Data Science
Comisión 14100

Grupo: Juan Gabriel Jara, Karina Rosa, Juan Ignacio Garcia

Septiembre, 2021
Entrega 1

Contenidos

1. Escenario de trabajo
2. Posibles análisis
3. Fuente de datos
4. Stack tecnológico
5. Diagrama de tablas Base de Datos OLIST
6. Diccionario de features general
7. Hallazgos EDA
8. Análisis univariado y bi variado
9. Modelos y medidas - Arbol de decisión
10. Natural Language Process (NPL)

1. Escenario de Trabajo

Trabajaremos con un conjunto de datos públicos de comercio electrónico brasileño de pedidos realizados en Olist Store. El conjunto de datos tiene información de 100k pedidos de 2016 a 2018 realizados en múltiples mercados en Brasil.

Sus características permiten ver un pedido desde múltiples dimensiones: desde el estado del pedido, el precio, el pago y el desempeño del flete hasta la ubicación del cliente, los atributos del producto y finalmente las reseñas escritas por los clientes.

El objetivo de esta primera entrega es realizar el EDA del dataset con el fin de analizar las features disponibles y la capacidad de presentar un algoritmo de predicción supervisado. Para ello deberemos:

- Realizar la preparación de los datos para obtener el dataset final
- Realizar análisis de los features para poder generar las primeras conclusiones sobre nuestro caso de estudio.

2. Posibles análisis

PNL

Este conjunto de datos ofrece un entorno supremo para analizar el texto de las reseñas a través de sus múltiples dimensiones.

Agrupación (Clustering)

Algunos clientes no escribieron una reseña. Pero, ¿por qué están felices o enojados?

Predicción de ventas:

Con la información de la fecha de compra, podrá predecir las ventas futuras.

Rendimiento de entrega:

También podrá trabajar en el desempeño de la entrega y encontrar formas de optimizar los tiempos de entrega.

Calidad del producto
Descubrir las categorías de productos que son más propensas a la insatisfacción del cliente.

3. Fuente de datos

- Data set original proveniente de Kragle

  https://www.kaggle.com/olistbr/brazilian-ecommerce

  Licencia de uso:
  CC BY-NC-SA 4.0

- Archivos del dataset

  Archivos separados por comas.
  Esquema de base de datos relacional
  - Un pedido puede tener varios artículos.
  - Cada artículo puede ser realizado por un vendedor distinto.
  - Todo texto que identifica tiendas y socios fue reemplazado por los nombres de las grandes casas de Game of Thrones.

4. Stack tecnológico

Initial Stack

- SQL server 2019 - OLIST DB
- ELT to load csv files into SQL DB
- Jupyter notebook

5. Diagrama de base de datos

6. Diccionario de features

olist_customers	
customer_id	Identificador cliente en data set
Customer_Unique_id	Identificador único del cliente
customer_city	Ciudad
customer_state	Estado

olist_orders	
order_id	Identificador único del pedido
customer_id	Identificador cliente
order_status	Estado del pedido
order_purchase_timestamp	Fecha y hora de la generación del pedido
order_approved_at	Fecha y hora de la aprobación del pedido
order_delivered_carrier_date	Fecha y hora de la entrega del pedido a distribución
order_delivered_customer_date	Fecha de la entrega del pedido a cliente
order_estimated_delivery_date	Fecha de la entrega estimada del pedido a cliente

6. Diccionario de features (Cont)

olist_order_items	
order_id	Identificador único del pedido
order_item_id	Identificador único de la línea de pedido
product_id	Identificador del producto
seller_id	Identificador del vendedor
shipping_limit_date	Fecha límite de entrega del vendedor
price	precio
freight_value	Valor del flete

olist_order_items	
order_id	Identificador único del pedido
payment_sequential	Identificador del medio de pago usado (puede ser más de un medio de pago)
payment_type	Método de Pago
payment_installments	Número de cuotas
payment_value	Valor de la transacción

6. Diccionario de features (Cont)

olist_order_reviews	
review_id	Identificador único del review
order_id	Identificador único del pedido
review score	Calificación (1-5) del pedido
review_comment	Comentario de la revisión
review_creation_date	Fecha creación de la encuesta
review_answer_timestamp	Fecha respuesta de la encuesta

olist_geolocation	
geolocation_zip_code_prefix	primeros 5 dígitos del código postal
geolocation_lat	Latitud
geolocation_lng	Longitud
geolocation_city	Ciudad
geolocation_state	Estado

