\ de\ despacho - Fecha\ de\ compra)}{Total\ de\ compras\ despachadas}$$

Este KPI es muy similar al anterior, aunque en este caso lo que se busca es medir el tiempo de despacho. Esta parte del proceso no está controlada en su totalidad por la empresa, ya que son los vendedores quienes despachan las compras. Sin embargo, esto no es fundamento para no medirlo e intentar mejorarlo, ya que la empresa y los vendedores deben trabajar como un todo para lograr que la experiencia del usuario sea satisfactoria. Adicionalmente, la empresa tiene el poder de dictar los términos y condiciones a los que los vendedores deben atenerse, por lo que los tiempos de logística pueden llegar a seguir un estándar.

- **Entregas a Tiempo**

Este KPI está asociado al objetivo específico de la empresa de lograr que el 95% de los envíos lleguen dentro del plazo estimado, para el semestre siguiente.
La métrica utilizada responde a la siguiente fórmula:

$$Porcentaje\ de\ entregas\ a\ tiempo = \frac{Cantidad\ de\ entregas\ a\ tiempo}{Total\ de\ entregas\ realizadas} * 100$$

Este KPI nos permite monitorear el desempeño de toda la cadena de logística. Si el valor que nos arroja está por fuera del planteado como objetivo, quiere decir que algún eslabón de la cadena puede estar fallando. Las entregas a tiempo son un factor muy importante para el comprador, y las demoras influyen muy negativamente en la intención de compra y valoración del servicio. Por ende, poder evaluar si se está llegando a este objetivo del 95% de los envíos a tiempo es vital para la empresa.
El análisis de la logística también se verá apoyado por otras métricas y gráficos que permitirán tener un mejor panorama de la situación de la empresa en este punto.

Análisis

Análisis de Ventas

1) Se consideran las ventas entre el 01/01/2017 y el 01/09/2018, rango de tiempo que expone de forma más clara la operatoria de la empresa.

2) La tendencia general de ventas, tanto en ingresos como en cantidades, es ascendente. Durante 2017 la línea de tendencia tiende a ser lineal, mostrando una pendiente de aproximadamente 45 grados. En los 2 primeros trimestres del 2018 se comienza a aplanar. El último trimestre registrado cuenta con un mes menos de datos, pero en términos generales sigue el mismo comportamiento.

3) La variación porcentual, tanto en ingresos como en cantidades, a lo largo de los trimestres ha fluctuado sin un patrón evidente, aunque con un crecimiento sostenido hasta 2018.

4) Observando el volumen de operaciones, deducimos que se trata de una empresa en sus inicios, y es obvio que esto repercute en las variaciones porcentuales de periodo a periodo, ya que incrementos o decrementos pequeños generan grandes fluctuaciones.

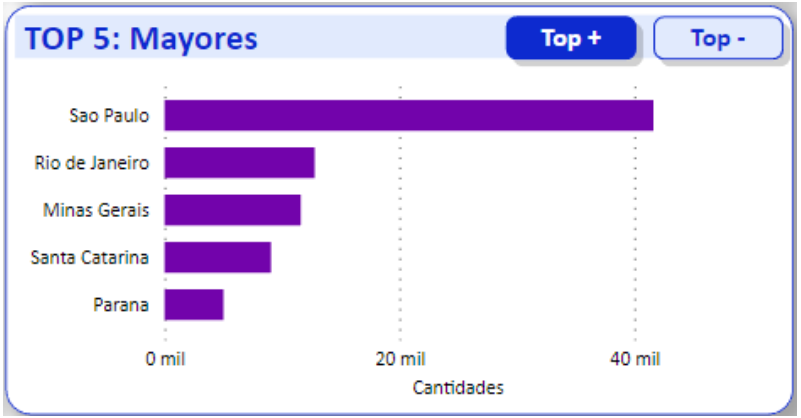
5) Se observa un estancamiento (incluso una leve disminución) en las ventas en el primer semestre de 2018. Las cantidades vendidas tienen a estabilizarse alrededor de las 6500 unidades y los ingresos alrededor de $1.000.000.

6) Si observamos la variación semestral aquí también se observa que el crecimiento porcentual se va desacelerando. Para las cantidades, pasa de un 108,68% para el 2do semestre de 2017, a un 35,08% para el 1er semestre de 2018. En cuanto a los ingresos, pasa de un 103,07% para el 2do semestre de 2017, a un 37,53% para el 1er semestre de 2018.

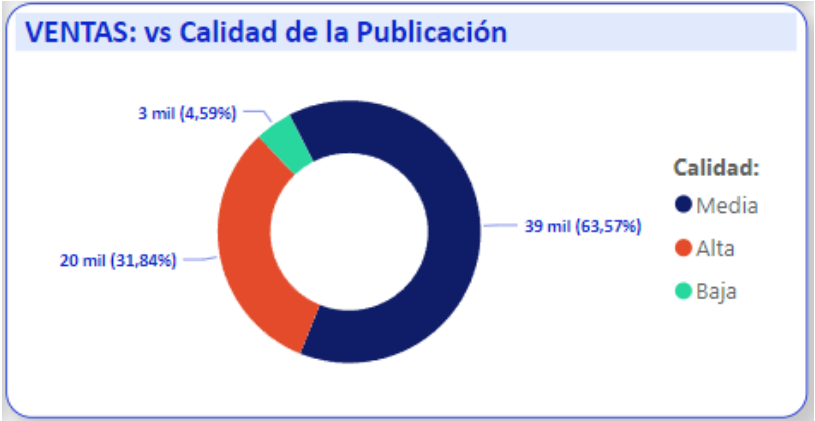
7) En el 2do trimestre de 2018 se observa un retroceso de casi el 5% en las cantidades vendidas respecto del trimestre anterior. Sin embargo, los ingresos reportados son mayores para dicho periodo.

