

1. Problem Space

An eCommerce company based in Brazil is looking for partner to work on a data science and machine learning project. The company is one of the leading companies in Brazil, and sell a wide-range of products. In the moment they are particularly interested in boosting their online sales along with online reputation.

As part of this, they want you to build a prediction model that will help predict the customer review score, which will allow them to particularly target the customers who are more likely to leave good reviews with emails and targeted messages and therefore, get more good reviews.



1. Problem Space



Smartphone Motorola Moto G6 Play Dual Chip Android Oreo - 8.0 Tela 5.7" Octa-Core 1.4 GHz 32GB 4G Câmera 13MP - Índigo

(Cód.133453169) ★★★★★ (215)



Caixa de Som ANKER SoundCore Bluetooth 12W - Preta
+ R\$ 429,99

pegue na loja hoje!

Pegue na loja mais próxima, no mesmo dia :)
Sujeito à alteração de preço. [Saiba mais](#)

[ver lojas](#)

Escolha uma loja abaixo e compre



olist
R\$ 1.299,00
R\$ 26,04 - 7 a 10 dias úteis

vendido e entregue por **olist**

R\$ 1.299,00

10x de R\$ 129,90 s/ juros

comprar

Corra! Temos apenas 5 no estoque



on...ra
R\$ 1.069,90
R\$ 38,32 - 7 a 10 dias úteis



mel...cê
R\$ 975,00
R\$ 22,94 - 5 a 6 dias úteis

Mais opções deste produto a
partir de **R\$ 959,00**



R\$ 1.299,00 em até 12x de R\$ 108,25 s/ juros

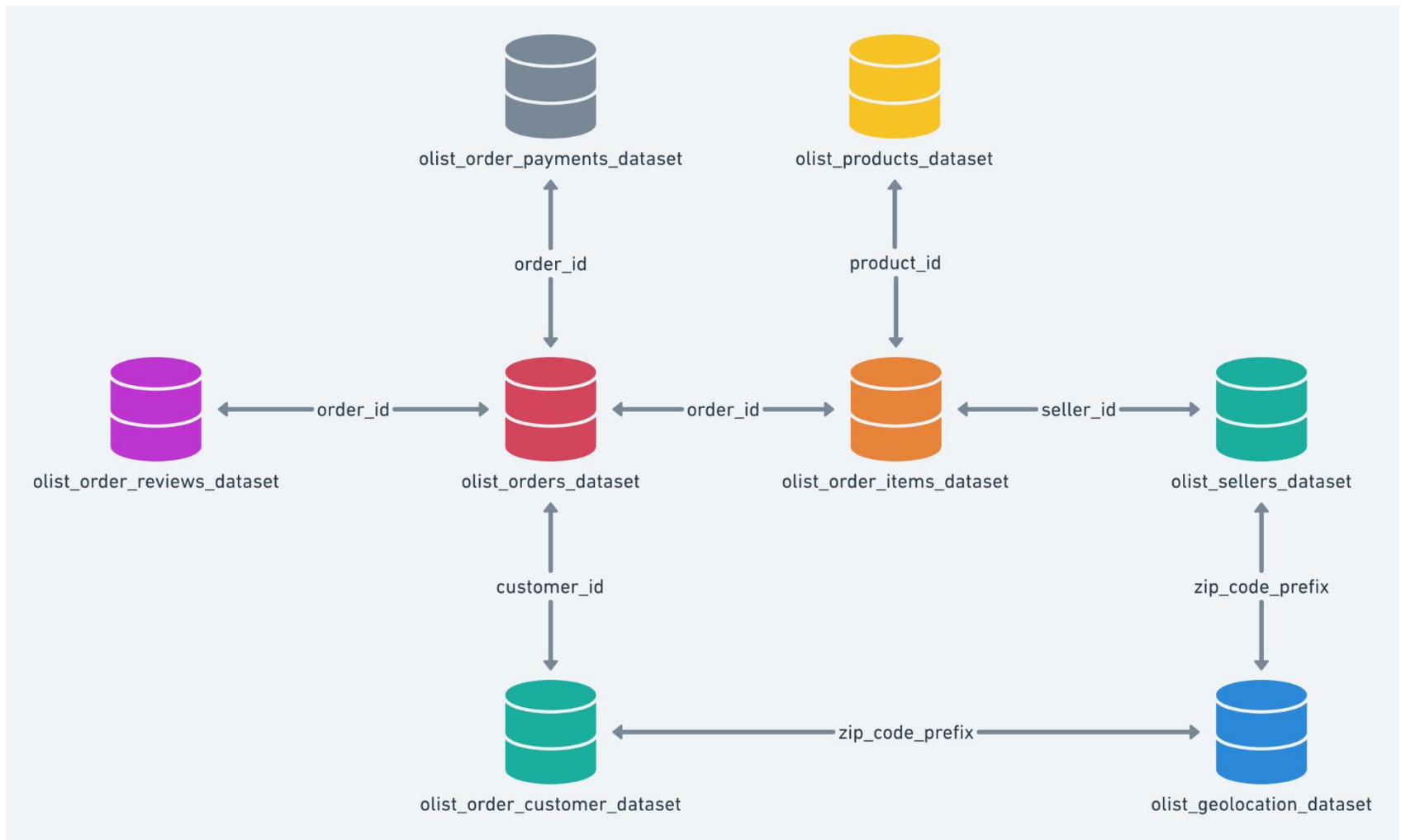
R\$ 1.299,00 no cartão ; em até 24x de R\$ 54,12 s/ juros

