Warwick Business School

Data Science & Generative AI

Dr Michael Mortenson
Associate Professor (Reader)
michael.mortenson@wbs.ac.uk

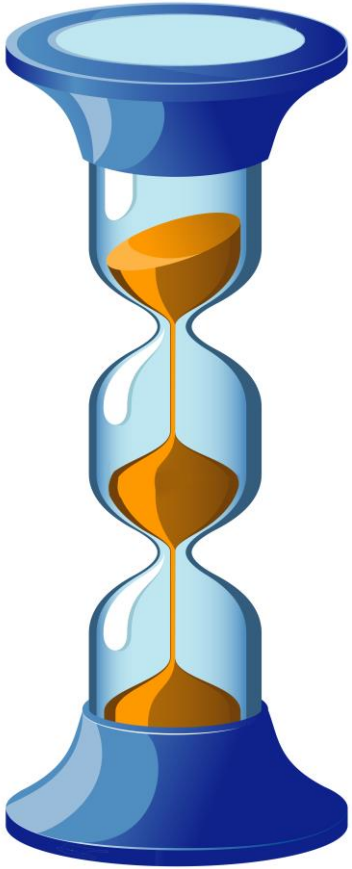
Session 2: Data Cleaning and Feature Engineering (Seminar)

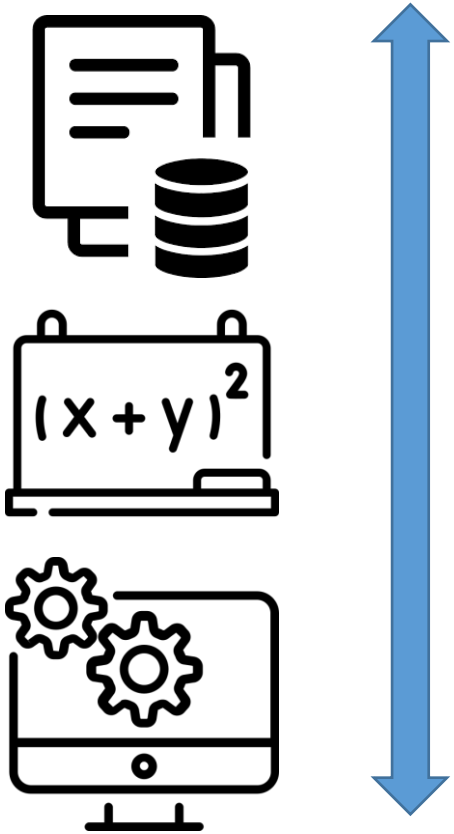
1.1 A Data Supply Chain

Data Processing, Engineering & Storage

Data Analytics and Modelling

Implementation / Decision Making





1.2 From Data to Features

"Feature engineering is the process of transforming raw data into relevant information for use by machine learning models. […] Because model performance largely rests on the quality of data used during training, feature engineering is a crucial preprocessing technique that requires selecting the most relevant aspects of raw training data for both the predictive task and model type under consideration"

Murel J (2024). What is Feature Engineering?
https://www.ibm.com/topics/feature-engineering

1.3 ANNOUNCEMENT!

- Unfortunately we cannot run next week's lecture here in person for operational reasons.
- I apologise unreservedly for this.
- Instead we will run the lecture as a synchronous online session (via Kaltura) on **Friday 24th at 09:00**.
- If this presents significant challenges for anyone I am more than happy to also record the lecture, and release it on Monday, as well as running the synchronous session Friday.

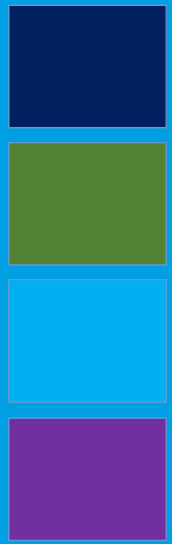
Session Structure

Introduction

Data Cleaning Exercise

Feature Engineering Exercise

Discussion



2.1 Data Cleaning

2.1 Data Cleaning

- Please access the file “Data2Clean.xlsx” from my.wbs.
- Our task is twofold:
  1. Identify the potential issues in the data;
  2. Think about how we might clean these data. (You do not need to actually perform any cleaning steps just consider what you might do in practice).