8) El mes de mayores ingresos y mayor cantidad vendida es el de noviembre de 2017. El 24 de dicho mes hubo un Black Friday, que fundamenta ese pico de ventas.

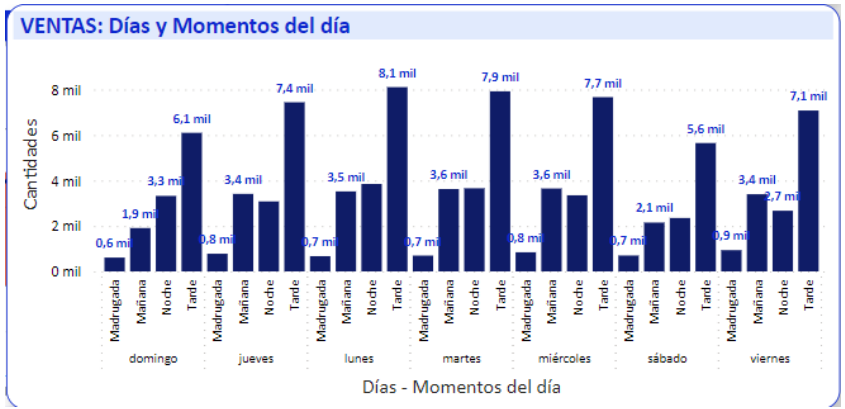
9) La Región Sudeste es la que más ventas reporta, y 3 de sus estados están dentro del Top 3 de mayores ventas. Esos estados son Sao Pablo, Rio de Janeiro y Minas Gerais. La Región Sur le sigue en importancia, mientras que la región Norte es la de menor alcance.



10) Asignamos una puntuación para las publicaciones de productos. Esa puntuación tiene en cuenta la cantidad de fotos y la longitud de la descripción incluidas en la publicación del producto. A más fotos y/o más descripción mejor se considera la publicación. Categorizamos los resultados en “Baja”, “Media” y “Alta”. De la comparativa obtenemos que las publicaciones con más ventas son las categorizadas como “Media”, con un 100% de ventas más a la categoría “Alta”. Las publicaciones de calidad “Baja” representan menos del 5% de todas las ventas. Podemos deducir que la cantidad de fotos y descripción debe ser mesurada, ya que la categoría “Media” tiene un efecto considerablemente superior a la categoría “Alta”. Una “Baja” calidad de la publicación conlleva a ventas ínfimas de ese producto.

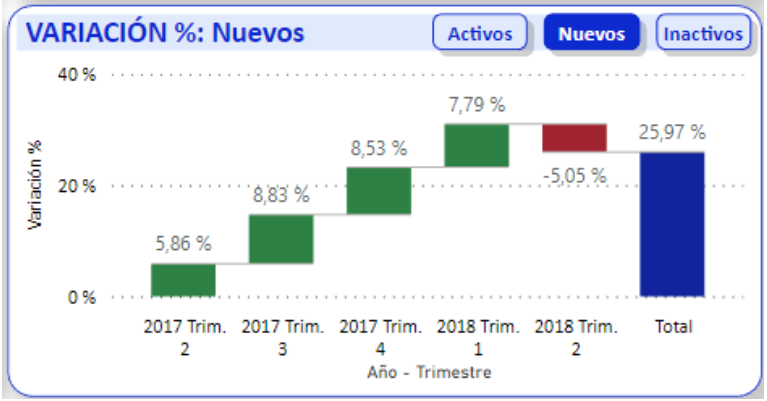


11) Dividimos las ventas por días de la semana y por momento del día. La tarde parece ser el momento preferido de compra, reportando más del 100% de ventas con respecto a la noche y la mañana (que tienen cantidades parecidas). El día con más ventas es el lunes, y va disminuyendo a medida que se acerca al fin de semana. Los sábados son el día menos elegido para comprar.