[formas de parcelamento](#)



Este produto é vendido por uma loja parceira.

2. Data



2. Data

olist_order_reviews_dataset

1. review_id: Unique review ID (primary key)
2. order_id: Unique order ID (foreign key)
3. review_score: 1-5* (target)
4. review_comment_message: Any comments left by the customer (in Portuguese)
5. review_creation_date: The date the survey (email notification) was sent to the customer
6. review_answer_timestamp: When the customer completed the review

2. Data

olist customers dataset

1. customer_id: Unique ID related to the order. Each new order generates a unique customer ID even if this customer has bought from the store before (foreign key)
2. customer_unique_id: Unique ID (one per customer) linked to their profile and order history (primary key)
3. customer_zip_code_prefix: First five characters of their zip code (post code)
4. customer_city: The name of the city the customer lives in (Brazilian)
5. customer_state: The state which the customer lives in (Brazilian)



2. Data

olist geolocation dataset

1. geolocation_zip_code_prefix: The first five digits of the zip code (primary key)
2. geolocation_lat: Latitude
3. geolocation_lng: Longitude
4. geolocation_city: The name of the city the customer lives in (Brazilian)
5. geolocation_state: The state which the customer lives in (Brazilian)



2. Data

olist order items dataset

1. order_id: Unique ID of the order (foreign key)
2. order_item_id: Each item purchased in the same order (primary key)
3. product_id: The unique ID of each product purchased (foreign key)
4. seller_id: The unique ID of the seller (foreign key)
5. shipping_limit_date: The date the seller will send the order to the logistic partner
6. price: Item price
7. freight_value: The item freight value (if an order has multiple items the freight value is split between them)



2. Data

olist payments dataset

1. order_id: Unique ID of the order (foreign key)
2. payments_sequential: A customer may pay with more than one method. This is the sequence of payments.
3. payment_type: Method of payment
4. payment_installments: If the payment is in instalments (multiple payments over time) this is the number of installments.
5. payment_value: Transaction value

2. Data

olist_order_dataset

1. order_id: Unique ID of the order (primary key)
2. customer_id: An ID unique to each order. This provides access to the unique customer ID in the customer dataset
3. order_status: Delivered, shipped, etc.
4. order_purchase_timestamp: The time of the purchase.
5. order_approved_at: The time the payment was authorised
6. order_delivered_carrier_date: When the order was posted
7. order_estimated_delivery_date: Show the date the customer was advised that the order would be delivered at the time of purchase



2. Data

olist product dataset

1. product_id: Unique ID of the product (primary key)
2. product_category_name: Name of the product category
3. product_name_lenght: Number of characters extracted from the product name
4. product_description_lenght: Number of characters extracted from the product description
5. product_photos_qty: Number of product photos online
6. product_weight_g: Product weight in grams
7. product_length_cm: Product length in centimetres
8. product_width_cm: Product width in centimetres



2. Data

olist sellers dataset

1. seller_id: Unique ID of the seller (primary key)
2. seller_zip_code_prefix: The first five digits of the seller's zip code (post code)
3. seller_city: City where the seller is based
4. seller_state: State where the seller is based



2. Data

olist sellers dataset

1. product_category_name: Product category name in Portuguese
2. product_category_name_english: Product category name in English

WARNINGS!!!!

1. An order may have multiple items
2. Each item may be fulfilled by different sellers
3. All text identifying stores and partners have been replaced with Game of Thrones great houses



3. Task

The business wants to have more reviews on their website, but they particularly want good reviews! They want to be able to target customers who otherwise would not write a review with an email and special offer prompting them to do so. However, they want this to be targeted at the “right” kind of customers.

You will need to build one models that can predict the customer review score (1-5*). You can treat it as either binary task, multi-classification or regression task. But remember your client want more good reviews.

Although not much cleaning is required, you will need to merge some datasets. Also think hard about the features you want to include and any aggregation or transformations you want to perform.