6. Diccionario de features (Cont)

olist_sellers	
seller_id	Identificador del vendedor
seller_zip_code_prefix	primeros 5 dígitos del código postal
seller_city	Ciudad
seller_state	Estado

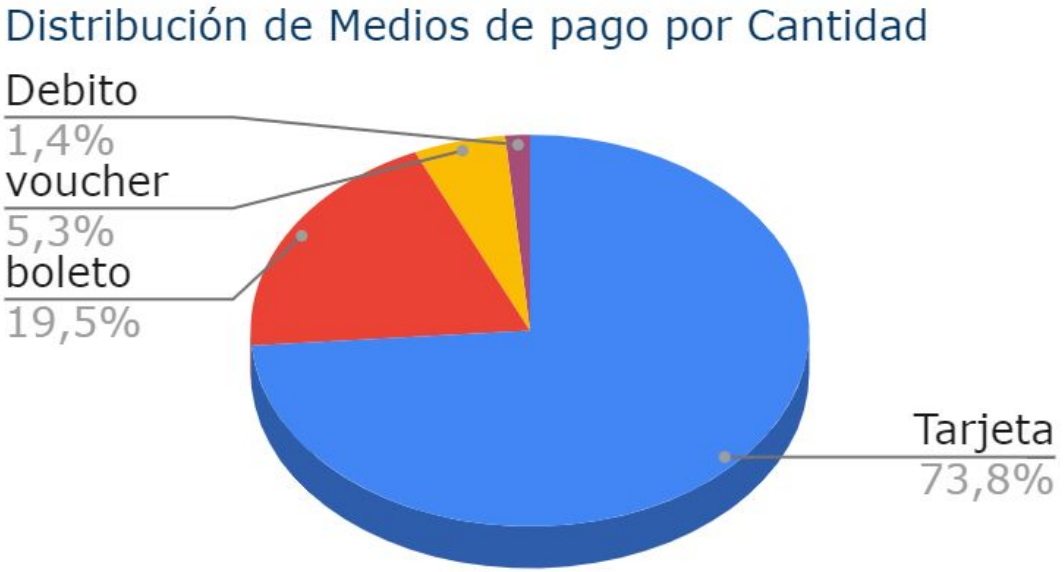
7. Hallazgos EDA

- Se pudo construir un modelo relacional sobre SQL Server con claves primarias y relaciones
- Registros y features por dataset

Órdenes: (99441, 8)
Ítems por orden: (112650, 7)
Productos: (32951, 9)
Sellers: (3095, 4)
Pagos: (103886, 5)
Customers: (99441, 5)
Reviews: (100000, 7)
Categoría Productos: (72, 2)

- No se analizan series de tiempo. Algunos campos contienen valores nulos y se asume que son pedidos en distribución.
- La desviación estándar del precio de cada ítem de orden es de 18363 reales. Con el mínimo en 85 reales y el máximo en 673500 reales

8. Análisis Univariado, bivariado

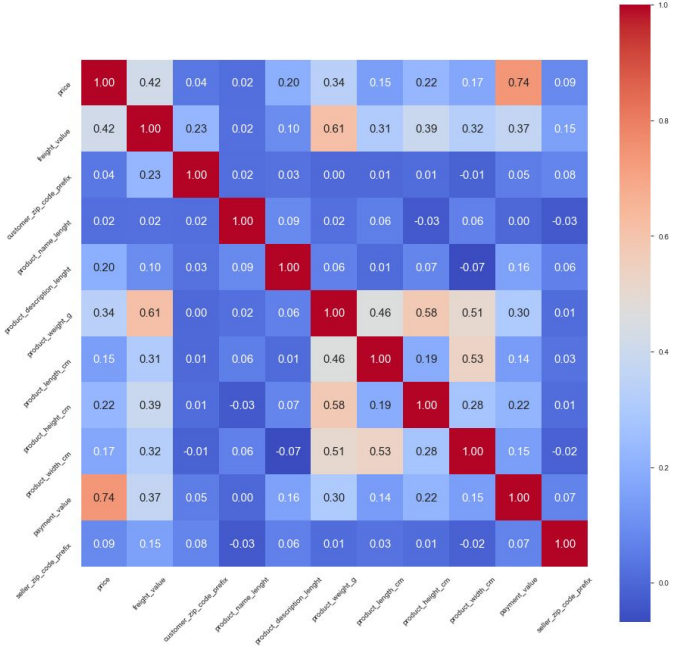


Se sugiere optimizar los canales de pago para el pago con tarjeta de crédito y orientar las campañas de marketing a promociones buscando alianzas con distintos bancos.