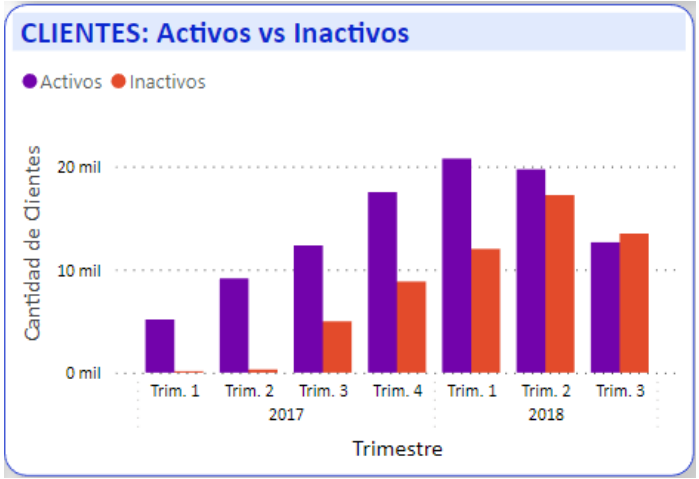


Análisis de Clientes

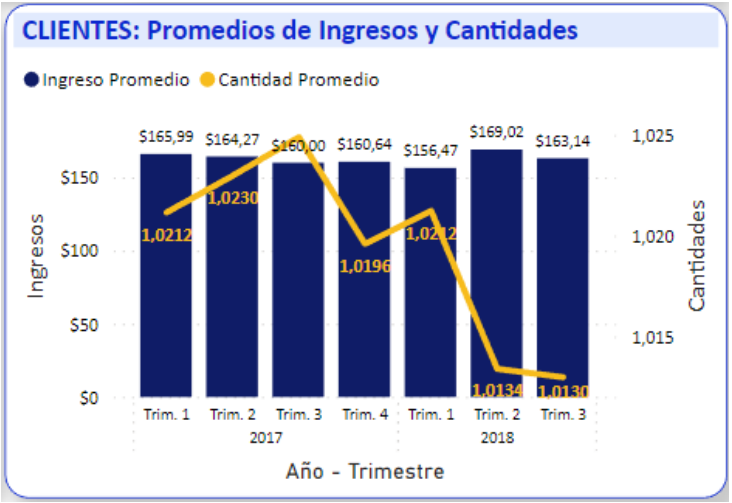
1) A lo largo del tiempo de operatoria, la empresa ha adquirido un total de 95,8 mil clientes.
2) La cantidad de clientes nuevos por trimestre ha ido en aumento hasta el primer trimestre de 2018, mostrando un crecimiento porcentual de aproximadamente 8% por trimestre. En el segundo trimestre de 2018 se observa una caída porcentual del 5% respecto del trimestre anterior.



3) Definimos a los clientes inactivos como aquellos que no han realizado compras en los últimos 6 meses. Al igual que con la cantidad de nuevos clientes, podemos observar que los clientes inactivos también han ido aumentando, incluso a una tasa mayor que los nuevos clientes. Para el tercer trimestre de 2018, y por primera vez, la cantidad de clientes inactivos supera a la cantidad de nuevos clientes.

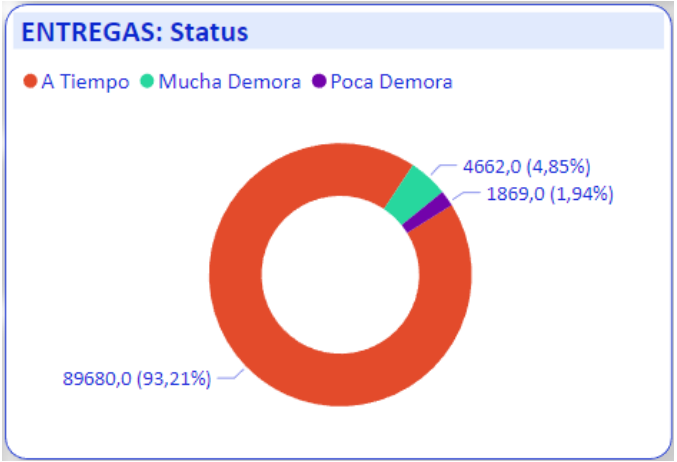


4) El promedio de cantidades compradas por cliente activo es cercano a la unidad en todos los trimestres. El promedio de ingresos por cliente activo varía en el rango de los $150 a $170 aproximadamente, y hasta el segundo trimestre del 2018, cuando hace un pico, venía decreciendo paulatinamente.

