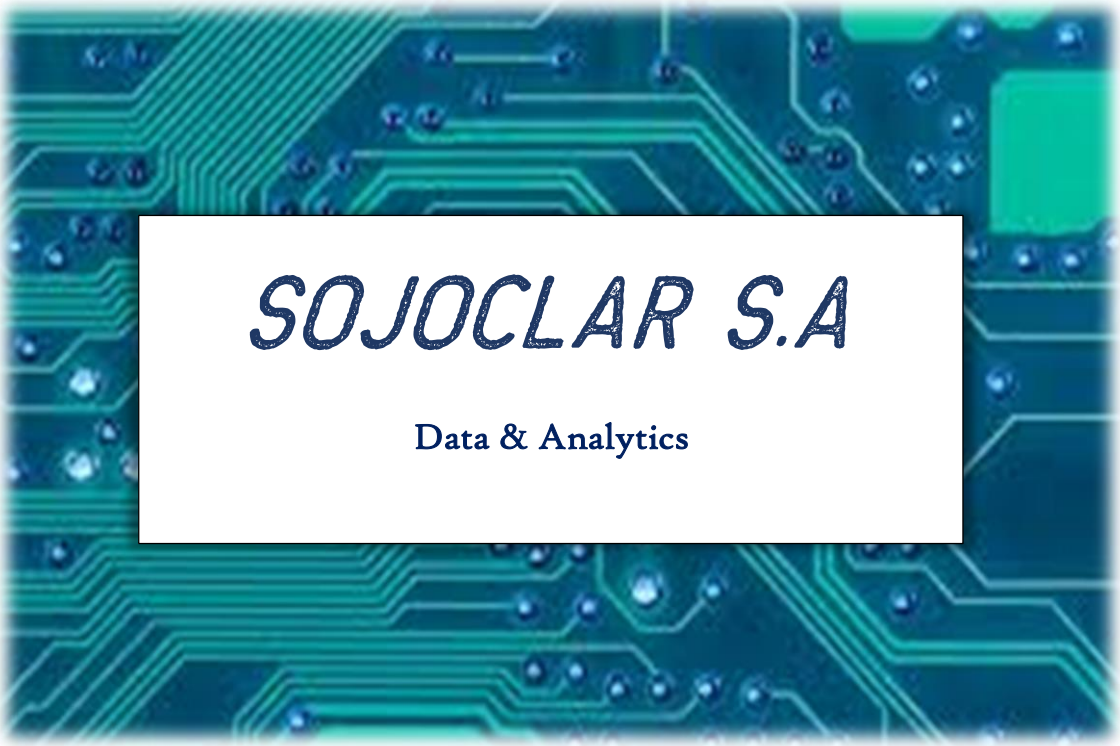


Proyecto grupal - Grupo 7

Carballo C. - Cuscueta S. – Deiloff J. - Gonzalez A

Contenido

Introducción:

En base a la necesidad de nuestro cliente, de tomar una decisión acerca de la posibilidad de expandir su negocio (plataforma de compra-ventas vía web) hacia Brasil, nuestra empresa diseñó un esquema de trabajo, que nos permitirá realizar una exhaustiva exploración del mercado de E-Commerce brasileño y a partir de la información obtenida, lograr desarrollar estrategias orientadas a facilitar la inteligencia del negocio. Para llevar adelante estas operaciones, nos basamos en los datos proporcionados por una fuente de acceso libre, denominada “Olist”.

Conformación del equipo: Roles y funciones

Para elevar la eficiencia de este proyecto, las tareas serán divididas entre los integrantes del equipo de trabajo, teniendo en cuenta las habilidades y preferencias de cada miembro del mismo. Dependiendo de la complejidad de los procesos, se otorgará a los integrantes, la libertad de sumarse a otras divisiones, optimizando, de este modo, los tiempos de ejecución de las tareas asignadas. A su vez, los avances serán auditados por todos los integrantes del equipo de trabajo.

División de Ingeniería de datos: esta área tendrá como función, encargarse de todo el proceso de ETL, entendiéndose al mismo, como la extracción de los datos, su transformación (limpieza y normalización) y su posterior carga en una base de datos. Staff: Jonathan Deiloff; Ariel González. Auxiliares: Claudio Carballo; Sofía Cuscueta.

División de Análisis de datos: las funciones de este sector se centrarán en el análisis de los datos, generación y visualización de reportes. Preparación de demos. Staff: Claudio Carballo; Sofía Cuscueta. Auxiliares: Jonathan Deiloff; Ariel González.

División de Machine Learning: de esta área, dependerá la construcción de modelos de aprendizaje automático para la posterior generación de predicciones en función de los objetivos propuestos. Staff: Jonathan Deiloff; Ariel González; Claudio Carballo; Sofía Cuscueta.

Metodología de trabajo

El proyecto se desarrollará utilizando metodología Scrum. Cada día se realizará una reunión de equipo, de unos quince minutos de duración, donde se debatirá la necesidad o no, de modificar algún detalle del proyecto y se hará un seguimiento del cumplimiento de los objetivos propuestos para el día anterior y el corriente. Así también, al finalizar cada semana, se dispondrá de un espacio para revisar de manera un poco más global, los avances del proyecto efectuados durante la totalidad de la misma. Además de los encuentros diarios y semanales, se fijarán reuniones con el cliente, de modo que el mismo, pueda conocer los avances del trabajo y realizar un feedback de lo propuesto. Si bien se confeccionará un cronograma de tareas, para proporcionar al equipo lineamientos y objetivos diarios, estos sin embargo, estarán siempre abiertos a adaptación, en función de los cambios que fueran surgiendo a lo largo del proceso, ya sea por nuevas necesidades, o ideas que nacieran en el transcurso, tanto de parte del equipo como del cliente.

Objetivos

Objetivo principal: “Evaluar la viabilidad y beneficios de abrir una empresa de E-Commerce en Brasil”.

Objetivos específicos:

- Generar datos con alto grado de calidad e integridad
- Confeccionar un Data Lake y Data Warehouse eficiente y escalable
- Analizar la información recabada y proponer KPIs
- Predecir sucesos beneficiosos para la empresa
- Confeccionar un reporte productivo y visualizable

Tareas a realizar en función de cada objetivo

Objetivo 1: “Generar datos con alto grado de calidad e integridad”

- Carga de los dataset a un editor de código.
- Exploración de los dataset, buscando datos nulos, errores, valores outliers, denominaciones no estandarizadas, etc.
- Construcción de algoritmos para automatizar el proceso de transformación de los datos.
- Transformación de los datos: limpieza y normalización.

Objetivo 2: “Confeccionar un Data Lake y Data Warehouse eficiente y escalable”

- Elaboración de un Data Lake, Data Warehouse y una base de datos desde pipelines programables, eficientes y escalables.

Objetivo 3: “Analizar la información recabada y proponer KPIs”

- Análisis detallado de los datos, generando información que permita la formulación de recomendaciones y estrategias de negocios al cliente.
- Propuesta de KPIs y métricas.

Objetivo 4: “Predecir sucesos beneficiosos para la empresa”

- Evaluación de los datos y decisión de un target para aplicar herramientas de aprendizaje automático.
- Elección de algoritmos de machine learning apropiados.
- Implementación del/los modelos.
- Análisis de los resultados.

Objetivo 5: “Confeccionar un reporte productivo y visualizable”

- Elección e implementación de representaciones gráficas de la información recabada.
- Generación de un dashboard estético, que permita visualizar de manera sencilla y atractiva los resultados del proyecto.
- Elaboración de un storytelling.

Alcances del proyecto:

La misión de este proyecto será ofrecer al cliente, información de calidad, que le permita tener argumentos sólidos, para tomar decisiones en base al objetivo de negocios de su empresa. Dicha información será obtenida por el equipo, tras aplicar procesos de ingeniería y análisis de datos.

Riesgos:

Un riesgo con el cual podríamos encontrarnos, es el extraer conclusiones erróneas por un mal entendimiento de la información con la que contamos. Para disminuir la probabilidad de ocurrencia de este hecho, decidimos evitar el uso de información que no contribuya a la performance del trabajo, por ejemplo, datasets o sus fragmentos, que consideremos, no hacen ningún aporte constructivo al proyecto, ya sea por su irrelevancia o por la cantidad de información confusa, incompleta y/o errónea que contengan.

Stack tecnológico propuesto

- Trabajo diario: Google Docs; GitHub; Trello; Google Meet
- Ingeniería de datos: Python; MySQL
- Análisis y Visualización de datos: Python; Power BI.
- Machine Learning: Python

Cronograma de trabajo

SEMANA 1

En la primera semana, el énfasis estará puesto en el diseño y la planificación del trabajo y en la exploración de los datos. Todas las divisiones trabajarán a la par.

Lunes: selección de equipos de trabajo

Martes:

➔ Definición de roles y funciones de los miembros del equipo

➔ Planificación de la ejecución del proyecto

➔ Formulación de un cronograma de trabajo

➔ Inicio de la redacción de un informe escrito

Miércoles:

➔ Análisis de los Datasets

➔ Definición de posibles KPIs a analizar

➔ Continuación de la redacción de un informe escrito

Jueves:

➔ Elaboración de informe preliminar de calidad del dato

➔ Continuación de la redacción de un informe escrito

Viernes:

➔ Revisión y pulido del material entregable

➔ Presentación de avances (entrega de informes)

SEMANA 2

Lunes:

➔ Carga inicial y limpieza de los datasets

➔ Automatización del proceso

➔ Documentación del procedimiento

Martes:

➔ Limpieza y normalización de los datasets

➔ Documentación del procedimiento

Miércoles:

➔ Etapa de carga en una base de datos

➔ Documentación del procedimiento

Jueves:

➔ Construcción de un modelo entidad - relación

➔ Redacción de un informe escrito

Viernes:

➔ Revisión y pulido del material entregable

➔ Presentación de avances

SEMANA 3

Lunes:

➔ Automatización de carga incremental de datos.