8. Análisis de correlaciones

Conviene estudiar mas en detalle la relación entre:

- order_item_id y payment_value (linea de pedido y valor del pago)
- product_weight y freight_value (peso del producto y valor del flete)
- price y payment_installments (precio y cuotas)



9. Modelos y medidas - Árbol de decisión

Intentamos predecir el valor de la review partiendo de las variables

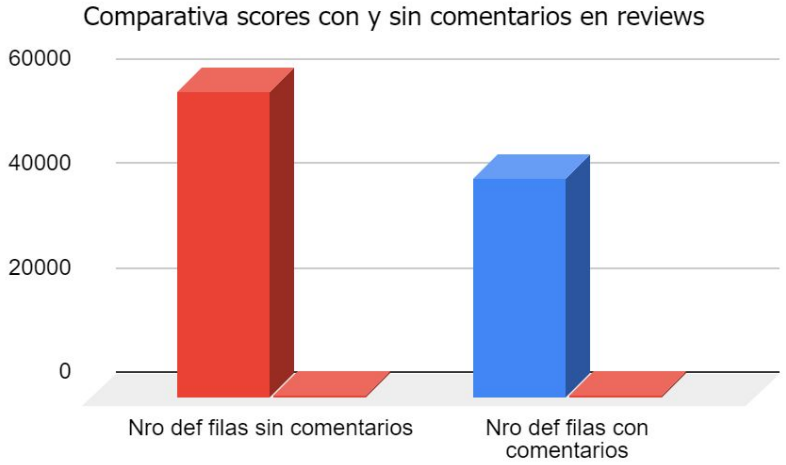
- "order_purchase_timestamp",
- "order_approved_at",
- "order_delivered_carrier_date",
- "order_delivered_customer_date",
- "order_estimated_delivery_date"

% de aciertos sobre el set de entrenamiento: 56,05%
% de aciertos sobre el set de evaluación: 56,48%

1. Pese a la casi nula correlación que existía en las variables, luego de entrenar el modelo logramos que el mismo reproduzca con un ***55% de acierto*** la calificación de la review del cliente. Es probable que este resultado se haya logrado gracias a la gran cantidad de observaciones con las que cuenta el dataset.
2. También es importante destacar que parece que 3 variables serían inútiles en el modelo, bastaría con utilizar 'order_purchase_timestamp' y 'order_delivered_customer_date' para alcanzar resultados similares.
3. Debido a la poca diferencia que se aprecia entre los porcentajes de aciertos entre el conjunto de 'train' y el de 'test' podríamos afirmar que nuestro modelo **no sufre** de overfitting.

10. Natural Language processing

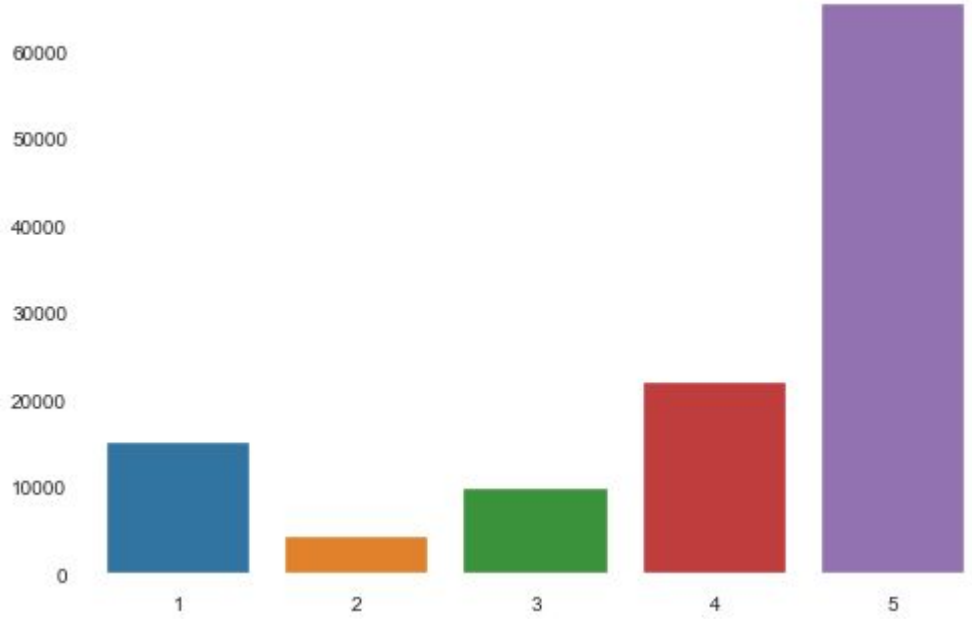
Intentamos predecir el valor de la review partiendo de la variable **comment_review**