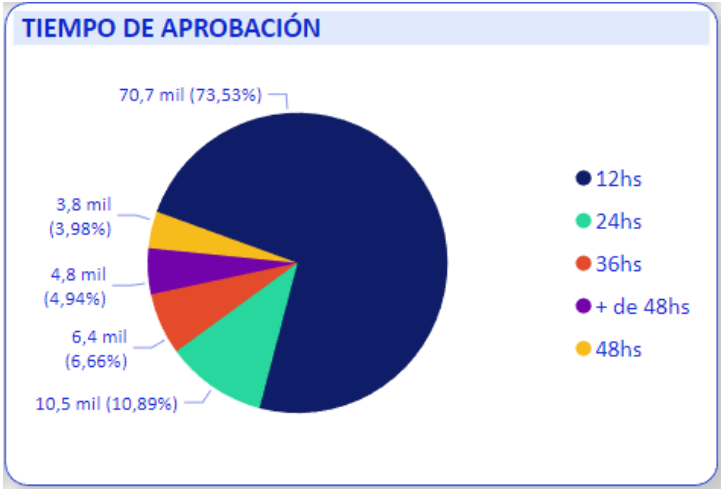


Análisis de la Logística

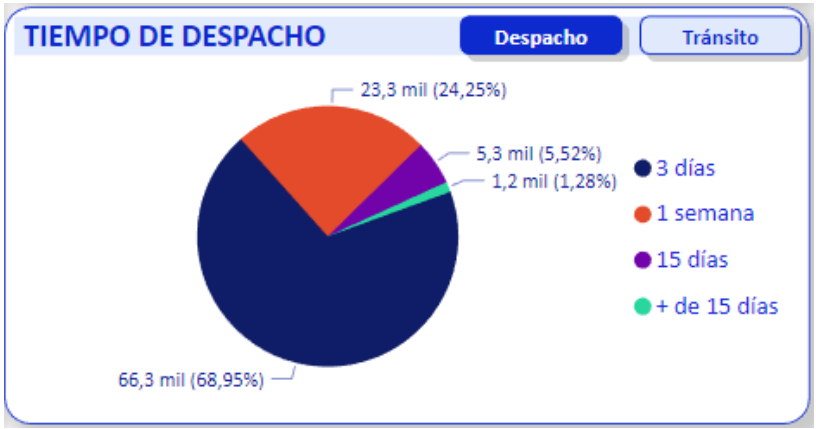
1) Se han despachado y entregado 96,2 mil órdenes.
2) El 93% de los envíos han llegado a tiempo, dentro del plazo estimado. De las entregas por fuera del tiempo estimado, el 2% se realizó dentro del plazo de 3 días (Poca Demora), mientras que el 5% tuvo 4 o más días de demora (Mucha Demora).



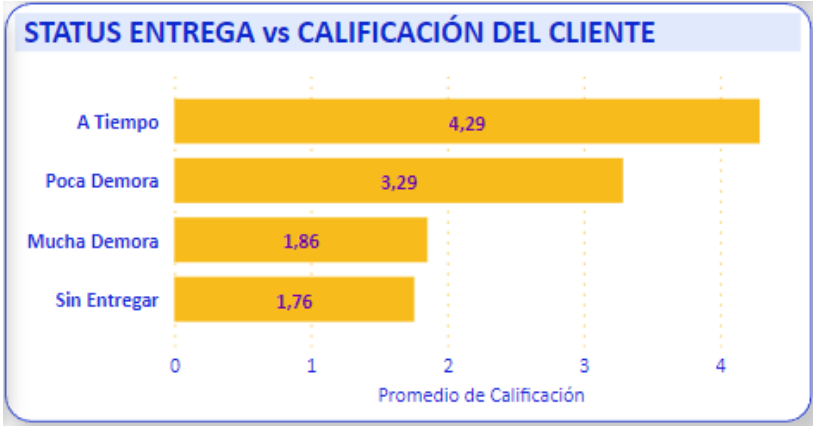
3) Respecto del tiempo de aprobación de la compra, el 73% de las órdenes es aprobada dentro de las 12 horas de realizada la compra. El 11% toma 24 horas en ser aprobada, mientras que el 7% se hace en 36 horas. El 9% restante tarda más de 36 horas en ser aprobada.



4) Un tercio de los vendedores despacha los productos hacia el repartidor dentro de los 3 días desde el momento en que el comprador efectúa la compra. Un 24% lo hace en una semana, mientras que el restante lo hace en más de 7 días.



6) Cruzando datos con las calificaciones de las compras por parte del cliente, notamos las entregas a tiempo tienen las mejores calificaciones, mientras que, a mayor demora de la entrega, menor es la puntuación que recibe el vendedor. Es claro que el tiempo de entrega es un factor determinante en la satisfacción del cliente.



Argumentos para la instalación de un Hub en Río de Janeiro

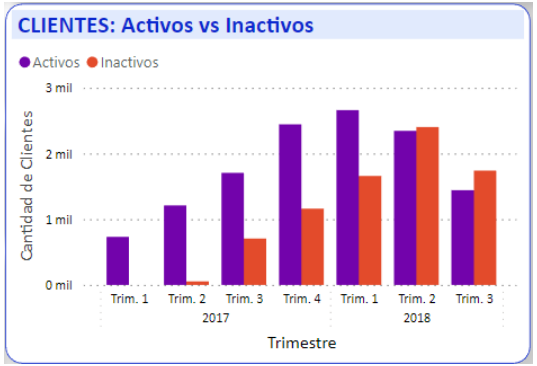
Río de Janeiro se lleva nuestra atención cuando notamos que las entregas a tiempo no solo están por debajo de la media del país, sino que también están por debajo del porcentaje de entregas a tiempo de los Estados vecinos, tanto de zona Sudeste, Centro y Sur. Por esta razón, y siendo que su capital se encuentra muy cerca de Sao Paulo y que además es el segundo Estado de Brasil en cantidad de habitantes, decidimos analizar más a fondo a Río de Janeiro.

Encontramos que:
1) Las entregas a tiempo están a más de 5 puntos por debajo de la media del país, que es de 93,21%.