➔ Querys para construcción de reportes

➔ Documentación del procedimiento

Martes:

➔ Querys para construcción de reportes

➔ Documentación del procedimiento

Miércoles:

➔ Visualización de resultados

➔ Construcción de un dashboard

➔ Documentación del procedimiento

Jueves:

➔ Construcción y ejecución de algoritmos predictivos

➔ Visualización de los resultados

➔ Redacción de un informe escrito

Viernes:

➔ Revisión y pulido del material entregable

➔ Presentación de avances

SEMANA 4

Lunes:

➔ Planificación del storytelling

➔ Revisión y pulido del informe final

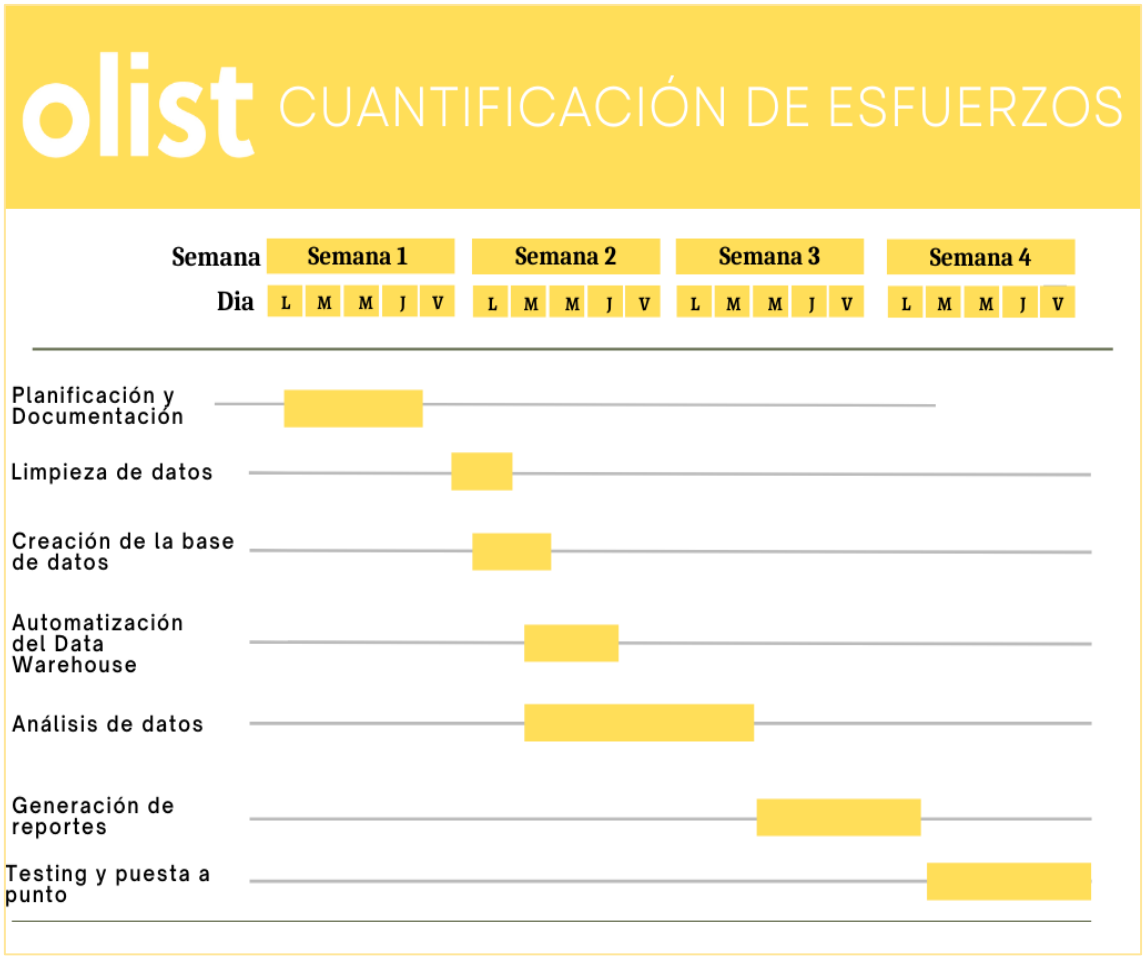
Martes:

➔ Preparación de la demo final

Miércoles:

➔ Presentación del proyecto

Estimación de esfuerzos:



Material entregable

Cada semana se entregará al cliente un informe de los avances del proyecto

Entregable Semana 1

- Informe donde conste la estructura y metodología de trabajo del equipo, los objetivos propuestos para encarar el proyecto, la diagramación de un cronograma de actividades y la estimación de esfuerzos (diagrama de Gantt.

- Extra: Informe preliminar de calidad de los datos

Entregable Semana 2

- Informe de las tareas realizadas hasta la fecha (proceso de ETL)

Entregable Semana 3

- Informe de los avances realizados hasta la fecha (análisis de la información, predicciones, construcción de reporte gráfico)

Entregable Semana 4

- Informe final completo

- Presentación audiovisual

Desarrollo del proyecto

Informe preliminar de calidad del dato

En una primera instancia, se realizó la carga a un editor de código, de un conjunto de datos, pertenecientes a un dataset de acceso público, denominado “Olist: E-Commerce Public dataset”. El mismo estaba conformado por 11 tablas.

Luego, se exploró cada una de las tablas y se confeccionó un diccionario para comprender el contenido de los datos allí alojados.

Diccionarios de datos:

olist_customers_dataset

columna	Descripción	Tipo de dato	Ejemplo
customer_id	Nº ID cliente, key para el dataset de orders	Texto y número	06b8999e2fba1a1fbc88172c00ba8bc7
customer_unique_id	ID único de cliente	Texto y número	861eff4711a542e4b93843c6dd7febb0
customer_zip_code_prefix	primeros 5 dígitos del código postal	Número entero	14409
customer_city	ciudad donde reside el cliente	texto	Franca
customer_state	estado donde reside el cliente	texto	SP

olist_geolocation_dataset

Columna	Descripción	Tipo de dato	Ejemplo
geolocation_zip_code_prefix	primeros 5 digitos del codigo postal	Número entero	1037
geolocation_lat	coordenadas de latitud	Numero flotante	-23.545621
geolocation_lng	coordenadas de longitud	Numero flotante	-46.639292
geolocation_city	nombre de la ciudad	Texto	sao paulo
geolocation_state	nombre del estado	Texto	SP

olist_order_items_dataset

Columna	Descripcion	Tipo de dato	Ejemplo
order_id	ID unico de orden de compra	Numero y texto	00010242fe8c5a6d1ba2dd792cb16214
order_item_id	ID secuencial de identificacion de productos incluidos en la misma orden	Numero	1
product_id	ID del producto	Numero y texto	4244733e06e7ecb4970a6e2683c13e61
seller_id	ID del vendedor	Numero y texto	48436dade18ac8b2bce089ec2a041202
shipping_limit_date	Fecha limite de entrega del producto	Fecha	2017-09-19 09:45:35
price	Precio final del producto	Numero flotante	58.90
freight_value	Precio de la logistica por la entrega del producto (si una orden tuvo mas de un producto, el costo se divide por la cantidad de productos)	Numero flotante	13.29

olist_order_payments_dataset

Columna	Descripcion	Tipo de dato	Ejemplo
order_id	ID unico de orden de compra	Numero y texto	b81ef226f3fe1789b1e8b2acac839d17
payment_sequential	Si el cliente uso mas de un medio de pago, se genera una secuencia para unificar los pagos	Numero entero	1
payment_type	Medio de pago utilizado	Texto	credit_card
payment_installments	Numero de cuotas en las que se dividio el pago	Numero entero	8
payment_value	Valor de la transacción realizada	Numero flotante	99.33

olist_order_reviews_dataset

Columna	Descripcion	Tipo de dato	Ejemplo
review_id	ID unico de la reseña	Texto y numero	bc2406110b926393aa56f80a40eba40
order_id	ID de la orden asociada	Texto y numero	73fc7af87114b39712e6da79b0a377eb
review_score	Puntaje de la reseña	Numero Entero	4
review_comment_title	Titulo de la reseña en portugues	Texto	'10'
review_comment_message	Mensaje de la reseña en portugues	Texto	Recebi bem antes do prazo estipulado.
review_creation_date	Fecha de envio del pedido de reseña al cliente	Fecha	2017-04-21 00:00:00
review_answer_timestamp	Fecha de respuesta de la reseña del cliente	Fecha	2017-04-21 22:02:06

olist_orders_dataset

Columna	Descripcion	Tipo de dato	Ejemplo
order_id	ID único del pedido	Texto y numero	e481f51cbdc54678b7cc49136f2d6af7
customer_id	ID del cliente, key para el dataset de customer	Texto y numero	9ef432eb6251297304e76186b10a928d
order_status	estado del pedido	Texto	delivered
order_purchase_timestamp	fecha de la compra	Fecha	2017-10-02 10:56:33
order_approved_at	fecha de aprobación de la compra	Fecha	2017-10-02 11:07:15
order_delivered_carrier_date	fecha de publicación del pedido. Cuando se entregó al socio logístico.	Fecha	2017-10-04 19:55:00
order_delivered_customer_date	fecha real de entrega del pedido al cliente.	Fecha	2017-10-10 21:25:13
order_estimated_delivery_date	fecha estimada de entrega que fue informada al cliente en el momento de la compra	Fecha	2017-10-18 00:00:00

olist_products_dataset

Columna	Descripción	Tipo de dato	Ejemplo
product_id	ID del producto	Texto y numero	1e9e8ef04dbcff4541ed26657ea517e5
product_category_name	categoria del producto, en portugués	Texto	perfumaria
product_name_lenght	número de caracteres extraídos del nombre del producto	Numero Flotante	40.0
product_description_lenght	numero de caracteres extraídos de la descripción del producto	Numero Flotante	287.0
product_photos_qty	números de fotos publicadas del producto	Numero Flotante	1.0
product_weight_g	peso del producto medido en gramos	Numero Flotante	225.0
product_length_cm	longitud del producto medida en centímetros	Numero Flotante	16.0
product_height_cm	altura del producto medida en centímetros	Numero Flotante	10.0
product_width_cm	ancho del producto medido en centímetros	Numero Flotante	14.0