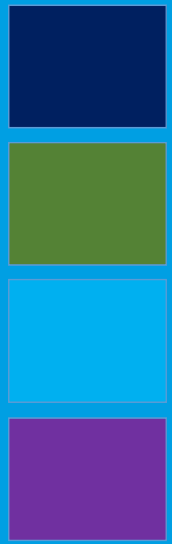
Session Structure

Introduction

Data Cleaning Exercise

Feature Engineering Exercise

Discussion



3.1 Data-driven Feature Engineering

- In data-driven approaches to feature engineering, we use statistical and programming methods to transform data based upon the values within it. For example:
  - Converting Boolean values such as `"DEAD"`, `"ALIVE"` to `0` and `1`.
  - Converting ordered lists to a dictionary of values. E.g. `["High School", "UG", "PG", "PHD"]` to `[0, 1, 2, 3]`.
  - Removing or modifying outliers in the data.
  - Dealing with missing values in the data.
  - Algorithmic methods such as dimension reduction (e.g. principal component analysis) or clustering (e.g. K-means).

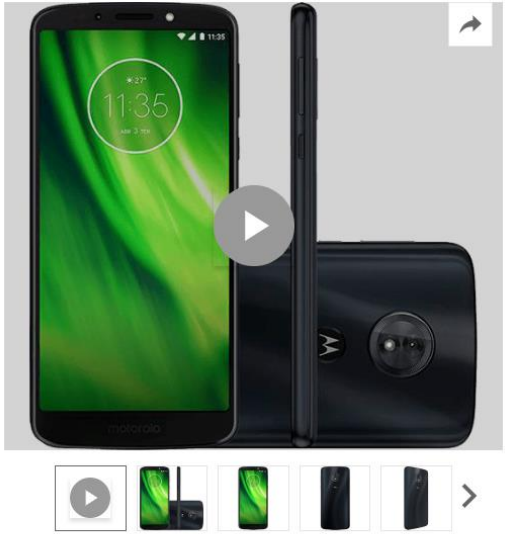
3.2 Theory-driven Feature Engineering

Theory-driven feature engineering involves injecting subject-matter expertise into raw data. This is usually achieved through:

- ✓ Facilitations/brainstorming workshops;
- ✓ Literature reviews;
- ✓ Contextualising data items.

Question: What data items may be predictive of whether a student will pass a module?