ENTREGAS A TIEMPO

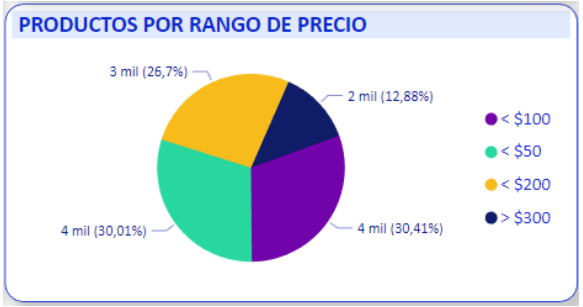
87,86 %

2) Los clientes inactivos vienen en aumento a tasas mayores que en otros lugares, e incluso es único lugar importante de Brasil en donde los clientes inactivos han sido superiores a los clientes nuevos en el 2do trimestre de 2018.

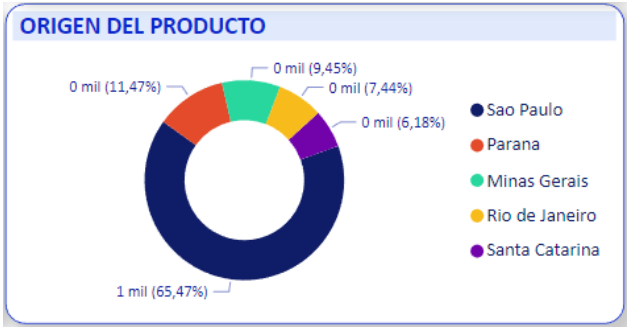


3) De las 5 categorías de productos más vendidas del país, todas son también las más vendidas en este Estado.

4) Los productos más vendidos son los de menor valor. Dos terceras partes cuestan menos de $100.



5) El 92% de los productos son adquiridos a vendedores de otros Estados, y en su mayoría son productos de menor valor. Los productos de mayor valor son adquiridos en mayor medida a vendedores locales.



6) Ya que los productos de menor valor son traídos en su mayoría de otro lado, los costos del flete son elevados en proporción. Se armó una métrica que es el ratio del valor flete sobre el valor de la orden, y se llegó a notar que en algunos casos el flete representa hasta el 60% del valor del producto. La media de este ratio es del 33%.

RATIO FLETE/ÓRDENES	TASA DE CLIENTES INACTIVOS
0,33	63,34 %

Sugerencia que nace del análisis de Río de Janeiro

Teniendo en cuenta que se trata del segundo Estado de mayor población de Brasil y, por tanto, un gran mercado potencial, no se puede dejar de sugerir alguna alternativa de acción para mejorar el desempeño de Olist en esta zona geográfica. Más aún, sabiendo que los Estados vecinos vienen teniendo mejores resultados.
Se sugiere la instalación de un centro de distribución o almacén de stock en la ciudad de Río de Janeiro o sobre rutas comerciales cercanas. De esta forma, se puede analizar la rotación de inventario de las categorías más vendidas, teniendo en cuenta datos históricos y de las ciudades más importantes, y mantener un stock de estos productos con el objeto de realizar entregas que lleguen más rápido al cliente, y a su vez disminuir, en el volumen, los costos de fletes. De esta forma el cliente obtendrá su paquete antes y más barato, aspectos que hemos encontrado cruciales para elegir y calificar el servicio. Poder abastecer mejor a una región tan importante como es esta implica mayor probabilidad de crecimiento para Olist en un futuro.

Machine Learning

Según el análisis que se realizó a los datos recibidos, se trabajó con el dataset orders, donde se encuentran las órdenes de compra, comprobamos que existen ventas hasta mediados de octubre del año 2018, razón por la cual decidimos hacer una predicción sobre las ventas hasta fines del mismo año. Por la distribución del mismo como serie temporal, optamos por una red neuronal auto regresiva utilizando la técnica de FORECASTING ya que es la más adecuada para predecir valores futuros con datos históricos. Esta red neuronal fue elaborada con la librería SKFOREST, la cual cuenta con una herramienta de backtesting, y para hacer las predicciones utilizamos el regresor de lightGBM.

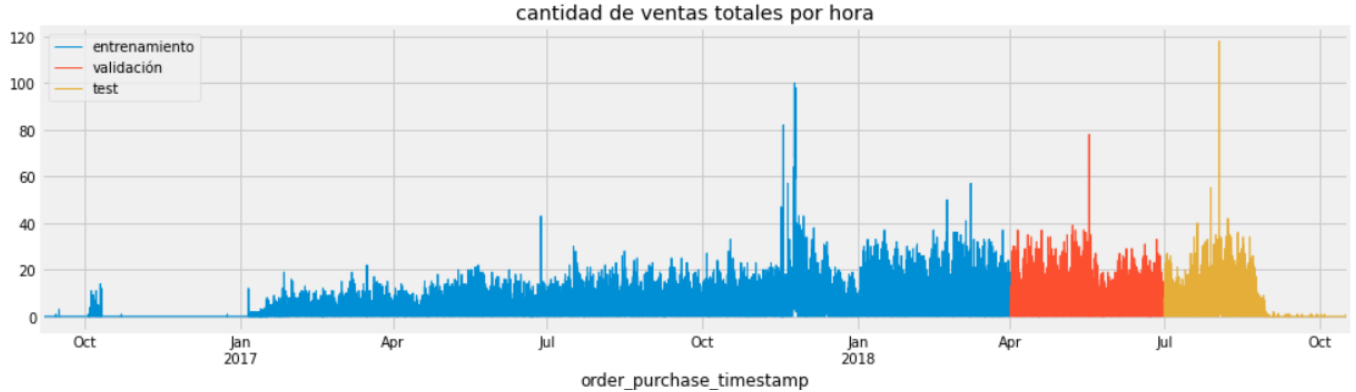
Las ventajas del machine learning aplicado al forecasting son:

- Facilidad para incluir variables exógenas además de las auto regresivas
- Permite incorporar comportamientos no lineales
- Elevada escalabilidad, aplicable cuando se dispone de grandes volúmenes de datos.