olist_sellers_dataset

Columna	Descripción	Tipo de dato	Ejemplo
seller_id	ID del vendedor	Texto y numero	3442f8959a84dea7ee197c632cb2df15
seller_zip_code_prefix	primeros 5 dígitos del código postal del vendedor	Numero Entero	13023
seller_city	nombre de la ciudad del vendedor	Texto	campinas
seller_state	estado(provincia) del vendedor	Texto	SP

product_category_name_traslation

Columna	Descripción	Tipo de texto	Ejemplo
product_category_name	nombre del producto	Texto	beleza_saude
product_category_nam	nombre del producto	Texto	health_beauty

olist_marketing_qualified_lead_dataset

Columna	Descripcion	Tipo de dato	Ejemplo
mql_id	ID del Marketing Qualified Leads	Texto	dac32acd4db4c29c230538b72f8dd87d
first_contact_date	primera fecha de contacto	Fecha	2018-02-01
landing_page_id	Informacion insuficiente	Texto y numero	88740e65d5d6b056e0cda098e1ea6313
origin	Origen de la búsqueda, como llegaron al formulario de contacto	Texto	social

olist_closed_deals_dataset

Columna	Descripcion	Tipo de dato	Ejemplo
mql_id	id del Marketing Qualified Leads	Texto y numero	5420aad7fec3549a85876ba1c529bd84
seller_id	ID del vendedor	Texto y numero	2c43fb513632d29b3b58df74816f1b06
sdr_id	Informacion insuficiente	Texto y numero	a8387c01a09e99ce014107505b92388c
sr_id	Informacion insuficiente	Texto y numero	4ef15afb4b2723d8f3d81e51ec7afefe
won_date	Informacion insuficiente	Feha	2018-02-26 19:58:54
business_segment	Segmento del negocio	Texto	pet
lead_type	Detalle del vendedor	Texto	online_medium
lead_behaviour_profile	Informacion insuficiente	Texto	cat
has_company	Informacion insuficiente	Texto	True
has_gtin	Informacion insuficiente	Texto	True
average_stock	Tipo de negocio	Texto	20-50
business_type	Informacion insuficiente	Texto	reseller
declared_product_catalog_size	Informacion insuficiente	Numero flotante	132.0
declared_monthly_revenue	Informacion insuficiente	Numero flotante	0.0

El siguiente paso, consistió en realizar consultas e iteraciones en las tablas, para detectar valores faltantes, aparición de valores outliers, errores de denominación de columnas, falta de normalización en la escritura de los datos, entre otras variables. Como resultado de este proceso, se hallaron los siguientes puntos, los cuales, creemos, pueden volverse un obstáculo para lograr una calidad de datos aceptable, motivo suficiente para proceder a su depuración, como próximo paso de este proyecto.

Errores encontrados:

- **olist_customers_dataset**

Esta tabla no contiene errores, solo el idioma de los campos podría modificarse, traduciéndose del inglés al español.

- **olist_geolocation_dataset**

Esta tabla, si bien, no presentó valores nulos, mostró registros duplicados, hallándose un total de 261.831 valores duplicados. Al igual que las demás tablas, esta se encuentra en idioma inglés.

- **olist_order_items_dataset**

Tipo de dato erróneo en el campo “shipping_limit_date”, el dato que debería ser de tipo fecha se presenta como tipo objeto. Idioma inglés.

- **olist_order_payments_dataset**

Tabla sin errores, a excepción del idioma.

- **olist_order_reviews_dataset**

Se detectaron en esta tabla un total de 145903 valores faltantes (21% del total de datos), los cuales pertenecen sólo a 2 de los 7 campos.

Distribución que genera un aumento en la significancia de estos faltantes, ya que, al analizarlos en función del total de valores por categorías, el porcentaje se vuelve muy elevado. En la categoría “review_comment_title” los valores nulos (87656) representan el 88.34% de los valores totales para dicho campo y en la categoría “review_comment_message” (58247) un 58.70%. Idioma inglés.

- **olist_orders_dataset**

No se hallaron valores duplicados, ni problemas de entendimiento en las denominaciones de los campos de esta tabla. Sólo debería cambiarse el tipo de dato en los campos con fechas, porque se encuentran como dato objeto. Sin embargo, el inconveniente principal, fueron los registros nulos, detectándose un total de 4908 valores faltantes, distribuidos en los campos “order_approved_at”(160); “order_delivered_carrier_date” (1783) y “order_delivered_customer_date” (2965). Estos nulos, sin embargo, representan sólo un 0.62% del total de datos (795528 registros totales). Al igual que en todas las tablas, los nombres de los campos están en inglés.

- **olist_products_dataset**

De igual forma a los Datasets ya analizados, además del idioma de los campos, los errores hallados solo fueron valores nulos. De un total de 287559 registros, 2448 fueron nulos (0.85%). Estos valores faltantes, se observaron, sobre todo, en las columnas “product_category_name”; “product_name_lenght”; “product_description_lenght” y “product_photos_qty”.

- **olist_sellers_dataset**

Esta tabla no presentó errores, sólo podrían modificarse los nombres de los campos, traduciéndose al español.

- **product_category_name_translation**

Tabla sin errores, sólo es necesario cambiar el idioma de las columnas.

- **olist_marketing_qualified_leads_dataset**

Esta tabla contiene solo 4 columnas de datos, con un total de 8000 valores en cada una de ellas (32000 registros totales). Entre los errores, sólo se detectaron 60 valores nulos en el campo “origin”. Valores duplicados no se hallaron. Los nombres de los campos se encuentran en inglés.

- **olist_closed_deals_dataset**

Esta tabla consta de 14 columnas de datos, con un total de 842 valores en cada una de ellas (11788 registros totales). Se detectaron 3299 valores nulos (27.99% del total de datos) distribuidos en 7 campos, siendo las categorías más afectadas “has_gtin”; “has_company”; “average_stock” y “declared_product_catalog_size”. No se encontraron registros duplicados, ni valores outliers. La mayor dificultad de la tabla es el entendimiento de la información de varios campos, como “sdr_id”; “sr_id”; “lead_behaviour_profile”; “has_company”, entre otros. A su vez, los nombres de los campos se encuentran en inglés.

Extracción, transformación y carga de los datos

Tras el análisis preliminar de los datos, continuamos con el siguiente paso del proceso, que según se había explicado, sería la transformación y carga de los datos.

Dada la reciente adquisición de servicios Cloud, por parte de la empresa, decidimos sumar estos recursos a nuestro trabajo, utilizando Amazon S3, para almacenamiento y backups de archivos y Amazon RDS como hospedador de la base de datos en la nube. Además de esta

incorporación, añadimos a nuestro stack tecnológico, el uso de Google Colaboratory, para facilitar el trabajo colaborativo y simultáneo.

La limpieza y normalización de los datos -con la finalidad de lograr una automatización del proceso- fue llevada a cabo en Python utilizando Pandas y Boto3. La base de datos MySQL fue generada y modificada mediante el uso de SQL Alchemy en Python y hospedada en una instancia de Amazon RDS. La administración de la base de datos, por su parte, se realizó mediante MySQL Workbench y DBeaver, dependiendo de las preferencias de cada miembro del equipo de trabajo.

Limpieza y normalización de los datos

Limpieza de registros duplicados y nulos.

Algunos registros nulos de poca relevancia analítica, fueron eliminados, mientras que otros, fueron reemplazados, por ejemplo, en la tabla “order_reviews”, los nulos de la columna “comment_message”, fueron cambiados por la palabra “nao”.

Los registros duplicados, fueron eliminados de las tablas:

- “order_items”: columna “order_id”
- “order_reviews”: columna “review_id”
- “geolocation”: columna “zip_code_prefix”.

Descarte de columnas y registros innecesarios.

- customer_unique_id
- payment_sequential

Normalización de nombres de columnas.

En la tabla “geolocation”, en la columna “city”, para evitar errores de escritura, se cambió el caracter conflictivo “ã” por “a”.

Los nombres de las tablas y de sus columnas fueron dejados en inglés, ya que su conservación no nos implicaba ningún inconveniente, pero en su mayoría fueron acortados, eliminando el nombre de la tabla que aparecía como palabra inicial.

Denominación original	Nuevo nombre
customer_zip_code_prefix	zip_code_prefix
customer_city	city
customer_state	state
geolocation_zip_code_prefix	zip_code_prefix
geolocation_city	city
geolocation_state	state
geolocation_lat	latitude
geolocation_lng	longitude

order_item_id	product_quantity
review_comment_message	comment_message
review_creation_date	creation_date
review_comment_title	comment_title
review_answer_timestamp	answer_timestamp
order_purchase_timestamp	purchase_timestamp
order_delivered_carrier_date	delivered_carrier_date
order_delivered_customer_date	delivered_customer_date
order_estimated_delivery_date	estimated_delivery_date
product_description_lenght product_name_lenght	description_length name_length
product_weight_g	weight_g
product_length_cm	length_cm
product_height_cm	height_cm
product_width_cm	width_cm
product_photos_qty	photos_quantity
seller_zip_code_prefix	zip_code_prefix
seller_city	city
seller_state	state

Corrección de los tipos de datos de las columnas.

Se realizaron múltiples modificaciones en los tipos de datos, por ejemplo, se convirtieron en VARCHAR, datos que estaban en formato TEXTO, y en DATETIME fechas que estaban también como texto.

Pensando en las necesidades de información que tendremos en el siguiente paso del proceso, agregamos algunas columnas y segmentaciones de datos que podrían ser de gran ayuda:

- Creación de una nueva columna en español para las sucursales latinoamericanas.