Muestra	Score
Observaciones	9.984
Media	3,81
std	1,36
min	1,00
25%	4,00
50%	5,00
75%	5,00
max	5,00

10. Natural Language processing

Intentamos predecir el valor de la review partiendo de la variable **comment_review**



10. Natural Language processing

Intentamos predecir el valor de la review partiendo de la variable **comment_review**

10. Natural Language processing

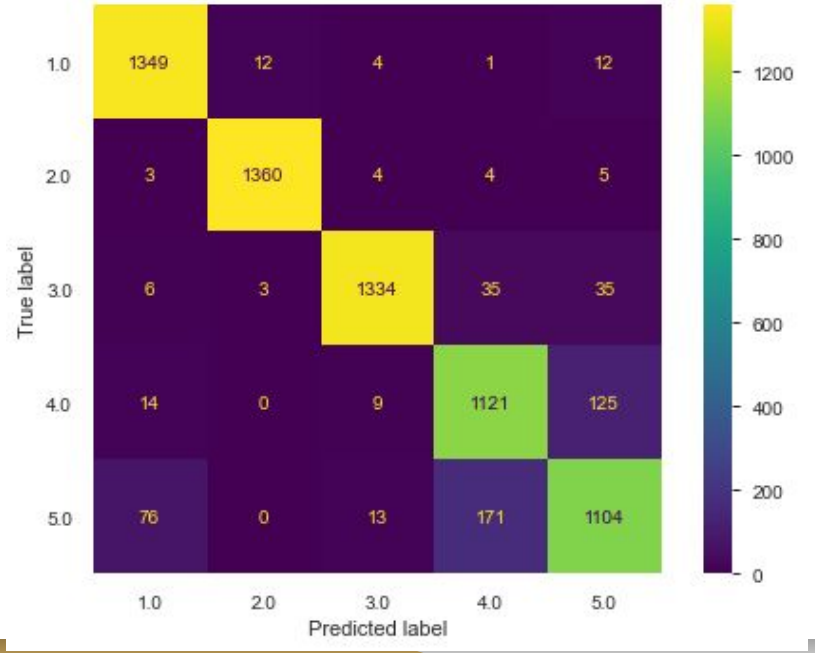
Intentamos predecir el valor de la review partiendo de la variable **comment_review**

10. Natural Language processing

Intentamos predecir el valor de la review partiendo de la variable **comment_review.**

Medidas

	precision	recall	f1-score	support
1.0	0.58	0.83	0.69	486
2.0	0.33	0.01	0.03	134
3.0	0.33	0.04	0.07	193
4.0	0.28	0.04	0.07	347
5.0	0.73	0.94	0.82	1336
accuracy			0.67	2496
macro avg	0.45	0.37	0.33	2496
weighted avg	0.58	0.67	0.59	2496



10. Natural Language processing

Intentamos predecir el valor de la review partiendo de la variable **comment_review.**
OVERSAMPLING CON RANDOM FOREST

```
print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
1.0	0.93	0.98	0.95	1378
2.0	0.99	0.99	0.99	1376
3.0	0.98	0.94	0.96	1413
4.0	0.84	0.88	0.86	1269
5.0	0.86	0.81	0.83	1364
accuracy			0.92	6800
macro avg	0.92	0.92	0.92	6800
weighted avg	0.92	0.92	0.92	6800

- Las métricas del oversampling demuestran cercanía al overfiting. Este resultado genera sospechas sobre la calidad del modelo con sobremuestreo principalmente por su alta calificación, con lo cual es necesario evaluar la precisión de las distintas clases.
- A menos que el cliente considere estos resultados normales, se sugiere trabajar sobre el feature de review de compra de la página web para ajustar los resultados (por ejemplo, scoring 5 sin comentarios)
- Cabe la posibilidad de que al sobre muestrear nuestro modelo se genera el overfitting.