**Análisis de la percepción de los clientes en un marketplace usando técnicas de ciencia de datos**

Manuel Alejandro Pachón Santana, Steven Andrés Llerena Navarro, Juan Sebastián Otálora Leguizamón
Universidad de los Andes, Bogotá, Colombia
{ma.pachons, s.llerena, j.otaloral}@uniandes.edu.co
Fecha de presentación: septiembre 20 de 2021

# Tabla de contenido



# Entendimiento del problema

## Contexto

Olist es una startup creada en Brasil en 2015 que facilita el comercio, logística y administración del capital de productos en las pequeñas y medianas empresas, por medio de un ecosistema tecnológico que permite aumentar la visibilidad de los productos de sus clientes en tiendas virtuales individuales o tiendas virtuales con base en grandes alianzas.

En los últimos años, Olist se ha planteado la meta de mejorar la experiencia y la percepción que sus clientes tienen después de realizar una compra, ya que las dinámicas de intercambios están evolucionando y Olist quiere estar preparada para enfrentar y mejorar la calificación del servicio luego de una compra.

## Objetivos

Definir los objetivos del proyecto de negocio que se usarán para la validación.

1. Identificar las causas de la percepción de los clientes luego de realizar una compra, a partir de la información obtenida por el sistema de información.
2. Mejorar los procedimientos que impacten en la percepción de los clientes

### Métricas

- Incremento en las tasas de calificación de los usuarios
- Cuantas áreas estratégicas usan el modelo para mejorar sus estrategias
- Número de estrategias basadas en datos
- Incremento del promedio de calificaciones obtenidas por los usuarios
- Evaluación de la calidad de las descripciones de los productos

## Ideación

### Arquetipo de usuario: Gerente de Customer Success

Mayara Scholze, es una mujer de 37 años con más de 12 años de experiencia comercial y en gestión del cliente

1. Comportamiento e ideas: Hace análisis generales y segmentados sobre hojas de cálculo que el equipo de TI le comparte. Hace algunos análisis descriptivos y define estrategias de implementación física con el fin tener una vista integral de la percepción del cliente.
2. Dolores: No entiende los patrones asociados a las calificaciones de los clientes, y económicamente es inviable ampliar su estrategia física. No sabe el impacto de sus análisis para proveer estrategias a los equipos de ventas, operación, delivery y pricing de Olist.

### Ideas de solución

1. Predecir el puntaje de calificación de un producto para entender como se ve impactado por las características del cliente, el producto, la orden y del proceso de entrega.
2. Determinar los clientes con compras frecuentes que dan baja calificaciones a sus productos para dar un incentivo o crear una estrategia focalizada con ellos.
3. Determinar las ciudades y rutas de operación con mayores tiempos de entrega de las ordenes, con el fin de decidir la necesidad de apertura de nuevos centros de distribución y logística.

### Selección de idea

- Predecir el puntaje de calificación de un producto para entender el impacto que las variables asociadas tienen sobre el puntaje.

Se elige porque es un proyecto viable pues la compañía cuenta con los datos necesarios e históricos para el proyecto, y tiene potencial de impactar más

perspectivas del negocio que tengan capacidad de accionar los planes de mejora de acuerdo con los insights.

### Prototipo

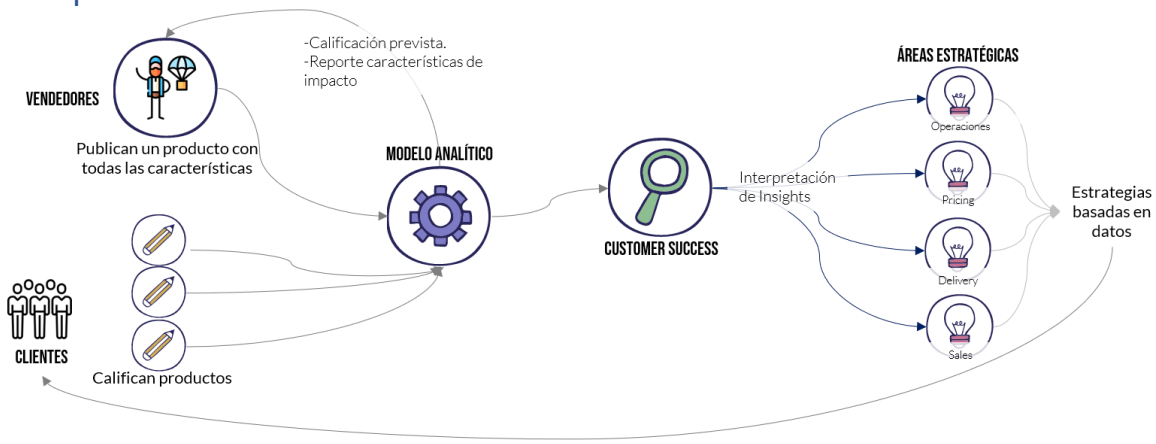


Figura 1. Prototipo storyboard: Predicción puntajes de calificación en compras

## Implicaciones

Respecto a las implicaciones éticas, riesgos legales o sesgos que puede tener el proyecto, se identifican algunos riesgos:

- Generación de clasificaciones basados en aspectos étnicos o sensibles de los clientes.
- Generar sesgos sobre categorías de productos que puedan afectar sectores de la industria determinados.
- Que la geolocalización y datos demográficos permitan individualizar los clientes.
- Uso de datos no autorizados en las políticas de tratamiento y uso de datos.

## Enfoque analítico

### Hipótesis

1. La descripción del producto afecta su puntaje de calificación

### Tipos de análisis predictivo

Regresión Lineal ya que se busca predecir el puntaje de calificación (variable numérica) con base en las variables identificadas, teniendo los procedimientos descriptivos, de correlación y significancia estadística.

## Entendimiento de los datos

La fuente de datos de Olist tiene 9 dataset, los cuales contienen información de órdenes, productos, localización geográfica, vendedores, clientes, pagos y

calificaciones, las cuales podrán ser consultadas en el siguiente enlace: https://www.kaggle.com/olistbr/brazilian-ecommerce

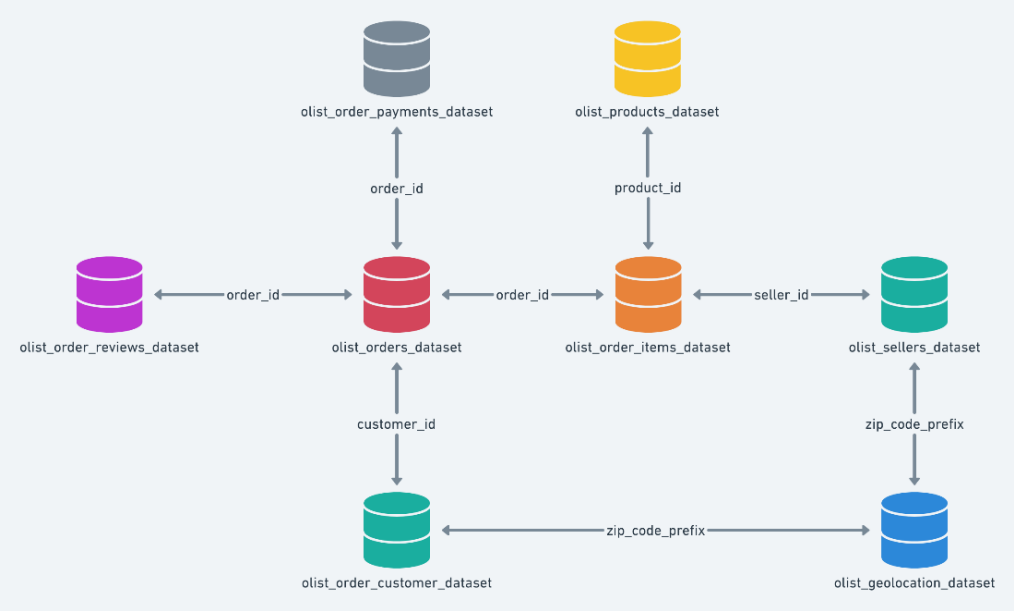


Figura 2. Esquema de datos - Brazilian E-Commerce Public Dataset by Olist

Para el problema que se está abordando, sólo son necesarios los dataset relacionados a los productos, órdenes y calificaciones (olist_order_reviews_dataset, olist_products_dataset y olist_order_items_dataset)

**Diccionario de datos**

- review_score: Calificación dada de 1 a 5 por parte del cliente en la encuesta de satisfacción.
- product_name_lenght: Cantidad de caracteres extraídos del nombre del producto
- product_description_lenght: Cantidad de caracteres extraídos de la descripción del producto
- product_photos_qty: Cantidad de fotos publicadas

**Perfilamiento de datos**

- Cantidad de atributos: 4 (3 numéricos y 1 categórico)
- Cantidad de observaciones: 100.785
- Nulos: 4227 (1.0%)
- Duplicados: 14.951 (14.8%)
- Outliers no observados

# Análisis de los datos

## Análisis descriptivo

Partiendo de la primera hipótesis planteada por el negocio respecto a determinar si la descripción del producto afecta su puntaje de calificación, se procede al análisis de las variables, *product_name_lenght, product_description_lenght* y *product_photos_qty*, y su comportamiento con la variable objetivo *review_score.*