- Creación de la columna 'region' en el dataframe 'geolocation' (agrupando los estados brasileños en 5 regiones)

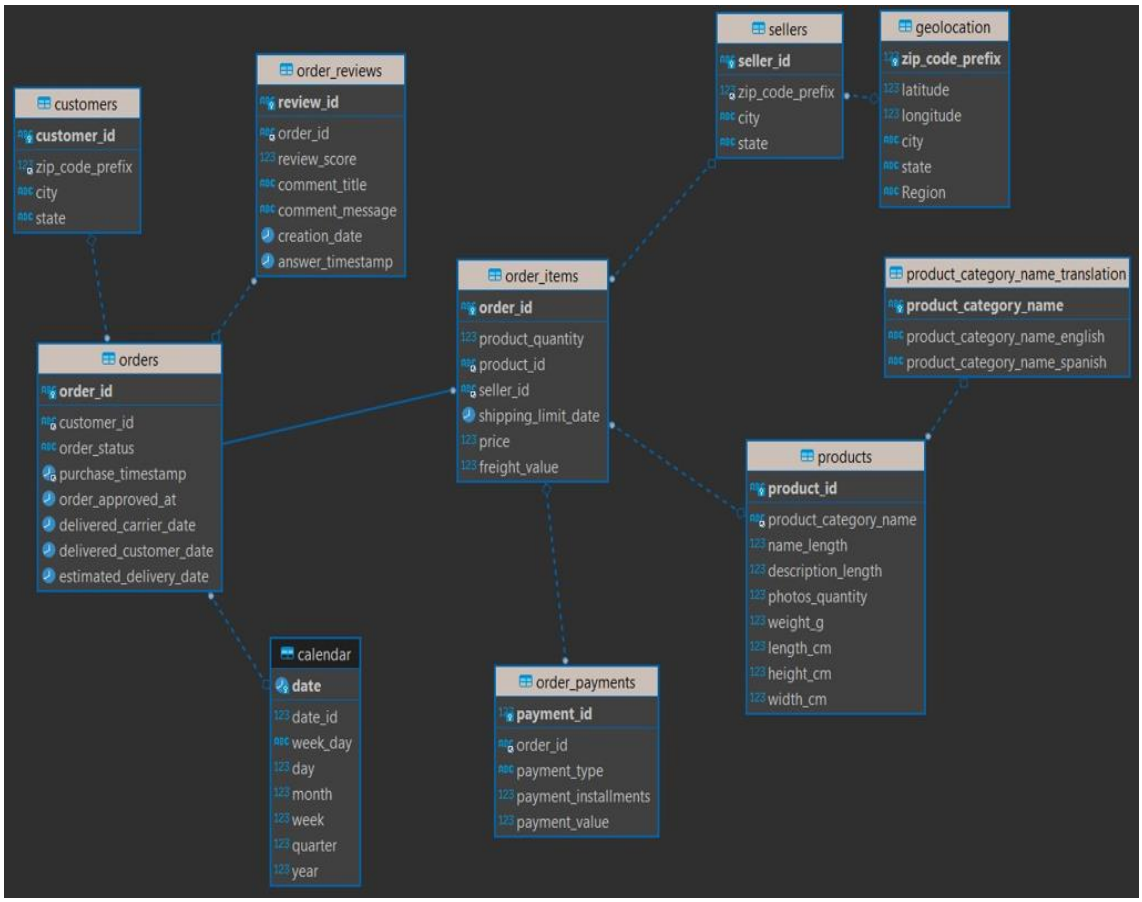
- Creación de una tabla Calendario

- Relacionamos los puntajes malos (menores a 3 puntos) dados por los clientes en las reviews, a problemas en la entrega de los productos. Para ello detectamos la palabra “entrega” en aquellos comentarios de clientes que habían dado puntajes malos a sus compras.

Finalmente, luego de las transformaciones realizadas sobre los datos, se crearon las llaves primarias y foráneas para generar un diagrama de entidad relación entre las tablas. Se determinó que en lugar de un “modelo estrella”, sería más apropiado un modelo “copo de nieve”.

Dado que de los 11 datasets originales, descartamos 2 por la mala calidad de sus datos (“olist_marketing_qualifield_lead_dataset” y “Olist_closed_deals_dataset”) y que añadimos una nueva tabla “Calendar”, el modelo presentado quedó conformado por 10 tablas. El mismo puede observarse en la imagen siguiente.

Modelo entidad–relación



Dado que, en el transcurso de la última semana, se realizaron algunos cambios, sobre todo en lo que respecta a la arquitectura de los datos, sintetizamos en el siguiente párrafo el proceso de ETL llevado a cabo hasta el momento:

Lectura de los archivos.csv pertenecientes al dataset Olist, (de acceso público en la web) mediante un editor de código, utilizando el lenguaje de programación Python. Normalización y limpieza de los datos usando la librería pandas. Pasaje de los dataframes obtenidos a una base de datos MySQL, mediante SQLalchemy trabajando “on cloud” por medio del

servicio Amazon RDS. Creación de llaves primarias y foráneas para la generación de un diagrama de relación-entidad.

Paralelamente se llevó a cabo, la unión de la base de datos a un bucket en Amazon S3 (Data Lake), el cual, a su vez, está conectado a un clúster de AWS, para procesar todos los datos cargados con sus relaciones en Delta Lake (Data Warehouse) donde se generan tablas optimizadas para análisis o Machine Learning.

Todo se trabajó en la plataforma Databricks, conectada a los servicios del ecosistema de AWS, para abastecerse de instancias físicas de procesamiento y almacenamiento. Esto permitió administrar mejor los clústeres, recursos de almacenamiento y nodos que ofrece AWS, corriendo nativamente en Spark. La plataforma en cuestión, para poder paralelizar los procesos, presenta su propia distribución de archivos similar a Hadoop, el DBFS, optimizada para Spark. A su vez, los notebooks de Databricks son similares a los Jupyter pero capaces de correr distintos tipos de lenguajes de programación en el mismo notebook (Python, Scala, SQL, R), favoreciendo el trabajo colaborativo y mejorando el entorno de trabajo integrándolo todo en la misma plataforma web. Su servicio Delta Lake ofrece una integración de Data Warehouse con Data Lake que permite el análisis de datos y la generación de dashboards interactivos.

Carga incremental de datos

Luego del feedback recibido por parte del cliente, de decidió añadir la posibilidad de que el proceso de ETL pudiera implementarse de manera automática, ante el ingreso de actualizaciones en los datos (deltas).

Para poder realizar pruebas y comprobar el buen funcionamiento del esquema propuesto, generamos un dataset sintético, logrando simular el ingreso de nuevos datos. Esta operación se desarrolló, mediante el uso de la librería sdv, la cual provee un modelo de aprendizaje automático, denominado “model GaussianCopula”, el mismo, fue entrenado con el dataset de Olist para que generara un nuevo conjunto de datos.

Tras implementar estos últimos pasos, pudimos dar por finalizado el pipeline. Obteniendo un modelo que, una vez que detecta nuevos archivos (deltas) en la carpeta del bucket de Amazon S3, inicia el proceso de ETL y actualización de la base de datos original con los nuevos datos. La base de datos antes de la actualización, genera una copia de seguridad que es almacenada al igual que los deltas, en el Data Lake.

Análisis de los datos

Análisis coyuntural económico y financiero de Brasil

Antes de proceder a evaluar la información que hemos procesado, decidimos ofrecer también, un breve estudio de la situación actual del mercado económico y financiero de Brasil. Dicho país, se presenta actualmente, como la octava economía del mundo y particularmente, la Bolsa de San Pablo representa la séptima bolsa de valores más grande e importante del planeta. Estos factores, son suficientes para dejar en claro que su coyuntura económica y financiera es un factor que puede condicionar la performance de cualquier empresa en su territorio, tanto en el corto como en el largo plazo.

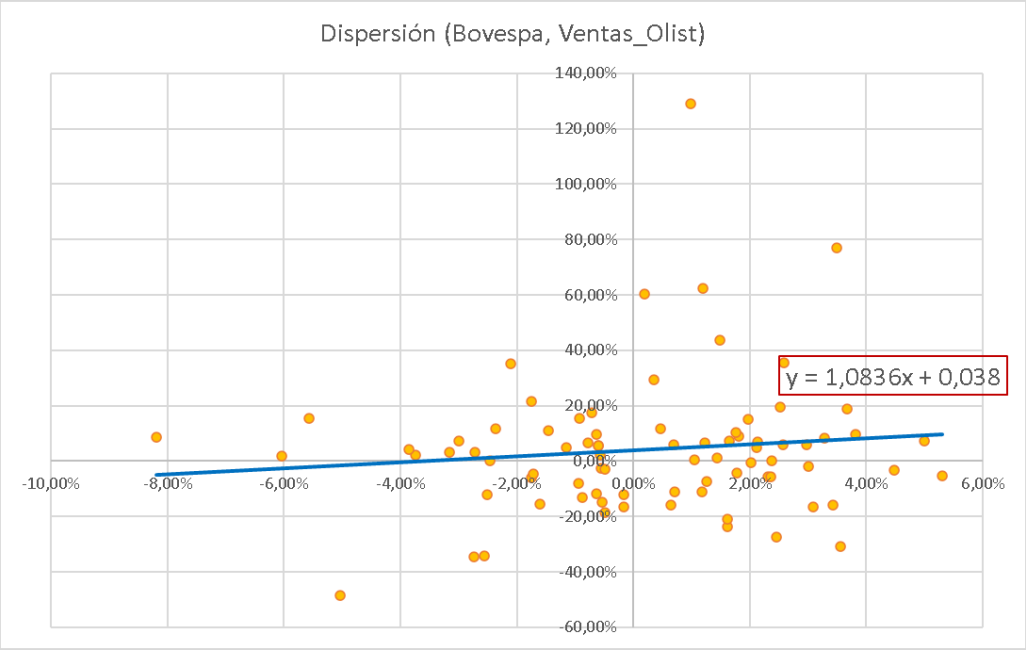
Según sondeos elaborados por Reuters (lo cuales se centran en 17 grandes índices) entre el 12 y el 24 de mayo del corriente año, se observó que la mayoría de las principales bolsas, tendrían dificultades para recuperar las pérdidas del año en curso a fines de 2022. Esperándose que casi todos terminen el año por debajo de sus máximos históricos y que sigan por debajo de ellos a mediados de 2023. La realidad brasileña no escapa a estos hechos, sumándose a ello, que este año habrá elecciones presidenciales en dicho país, por lo que se espera que el índice bursátil Bovespa suba menos de lo previsto, más aún teniendo en cuenta que en lo que va del año cayó un 20%.