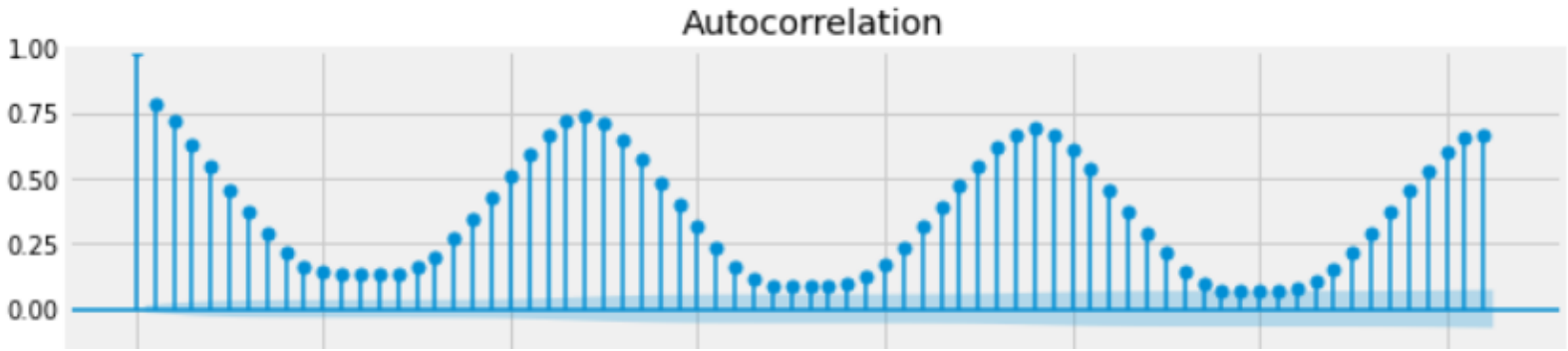
Para poder realizar el modelado, se tuvo que enfrentar algunos problemas como:

- Reestructurar los datos para adaptarlos a un problema de regresión
- A medida que las iteraciones avanzan, cada vez se utilizan más variables predichas por el modelo para seguir prediciendo.
- La validación cruzada no es aplicable a estos modelos, por los que se requieren procesos de validación como el backtesting

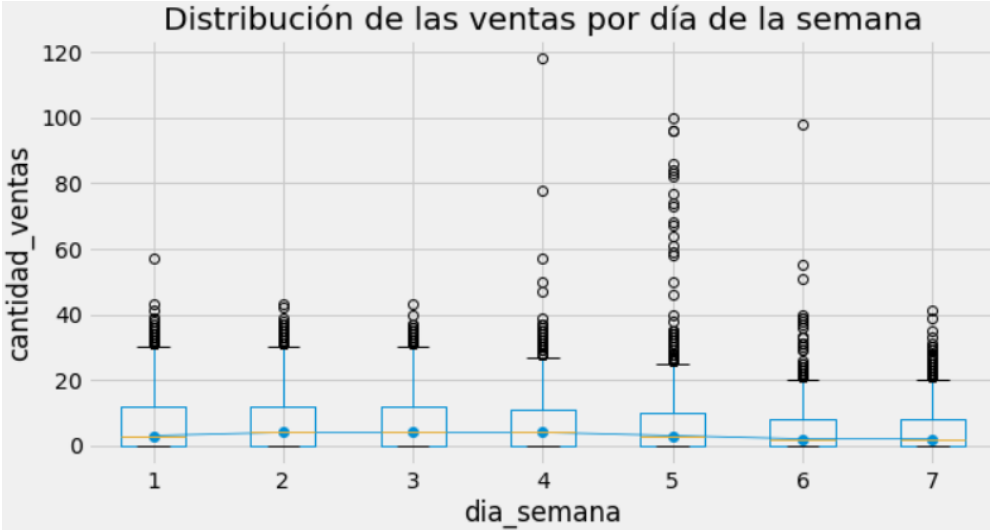
Para enfrentar el primer problema, la reestructuración, se crea una matriz haciendo conversiones para que cada valor este asociado a una ventana temporal que le precede (denominado lag) siendo de 1 hora cada una. Para poder entrenar la red de forma efectiva dividimos el dataset en intervalos de a 1 hora, tomando en total un set de entrenamiento de 19 meses, uno de validación y otro de testeo ambos de 3 meses.



Además, se analizó la correlación entre los lags y se pudo concluir que hay una directa relación entre lo que pasó el día anterior a la misma hora con el período actual bajo estudio



Para lograr una mayor precisión a la hora de predecir, tuvimos en cuenta variables exógenas como el día de la semana en la cual se realizó la compra, teniendo en cuenta si es un día laboral o no, lo cual mostró que las ventas van disminuyendo llegando al fin de semana.