3.3 Exercise: Company background



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

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**

vendido e entregue por **olist**

R$ 1.299,00
10x de R$ 129,90 s/ juros

comprar

Corra! Temos apenas 5 no estoque

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

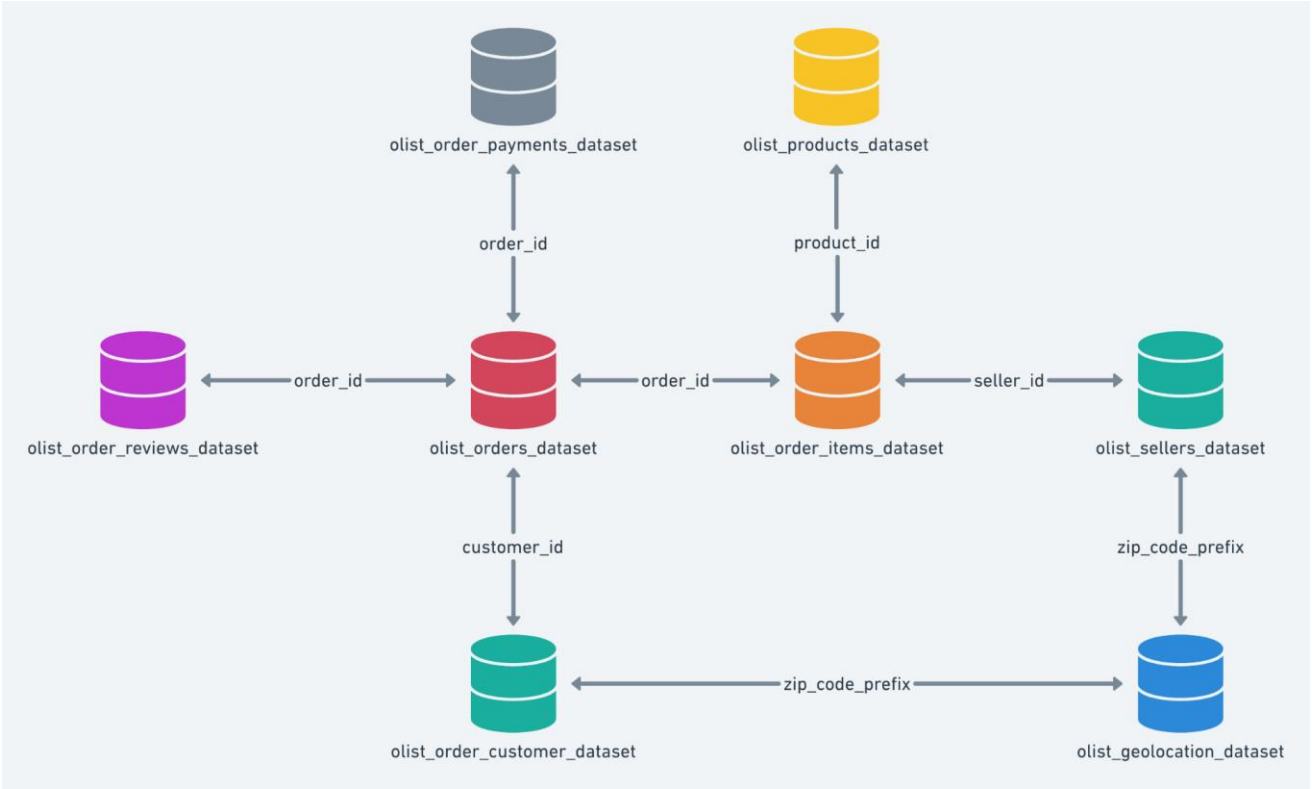
formas de parcelamento

:) Este produto é vendido por uma loja parceira.

3.4 Scenario

- The company has requested you to develop a model to predict which customers will leave positive reviews.
- This model will be used to then specifically target/incentivise these customers to write a review. (Maintaining a set of positive reviews is very important in online retail).
- For this exercise, we are looking through the data in order to identify which features may be predictive of this. You do not need to perform any feature engineering (data-driven or theory-driven) but to think what features you can create which may be predictive.

3.5 Dataset


olist_order_payments_dataset
olist_products_dataset
order_id
product_id
olist_order_reviews_dataset
order_id
olist_orders_dataset
order_id
olist_order_items_dataset
seller_id
olist_sellers_dataset
customer_id
zip_code_prefix
olist_order_customer_dataset
zip_code_prefix
olist_geolocation_dataset

3.5 Dataset

olist_order_reviews_dataset

1. review_id: Unique review ID
2. order_id: Unique order ID
3. review_score: 1-5*
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

3.5 Dataset

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
2. customer_unique_id: Unique ID (one per customer) linked to their profile and order history
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)

3.5 Dataset

olist_geolocation_dataset

1. geolocation_zip_code_prefix: The first five digits of the zip code
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)

3.5 Dataset

olist_order_dataset

1. order_id: Unique ID of the order
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

3.5 Dataset

olist_order_items_dataset

1. order_id: Unique ID of the order
2. order_item_id: Each item purchased in the same order
3. product_id: The unique ID of each product purchased
4. seller_id: The unique ID of the seller
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)

3.5 Dataset

olist_payments_dataset

1. order_id: Unique ID of the order
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 instalments.
5. payment_value: Transaction value

3.5 Dataset

olist_product_dataset

1. product_id: Unique ID of the product
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

3.5 Dataset

olist_sellers_dataset

1. seller_id: Unique ID of the seller
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

3.5 Dataset

WARNINGS!!!!!

1. The *customer_id* is not unique – *customer_unique_id* is. Each customer can have multiple *customer_id* records (1x per transaction)
2. An order may have multiple items
3. Each item may be fulfilled by different sellers
4. All text identifying stores and partners have been replaced with Game of Thrones great houses