En este sentido, entendemos que el riesgo sistemático o de mercado es un riesgo que no se puede eliminar o reducir, pero sí, es posible medirlo y

tratar de anticipar su impacto. Para eso usaremos el famoso coeficiente Beta. Este representa la sensibilidad de los cambios en los rendimientos de una acción (en nuestro caso al ser Olist una empresa no pública tomaremos como referencia sus ventas) con respecto a los cambios en el rendimiento en el mercado, cuando los rendimientos de una acción varían de forma similar a los rendimientos del mercado su Beta será similar a uno (1), cuando la Beta es mayor a uno (1), entonces el rendimiento de la empresa/acción es más volátil que el rendimiento del mercado en su conjunto

Estimación del Beta

Hay dos técnicas que son las más usadas para determinar el Beta, la primera, consiste en realizar una regresión lineal, ajustando una línea recta a los puntos dispersos, que en este caso serán los rendimientos de Olist y del índice bursátil Bovespa. Los datos que tomamos para analizar van desde el 1 de diciembre 2017 hasta el 1 de agosto de 2018, de estos extrajimos los rendimientos semanales de ambas variables. En el caso de Bovespa exportamos sus rendimientos de Yahoo Finance y en el caso de Olist tomamos los rendimientos de sus ventas semanales.



Con este método tomamos la pendiente de la recta, en este caso observamos que el Beta es de 1,08.

El segundo método lo realizamos usando una fórmula que contempla la relación entre la varianza del mercado (en este caso Bovespa) y la covarianza entre los rendimientos de Olist y el índice bursátil.

$$\beta = \frac{Covarianza_{Empresa\ ,Mercado}}{Varianza_{Mercado}}$$

Covarianza (empresa, mercado) = 0.00072

Varianza(mercado) = 0.00076

Beta = 1.09

Conclusión

Como puede verse, el Beta que obtuvimos de la empresa Olist (1.09) es afín al de aquellas compañías públicas de similares características. Por lo que su cálculo está dentro de parámetros racionales.

International Peers - MercadoLibre Inc.

Company Name	Ctry	Market Cap. last (mUSD)	Beta 1-Year	Year-To-Date Price Change (in local currency)
MercadoLibre Inc.	ARG	35 996	N/A	-46.0%
	International Peers Median		*1.22*	*-45.0%*
Vipshop Holdings Ltd.	CYM	5 699	1.27	13.8%
Rakuten Group, Inc.	JPN	7 341	1.16	-45.0%
Sea Limited	CYM	37 373	3.21	-67.9%
Global-e Online Ltd.	ISR	3 320	N/A	-66.1%
eBay Inc.	USA	25 902	1.03	-31.3%

Fuente: https://www.infrontanalytics.com/fe-ES/30245LA/MercadoLibre-Inc-/beta

Creemos que la expansión de la empresa a Brasil estaría acompañada por una leve reactivación postpandemia, en donde el mercado estaría especulando con las elecciones presidenciales del presente año. El Beta de la empresa Olist demuestra que ante un repunte de la economía

brasileña estaríamos frente a un escenario más que propicio para una expansión empresarial, por otro lado, y frente a un riesgo sistemático su impacto también será mayor en la empresa analizada.

Análisis y reportes

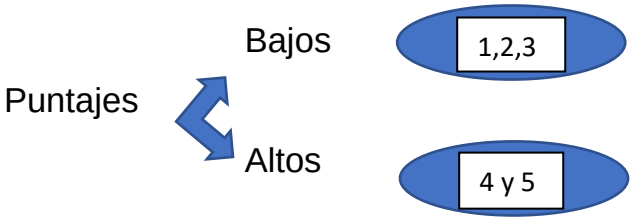
Regresando el foco a la información procesada por nuestra empresa, tras trabajar sobre el dataset Olist y luego de asegurarnos de contar con datos de calidad, comenzamos con el análisis de los mismos.

Relación entre puntuaciones negativas y las entregas de productos

Observando la información de las tablas, donde se encontraban las valoraciones de los clientes a las compras realizadas en la empresa de E-Commerce analizada (puntajes y comentarios), dilucidamos uno de los hallazgos más relevantes de este trabajo, el mismo consistió, en que parte de los puntajes bajos, dados por los clientes a su experiencia de compra, sucedieron (según lo detectado en sus comentarios) debido a problemas en la entrega de los productos comprados.

Cabe aclarar, que no todos los clientes que realizaron puntuaciones, dejaron comentarios escritos de su experiencia, motivo por el cual, no fue posible determinar la causa de los malos puntajes en estos casos.

Para llegar a la conclusión antes descripta, en primer lugar, se realizó una clasificación de los puntajes entendiendo como “puntajes bajos” o negativos, a aquellos que se encontraran entre 1 y 3 puntos de valoración de compra (review_score) y “puntajes altos” a los mayores a 3. Una vez realizada esta segmentación, se rastreó la existencia de la palabra “entrega” en los comentarios adjuntos de los clientes (review_comment_message), entendiendo que, si dicha palabra era detectada en un comentario relacionado a un puntaje negativo, el cliente no estaba conforme con la entrega del producto.



Así se obtuvo que, del total de compras registradas (99684), 14396 fueron puntuadas negativamente (14.44%), atribuyéndose esta valoración a problemas de entrega del producto, en un total de 694 casos (4.82%).

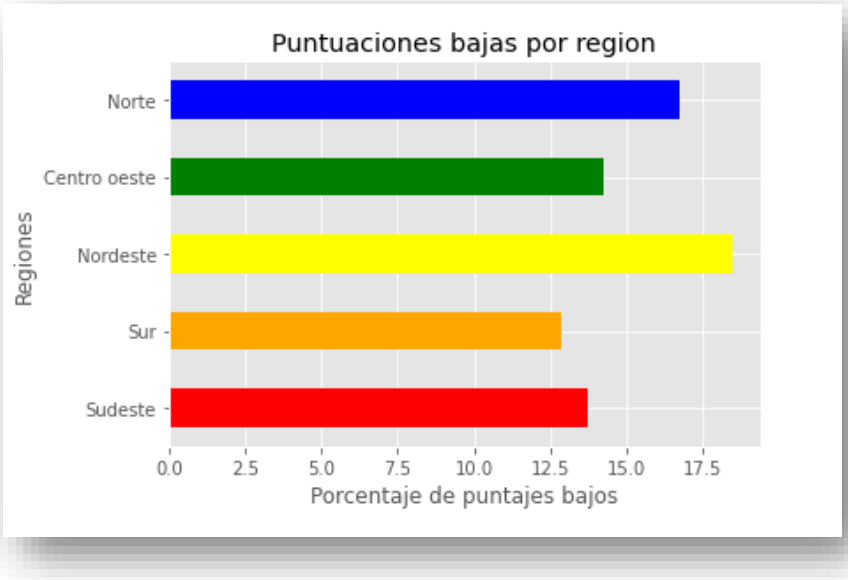
Los puntajes negativos también fueron relacionados a las categorías de productos. Siendo las categorías peores puntuadas “cama-mesa-banho”; “beleza-saude”; “informática-accesorios”; “moveis-decoracao” y “Esporte-lazer”.

Categoría de productos	Total de valoraciones negativas
Cama-mesa-banho	1514
Beleza-saude	1101
Informática-accesorios	1046
Moveis-decoracao	1004
Esporte-lazer	976

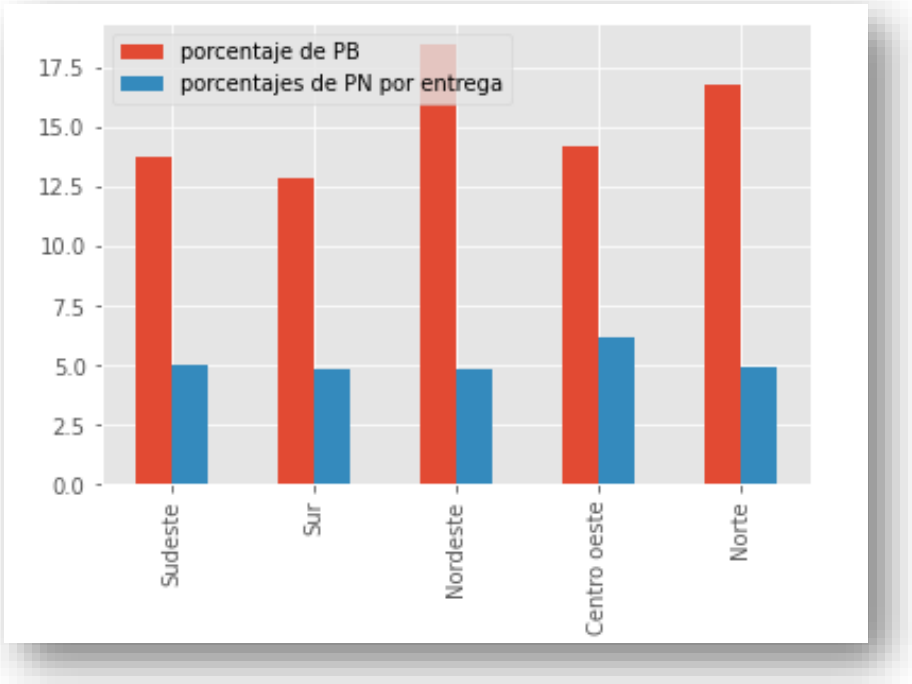
También se determinaron las regiones del país de donde provinieron la mayor cantidad de puntuaciones bajas, hallándose a la región Sudeste, seguida de la región Sur, como los lugares de procedencia de las compras peores puntuadas en términos absolutos, y a las regiones Nordeste y Norte en términos porcentuales.

Para realizar estos hallazgos, fue necesario, agrupar los datos de estado (state) en las 5 regiones en que se divide Brasil (Sudeste, Sur, Nordeste, Centro oeste y Norte), lo cual fue realizado en los avances dados en el análisis preliminar de los datos.

Gráfica de las puntuaciones negativas según cada región



Gráfica comparativa de los puntajes negativos por entregas y los puntajes bajos en general según cada región.



Accediendo a través del siguiente código QR, puede observarse la procedencia geográfica de las puntuaciones tanto negativas como positivas dadas por los clientes.

Código QR

El mismo procedimiento realizado con las regiones, se llevó a cabo con las ciudades brasileñas, obteniéndose como cabecera (en valores absolutos) la ciudad de Sao Paulo, con un total de 1911 puntuaciones negativas sobre 15444 valoraciones totales, lo que representó un 12.37% del total para dicha ciudad, y realizando la misma evaluación, pero en términos porcentuales, se encontró a la ciudad de Río de Janeiro como el origen de la mayor cantidad de puntajes malos.