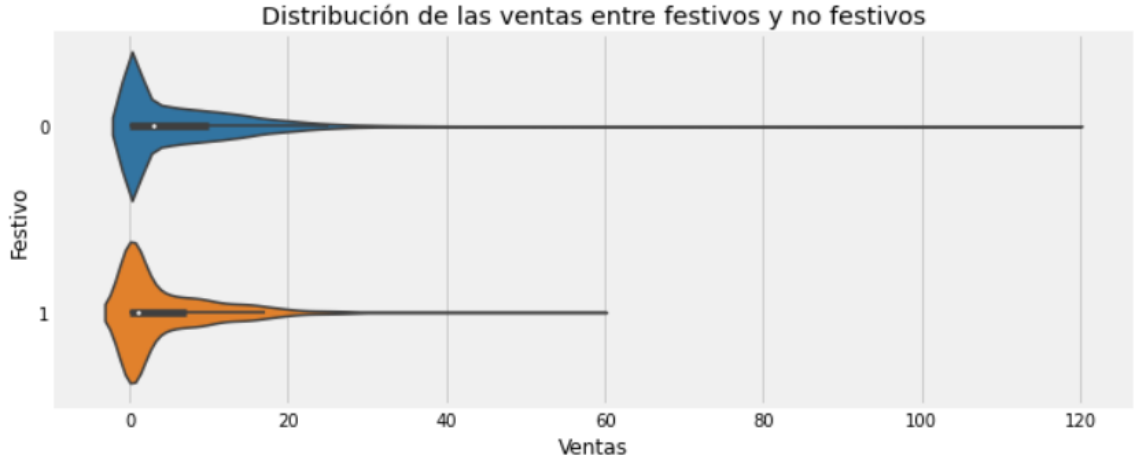
Otro factor que tuvimos en cuenta fue si es feriado nacional inamovible, listado a continuación

Feriados inamovibles

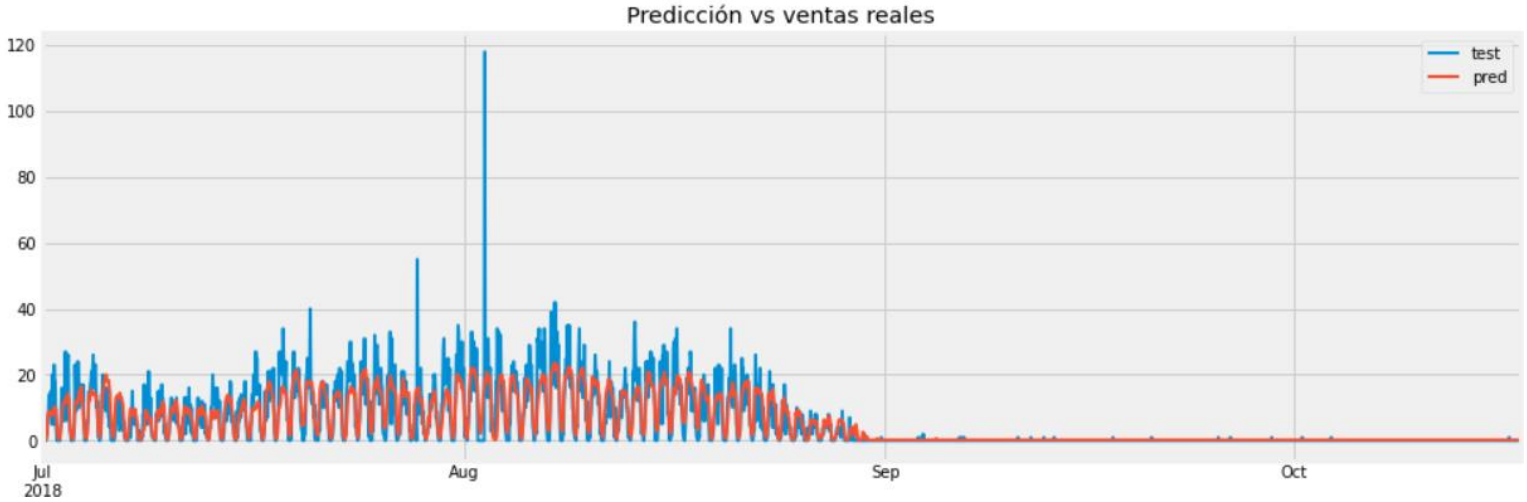
Los días no feriados se muestran en gris.