Iniciando con los análisis univariado, en la figura 3 se observa el comportamiento de la variable *product_description_lenght*, del que se puede concluir que no tiene una distribución de datos orientada por un patrón determinada, y que, tiene *outliers* que dentro del negocio es común para los productos que requieres extensas descripciones. Sin embargo, en los análisis posteriores se deberá determinar cómo usar estos datos para evitar sesgos e inconsistencias

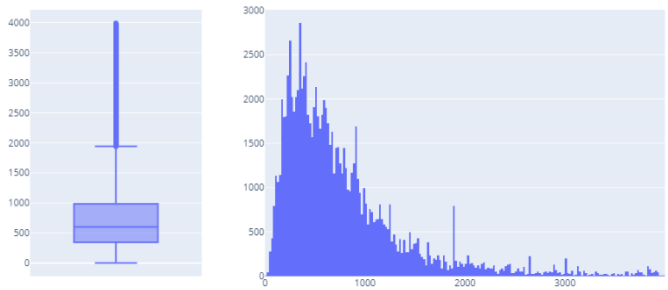


Figura 3. Boxplot e histograma de variable product_description_lenght

En la figura 4 se observa un contraste en los *outliers* de la longitud del nombre de producto, pues por lo general se prefieren etiquetas largas, pues las de longitud reducida son catalogadas como datos atípicos. En ese sentido, desde el negocio se podría suponer que los nombres largos proveen información más rápida al comprador.

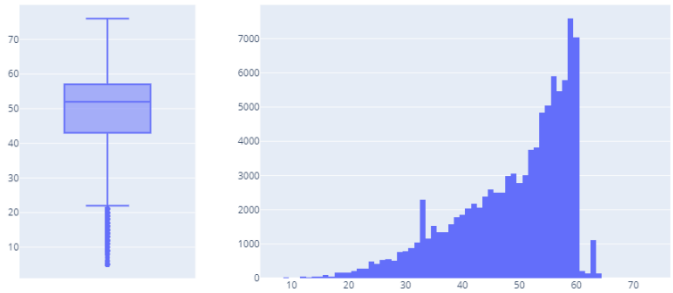


Figura 4. Boxplot e histograma de variable product_name_lenght

En presenta un comportamiento particular en la cantidad de fotos, pues la mayoría de los productos entregados fue publicada con pocas imágenes. Los *outliers* se presentan en productos con gran cantidad de ellas. De hecho, contrario a la perspectiva del negocio, parece no haber una relación directa de estas dos variables según la figura 5.

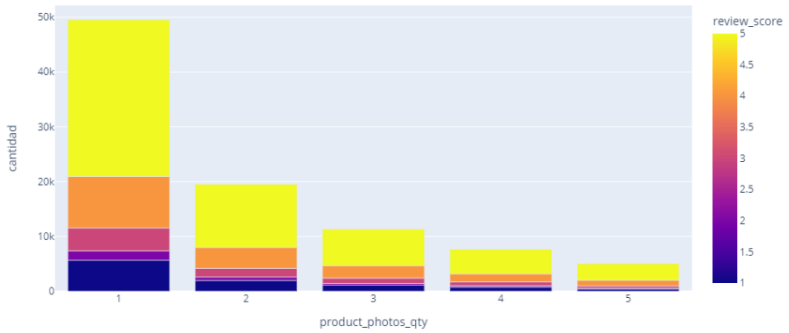


Figura 5. Gráfica de barras: Review Score y Cantidad de Fotos

En la figura 6 se pretende estudiar la relación entre la longitud del nombre, le tamaño de la descripción y la calificación de la compra, para determinar posibles patrones. Sin embargo, no se evidencia una relación clara que identifique una correlación de las variables. Tomando únicamente como referencia esta imagen no se puede afirmar que el *review score* de un producto aumente proporcionalmente con la longitud de las características descriptoras.