Ciudades	Valoraciones	Valoraciones negativas	Porcentaje
Sao Paulo	15444	1911	12.37
Rio de Janeiro	6853	1335	19.48
Belo Horizonte	2758	358	12.98
Brasilia	2125	303	14.25
Curitiba	1511	177	11.71

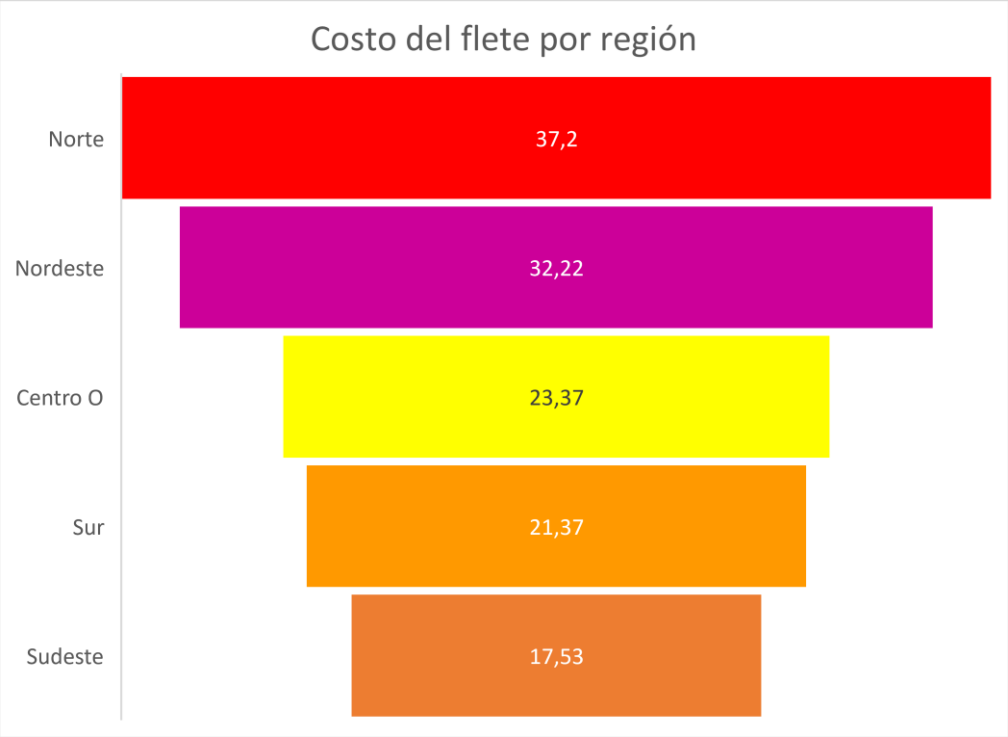
Demoras en los tiempos de entrega

En pos de averiguar si podíamos atribuir más casos de puntuaciones bajas a problemas en la entrega de los productos, investigamos la diferencia de tiempo, existente entre la fecha estimada de entrega del producto y la fecha en que este finalmente llegó al cliente objetivo. De este modo obtuvimos el promedio de diferencia de días, para cada región del país.

El área con mayor demora en las entregas fue la región Norte, la cual se encontraba en el segundo puesto en lo que respecta al origen de las puntuaciones negativas, por lo que inferimos que esta falla de logística podría ser una de las causas de las valoraciones bajas por parte de los clientes.

Costo del flete

Se evaluó el costo promedio del flete según la región del país de donde provenía la compra, hallándose a la región Norte como la de mayor costo de transporte y a la región Sudeste como la más barata.

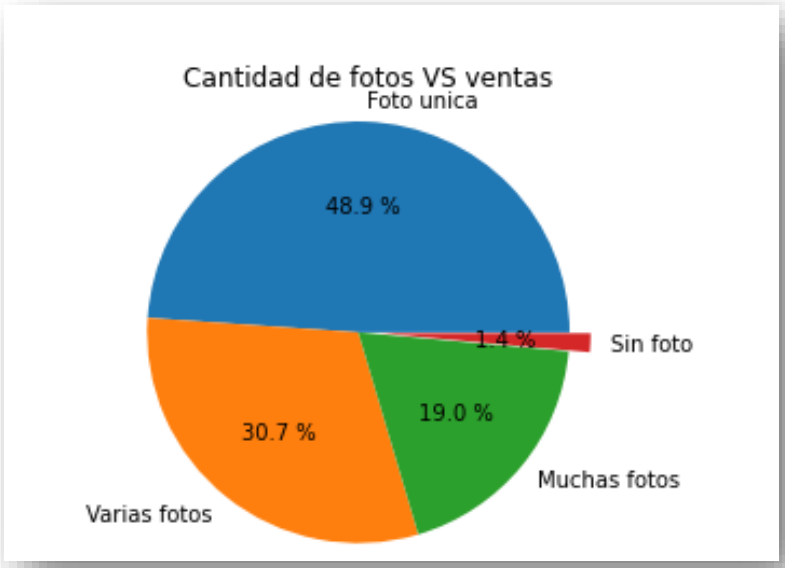


Cantidad de fotos VS ventas

Teniendo en cuenta la cantidad de fotos disponibles por producto, se realizó una segmentación en categorías, obteniéndose 4 grupos:

- Sin foto: no se cuenta con fotografías del producto.
- Foto única: sólo se dispone de una fotografía.
- Varias fotos: se cuenta con 2 o 3 fotografías.
- Muchas fotos: se dispone de 4 o más fotografías.

Luego se contabilizó la cantidad de productos ordenados, según cada categoría, hallándose los siguientes resultados:



Como puede observarse, los productos que carecían de fotografías, fueron los que menos ventas presentaron.

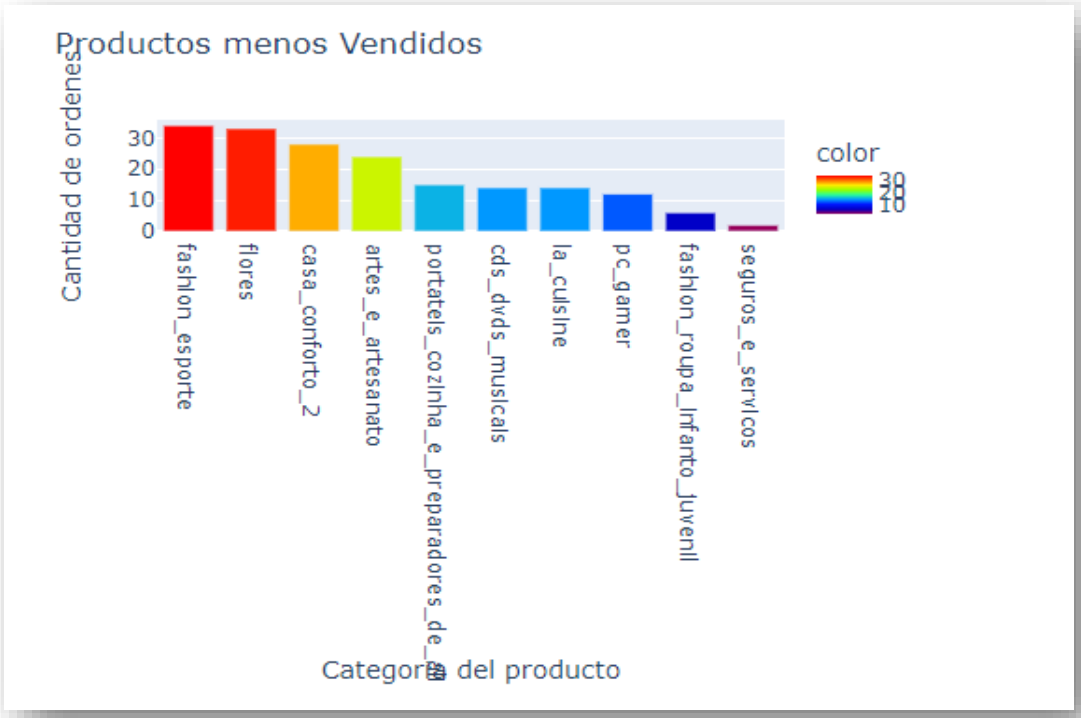
Ventas por mes

Analizando la cantidad de ventas registradas por cada mes del año (promediando los años de registros aportados por el dataset modelo), obtuvimos que el período que va desde el mes de marzo al mes de agosto, es el de mayores ventas. Siendo particularmente el mes de agosto quien lidera este ranking y el mes de septiembre su contraparte.

Horas de mayor cantidad de ventas

La franja horaria donde se registraron la mayoría de las ventas fue entre las 10 AM y las 9 PM, horario que no fue demasiado revelador.

Categorías de productos con mayor y menor cantidad de ventas



Como puede observarse en los gráficos, las categorías de productos con mayor cantidad de ventas fueron: cama_mesa_banho; beleza_saude ;esporte_lazer; informatica_acessorios y moveis_decoracao.

Y las categorías con menor cantidad de ventas: cds_dvds_musicais ;la_cuisine; pc_gamer; fashion_roupa_infanto_juvenil y seguros_e_servicos.

Cantidad de ventas según región geográfica

La región en la cual se registraron más ventas, fue la región Sudeste (68.65%), sucedida por la región Sur (14.22%), la región Nordeste (9.44%), la región Centro oeste (5.81%) y en último lugar por la región Norte (1.86%).

Predicciones generadas mediante el uso de aprendizaje automático

Tras identificar las categorías de productos con mayor cantidad de ventas, se decidió implementar un modelo de machine learning para predecir el flujo futuro de las ventas según dichas categorías.

El primer paso para realizarlo, fue seleccionar de todas las ventas, sólo aquellas que se hayan consumado. Luego se prepararon los datos, empezando por detectar valores outliers (los cuales no se contemplaron antes, dada su irrelevancia para la mayoría de los análisis realizados), y reemplazando estos valores por la media de ventas. A su vez, las fechas faltantes (por ausencia de ventas en ellas) se modificaron colocando el número cero.