Fecha	Día festivo
1. ene.	Año Nuevo
20. ene.	San Sebastián (sólo Río de Janeiro)
25. ene.	Fundación de la ciudad de São Paulo (sólo São Paulo)
21. abr.	Día de Conmemoración Tiradentes
23. abr.	San Jorge (sólo Río de Janeiro)
1. mayo	Día del Trabajo
12. jun.	Día de los Amantes
9. jul.	Revolución constitucional (sólo São Paulo)
15. ago.	Día de la Asunción
28. ago.	Control y prevención del Escalpelo
7. sept.	Día de la Independencia
12. oct.	Día del Niño
12. oct.	Aparición de la Virgen María
15. oct.	Día del profesor
28. oct.	Día de la Administración Pública
2. nov.	Día de Todos los Santos
15. nov.	Proclamación de la República
20. nov.	Día de la Recordación de la Gente Negra (no a nivel nacional)
24. dic.	Nochebuena
25. dic.	Día de Navidad
26. dic.	Día de San Esteban
31. dic.	Noche Vieja

Como puede verse a continuación, tanto en días festivos como no, la distribución de ventas se concentra alrededor de 0, pero teniendo en cuenta que el pico de ventas es la mitad de los días de semana, es importante resaltar que al haber muchos menos días festivos en total, es una cantidad importante a tener en cuenta.

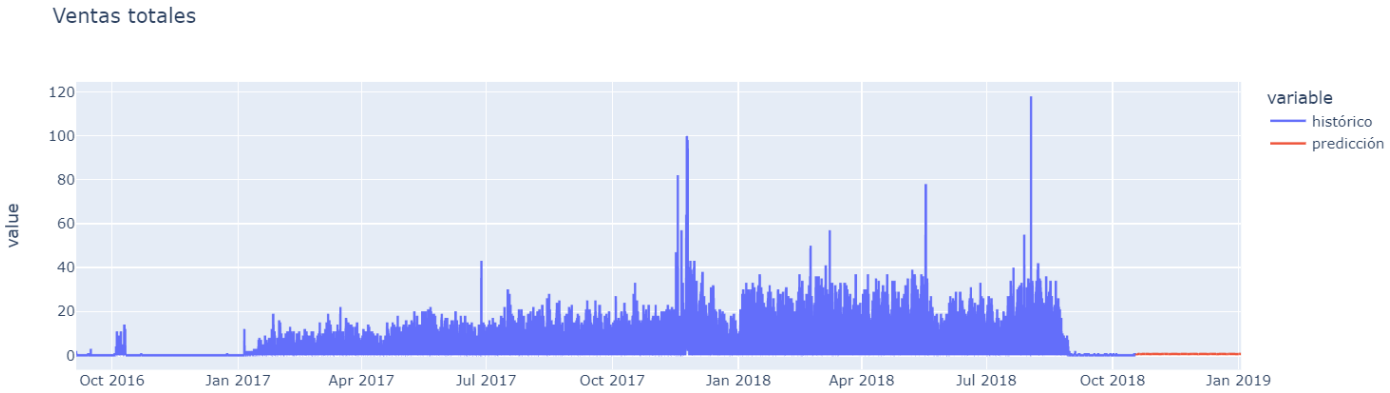


Para solventar el error que se acumula a medida que se utilizan las predicciones como datos históricos en el entrenamiento, se eligieron los mejores parámetros comparando 24 modelos con greed search, logrando un porcentaje de incertidumbre igual al 35%.

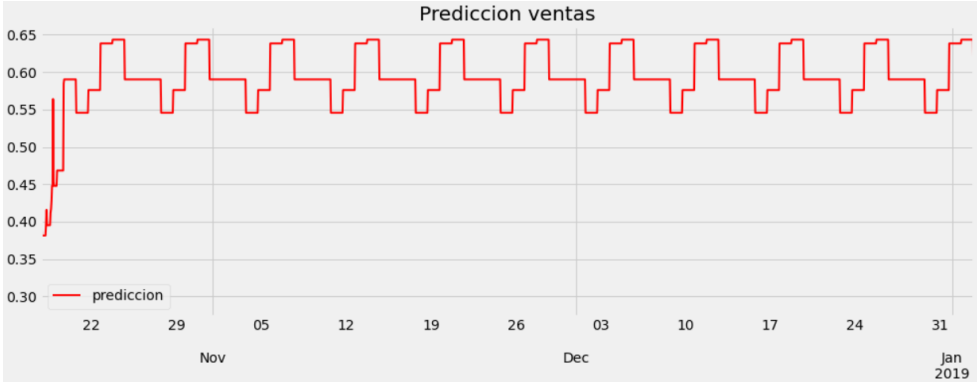


Predicción

La proyección entre mediados de octubre y principios de enero del 2019



La proyección entre mediados de octubre y principios de enero del 2019



En base a los resultados otorgados por el modelo llegamos a la conclusión de que hubo una baja en las ventas durante septiembre y continúa hasta principios de enero del año siguiente. Con una mayor ingesta de datos se podría predecir más tiempo, y se puede elegir el punto de partida donde se quiere predecir indicando cual es la ventana de 48hs previa al período donde quiere la predicción.

Conclusión

Este proyecto permitió desafiar nuestros conocimientos, soft skills y poner a prueba lo aprendido a lo largo del bootcamp.
Logramos mejorar día tras día y afianzar el trabajo en equipo, culminando de forma óptima los objetivos propuestos desde un principio.

Contacto

https://github.com/ronalcabrera/PG_Olist.git

LinkedIn:

https://www.linkedin.com/in/ronal-cabrera

https://www.linkedin.com/in/juan-valentin-fogliatti-06505852

https://www.linkedin.com/in/matias-martinez96

https://www.linkedin.com/in/macielaortiz