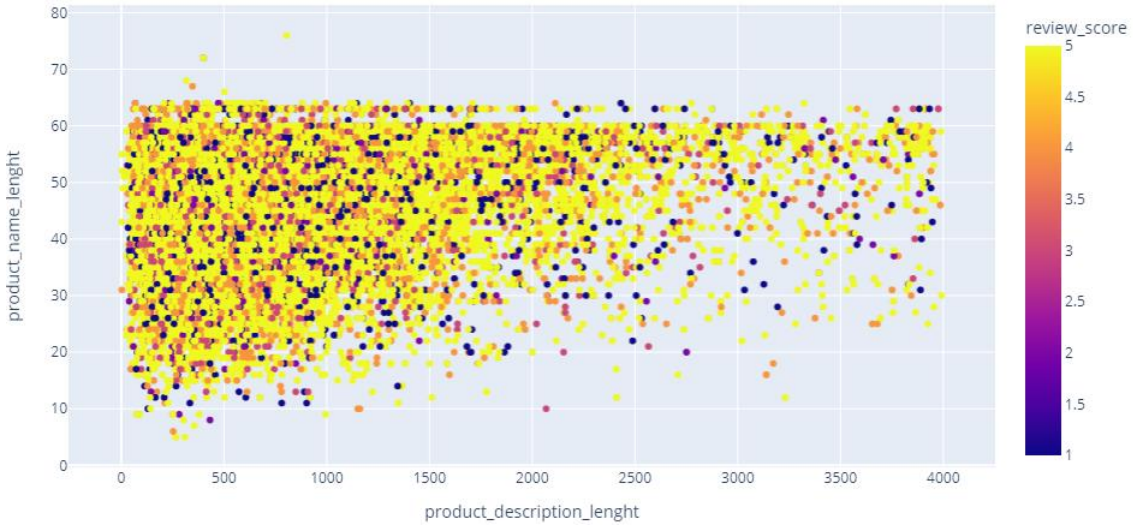


Figura 6. Gráfico de dispersión product_name_lenght vs product_description_lenght

En general no se han podido encontrar elementos descriptivos en datos o elaciones gráficas que sugieran la dependencia de la caracterización del producto en el Marketplace son su puntaje de calificación de compra. En ese sentido se e hará test estadístico de mayor rigurosidad para validar la hipótesis, y de manera previa se sugiere al cliente ampliar el espacio de búsqueda a otras variables que expliquen este comportamiento: tiempos, cantidad de órdenes del vendedor, etc.

**Nota**: Todo el análisis exhaustivo de estas variables, otro tipo de gráficas y métricas se encuentran en el notebook adjunto en el repositorio de GitHub.

### Validación de hipótesis

Teniendo en cuenta la hipótesis planteada y luego de la validación de estas con las variables de descripción (longitud del nombre, longitud de la descripción y cantidad de fotos) se identificó lo siguiente:

| Variable | Prueba t – test | ANOVA |
|---|---|---|
| **Longitud Nombre** | Rechazar Hipótesis nula | Rechazar Hipótesis nula |
| **Longitud Descripción** | Rechazar Hipótesis nula | Rechazar Hipótesis nula |
| **Cantidad de fotos** | Rechazar Hipótesis nula | Rechazar Hipótesis nula |

Es por esta razón que la longitud de la descripción de los productos no afecta en el puntaje de calificación que otorgan los clientes a los productos y la hipótesis inicialmente planteada es rechazada.

La ampliación de los procedimientos podrá ser consultados en el notebook 04_Proyecto_Entrega1_ConfirmaciónHipótesis.ipynb en GitHub.

## Conclusiones y acciones sugeridas

- Los datasets obtenidos de la fuente de datos que son relevantes para el ejercicio, no cuentan con duplicados y tienen muy pocos nulos que no afectan en gran medida a los datos.
- El dataset final construido cuenta con una buena calidad para abordar el problema propuesto.
- Existe una presencia importante de *outliers* en las variables que describen los productos, y no se puede identificar con facilidad alguna distribución que explique ese comportamiento. Además, se deberá evaluar la pertinencia de usar esos datos atípicos en el modelamiento predictivo.
- Se deberían explorar más variables que expliquen el comportamiento del puntaje de calificación, puesto que desde el análisis descriptivo no se ven patrones evidentes al evaluar únicamente los atributos que describen el producto.
- La descripción del producto no afecta la calificación que el usuario asigna en el sistema. Sin embargo, podría evaluarse semánticamente la descripción para verificar si una hipótesis similar sobre la descripción afecta la calificación.

## Repositorio GitHub

Para la consulta de los notebooks realizados puede utilizar el siguiente enlace: https://github.com/stevenllerenan/MINE4101-202120-CDA

## Contribuciones

A pesar de que una parte del trabajo fue dividida entre los integrantes del grupo, todos participaron en la revisión de este, aportando mejoras y/o correcciones.

Manuel Alejandro Pachón Santana – Líder de Analítica y Negocio: Estudio de negocio, identificación de problemas, planteamiento de hipótesis, confirmación de hipótesis y análisis descriptivo introductorio.

Steven Andrés Llerena Navarro - Líder de Datos: Definición del tipo de análisis predictivo, entendimiento de los datos, entendimiento de los datos/producto a nivel de calidad de los datos y el dataset resultante.

Juan Sebastián Otálora Leguizamón - Líder de Proyecto: Organización de encuentros de equipo, y asignación de actividades. Proceso de Ideación y prototipado, además del análisis descriptivo previo a la validación de hipótesis.