Luego se eligió un modelo de machine learning adecuado, utilizando en este caso, al XG Boost Regressor model en conjunto con la librería Forecaster, la cual es utilizada para entrenamientos y predicciones en series de tiempo.

Una vez creado el modelo, se dividieron los datos, en datos de entrenamiento y datos de prueba, siendo destinados para la primera acción un 70% de los mismos y para la segunda, el restante 30%.

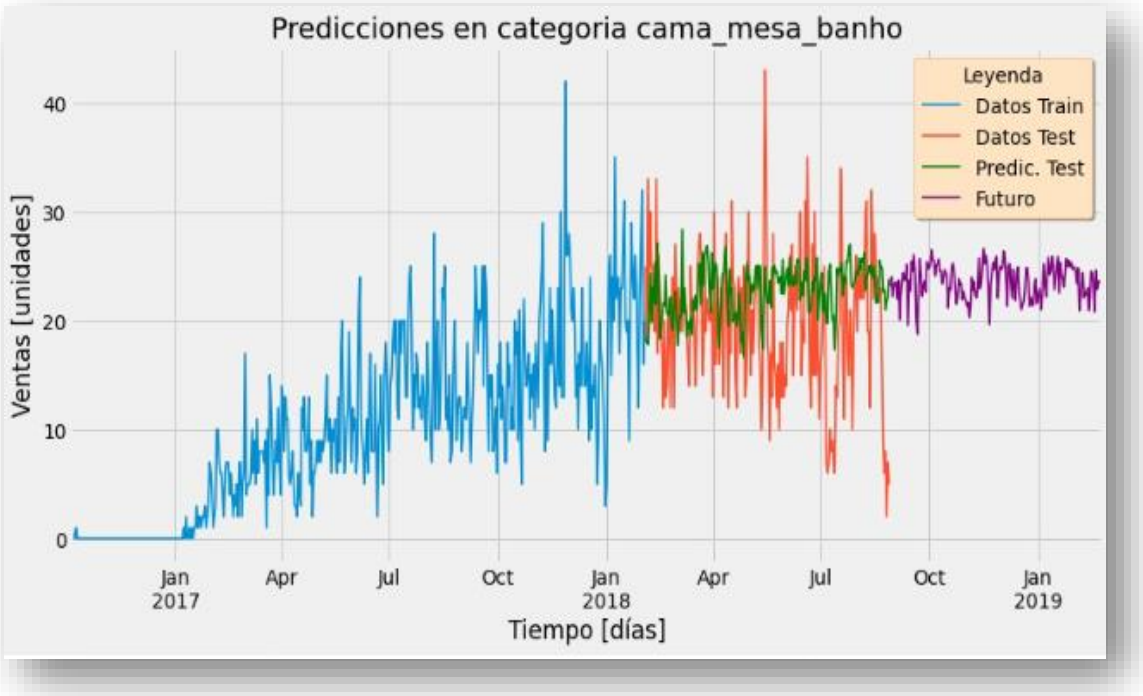
Para entrenar el modelo, se tomaron los 90 días anteriores a la última fecha registrada en el dataset. Luego de la etapa de entrenamiento, se testeó el funcionamiento del modelo, para una vez aprobada su eficacia, proceder a la predicción objetivo.

Finalmente se creó una función, que permitiera colocar la categoría de producto sobre la cual se desea conocer las ventas a futuro, y el plazo de tiempo en meses que se quiera evaluar. Esta función además de generar un gráfico con la predicción de ventas pedida, devuelve la cantidad de productos que serían necesarios para cubrir el stock de ventas que se espera en el plazo de tiempo evaluado.

Aplicando el modelo a las categorías de productos con mayor cantidad de ventas, se obtuvo que:

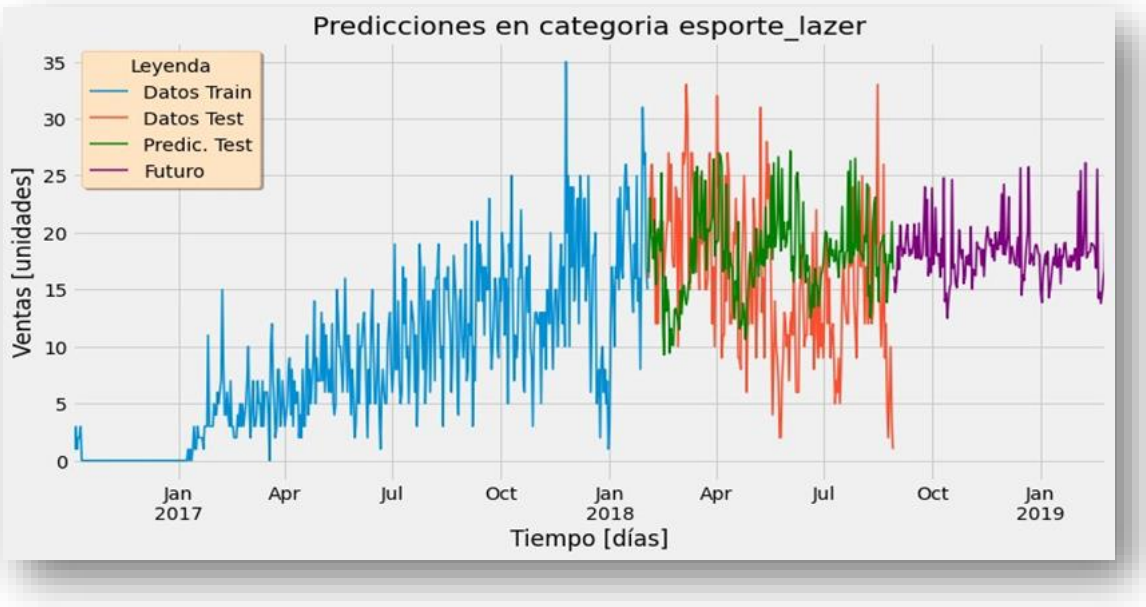
- ✓ Para la categoría cama_mesa_banho, se espera vender en un plazo de 6 meses, un total de 4246 productos.

Gráfica de predicción de ventas a 6 meses para la categoría cama_mesa_banho.



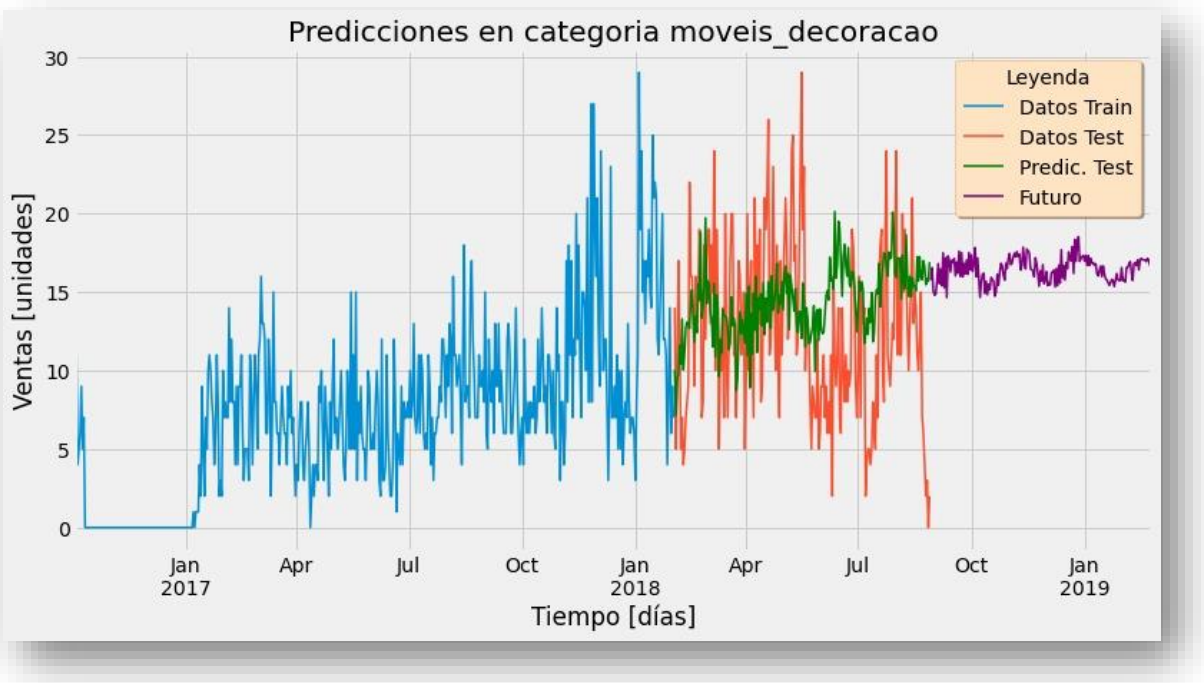
- ✓ Para la categoría esporte_lazer, se espera vender en un plazo de 6 meses, un total de 3282 productos.

Gráfica de predicción de ventas a 6 meses para la categoría esporte_lazer.



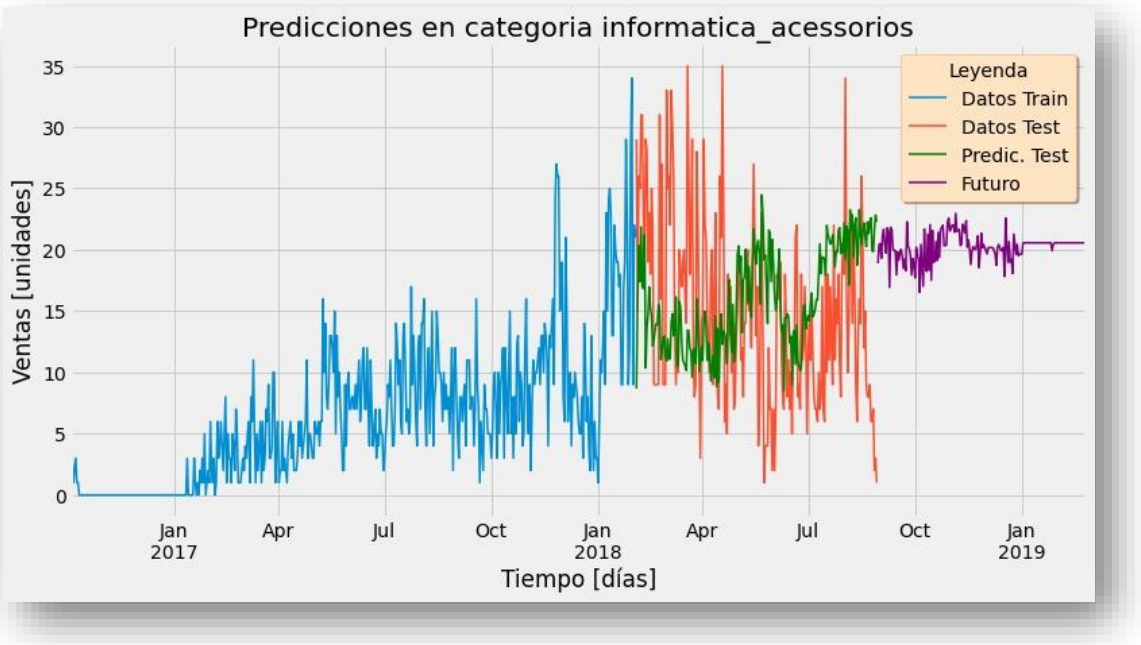
- ✓ Para la categoría moveis_decoracao, se espera vender en un plazo de 6 meses, un total de 2961 productos.

Gráfica de predicción de ventas a 6 meses para la categoría moveis_decoracao.



- ✓ Para la categoría informática_accesorios, se espera vender en un plazo de 6 meses, un total de 4246 productos.

Gráfica de predicción de ventas a 6 meses para la categoría informática accesorios.



Conclusiones

Luego de analizar la información recabada, podemos inferir que:

- ✓ Los problemas en las entregas de los productos pueden ser causantes de valoraciones negativas por parte de los clientes a sus experiencias de compra.
- ✓ El mayor porcentaje de puntuaciones bajas procede de los clientes pertenecientes a las regiones Nordeste y Norte de Brasil.
- ✓ Las 5 ciudades con porcentajes más altos de bajas valoraciones son Sao Paulo; Rio de Janeiro; Belo Horizonte; Brasilia y Curitiba.
- ✓ Las categorías de productos peores puntuadas son “cama-mesa-banho”; “beleza-saude”; “informática-accesorios”; “moveis-decoracao” y “Esporte-lazer”.
- ✓ Se registran demoras en las entregas de los productos.
- ✓ Las regiones con mayor demora son la región Norte y la región Sur.

- ✓ En el caso de la región Norte, se observa una coincidencia entre la presencia de demoras en la entrega de productos y el elevado número de puntuaciones negativas registradas por los clientes.
- ✓ El costo promedio del flete es más elevado en la zona Norte y más bajo en la zona Sudeste.
- ✓ La ausencia de fotografías del producto a vender, incide en la menor cantidad de compras del mismo.
- ✓ La mayor cantidad de ventas se registran en el período que va desde el mes de marzo hasta el mes de agosto.
- ✓ Agosto es el mes de más ventas y septiembre su contraparte.
- ✓ La franja horaria más activa para las compras va desde las 10 AM a las 9 PM.
- ✓ Las 5 categorías de productos con mayor cantidad de ventas son cama_mesa_banho; beleza_saude; esporte_lazer; informatica_acessorios y moveis_decoracao y las 5 con menor cantidad son cds_dvds_musicais; la_cuisine; pc_gamer; fashion_roupa_infanto_juvenil y seguros_e_servicos.
- ✓ La región con mayor porcentaje de ventas es la región Sudeste y la que menos ventas registra es la región Norte.

Recomendaciones

Finalmente, podemos recomendar al cliente las siguientes estrategias para lograr que la expansión de su negocio al mercado brasileño sea lo más performante posible:

Estrategias de logística

Se sugiere que, al momento de planificar la locación de centros de depósito, se tenga en cuenta, por un lado, que la ubicación del mayor número de clientes probablemente sea en la región Sudeste del país, motivo por el que sería óptimo acortar las distancias de transporte del producto hasta la residencia del cliente, disminuyendo costos de flete y

posibles demoras en los tiempos de entrega. Por otro lado, la misma estrategia, pero en este caso enfocada no en retener clientes, sino en conquistar nuevos y mejorar la experiencia de compra de quienes ya son parte del mercado E-commerce, sería interesante, evaluar la posibilidad de contar con una central de depósito en la zona más relegada en ventas, refiriéndonos en este caso, a la región Norte, región que en el mercado actual, no sólo presenta un pequeño porcentaje de compras en relación a las demás regiones, sino que sus pobladores son en quienes recae el costo de flete más elevado y quienes registran mayor número de valoraciones negativas, atribuidas en gran parte a demoras en las entregas de los productos.

Por otro lado, dado que la región Sur, se encuentra en segundo puesto en el ranking de ventas y a su vez, es una de las regiones donde los tiempos de entrega conllevan mayor demora, se recomienda evaluar la apertura de un depósito de productos en la misma.

En función de los puntos descriptos, se sugiere la implementación de los siguientes KPIs:

Disminución del tiempo de demora de las entregas.

Objetivo específico:

Reducir las demoras un 50% respecto a las observadas en el mercado actual, en el plazo de un año.

PDDO = PDDA / 2

PDDO = Promedio de diferencia de días objetivo.

PDDA = Promedio de diferencia de días en el mercado actual.

DD = Diferencia de días (Fecha de entrega del producto - Fecha estimada de entrega del producto).

Disminución de las puntuaciones negativas causadas por entregas.

Objetivo específico:

Reducir la aparición de puntuaciones negativas producto de problemas con las entregas, un 10% cada mes respecto al mismo mes del año anterior.

NPNMAAC = NPNMAA – 10% NPNMAA

NPNMAAC= Número de puntuaciones negativas por mes año actual.

NPNMAA= Número de puntuaciones negativas mes año anterior.

Estrategias de venta

Conociendo las categorías de productos con mayor y menor cantidad de ventas, se sugiere por un lado, evaluar la viabilidad de contar con un stock disponible para las categorías con altas ventas (lo cual es posible predecir en un corto plazo) y por otro lado, elaborar estrategias de marketing adecuadas, para que, en el caso de aquellos productos con gran caudal de ventas, se pueda lograr un crecimiento respecto al mercado de E-Commerce actual o mínimamente mantener dicho perfil, y en el caso de las categorías de productos más rezagadas, decidir si la demanda es lo suficientemente significativa como para justificar su presencia en la cartera del negocio, o si, por el contrario, sería mejor no contar con ellas. O bien, averiguar en estos casos, si el problema es el producto en sí, o si la falta de ventas del mismo depende de otros factores modificables, por ejemplo, una categorización poco llamativa de un producto, como podría decirse de la existencia de una categoría para “cds” y “dvds”, los cuales son formatos casi obsoletos en la actualidad (este grupo tal vez podría recategorizarse como “antigüedades”). A su vez, dada la posibilidad de que el problema no radique en las categorías, debería intervenirse mediante publicidad y otras estrategias para identificar y actuar mejorando los factores negativos que produzcan la baja en sus ventas.

Respecto a los meses en los que se registran las mayores cantidades de ventas, al igual que con las categorías de productos más demandadas, es necesario mantener siempre actualizadas las campañas de marketing, evitando que se produzcan grandes fluctuaciones en las ventas esperadas. Y para el caso particular de los meses de bajo caudal, sería óptimo redoblar las estrategias, para lograr un aumento considerable de las ventas, ya que, de otro modo, de continuarse dicho ciclo a lo largo de los años venideros, el negocio podría verse perjudicado.

Otro punto clave a tenerse en cuenta, es la importancia de que las publicaciones de productos a vender, contengan al menos una fotografía del mismo, ya que se detectó una tendencia significativa a obtener bajas ventas, cuando no se contaba con ninguna imagen del producto ofrecido.

Por otro lado, teniendo conocimiento de las regiones del país a donde se registra el mayor y el menor número de ventas, es posible también, plantear como una estrategia factible, el focalizar las publicidades, en captar a quienes residen en las zonas más rezagadas, por ejemplo, mediante descuentos específicos por región y/o con la implementación de envíos gratuitos, etc.

Estos mismos descuentos o estímulos, podrían utilizarse además, como estrategias para compensar a aquellos clientes que arrojasen puntuaciones negativas a sus compras. En el mercado analizado, se observó un gran número de puntuaciones y comentarios negativos en función de las experiencias de compra de los clientes, lo cual consideramos, es un apartado que no se debe dejar pasar por alto, ya que la voz del cliente, hoy en día, tiene un fuerte peso, generando que las referencias de un cliente actúen tanto positiva como negativamente, en la decisión de compra de las personas que ven sus valoraciones.

En función de los puntos hasta aquí planteados, se presentan los siguientes KPIs:

Aumentar las ventas del período de meses inactivos.

Objetivo específico:

Aumentar un 10% las ventas del período inactivo respecto al mismo período del año anterior.

$$\sum VPIAS = \sum VPIAI + ((\sum VPIAI * 10) /100)$$

VPIAI: ventas del período inactivo del año inicial.

VPIAS: ventas del período inactivo del año siguiente.

Incrementar las ventas en la región Norte

Objetivo específico:

Obtener luego de 3 meses un incremento del 5% en las ventas, respecto al mes inicial evaluado.

$$\sum VRN3M = \sum VRNI + ((\sum VRNI * 5) /100)$$

VRNI: ventas en la región Norte en el mes inicial.

VRN3M: ventas en la región Norte luego de 3 meses.

Aumentar el número de comentarios positivos de los clientes

Objetivo específico:

Incrementar la cantidad de reviews positivas un 5 % respecto al mismo mes del año anterior.

$$\sum RPMAS = \sum RPMI + ((\sum RPMI * 5) /100)$$

RPMI: reviews positivas en el mes inicial.

RPMAS: reviews positivas del mismo mes al año siguiente.

Mantener la brecha entre comentarios (+) y (-)

Objetivo específico:

Mantener la diferencia porcentual entre los comentarios positivos y negativos menor o igual al 1%.

$$(100 * (CN / (CN + CP))) \leq 1\ \%$$

CN: comentarios negativos. CP: comentarios positivos