

# When features go missing, Bayes' comes to the rescue

**Narendra Mukherjee**

Machine Learning Scientist, Tripadvisor

## About myself

- Machine Learning Scientist in Tripadvisor's B2C data science team
- Work in a variety of domains from sort & recommendations to NLP
- Previously: PhD in Neuroscience from Brandeis University; studied taste processing in the brain
- Long time Bayesian
- Extras: avid bicyclist, >10k miles logged in the Boston/Massachusetts area
- [narendramukherjee.github.io](https://narendramukherjee.github.io)

# About Tripadvisor

TRAVEL NOTICE: Learn more about COVID-19



+ Post Alerts Trips Sign in

- Hotels
- Vacation Rentals
- Things to Do
- Restaurants
- Write a review
- Travel Forums
- More

Find the good out there.

Where to?



Keep Planning



A good product sort is essential for a smooth user experience

# A good product sort is essential for a smooth user experience

Amsterdam: Tours and Tickets

When are you travelling?

All Things to Do **Tours** Day Trips Outdoor Activities Activities Concerts & Shows Food & Drink More

**Tours and Tickets**

- Featured Tours and Tickets (2)
- Local Experiences (2)
- Recommended Experiences (7)
- Classes, Cooking & Water Tours (6)
- Cultural & Theme Tours (3)
- Day Trips & Excursions (2)
- Food, Wine & Spirits (4)

More

**Price**

- \$0 - \$20 (14)
- \$20 - \$40 (29)
- \$40 - \$60 (22)
- \$60 - \$80 (14)
- \$80 - \$100 (4)

**Language**

- English (14)
- French (4)
- German (4)
- Italian (2)
- Japanese (0)
- Spanish (0)

**Popular Attractions**


- Anne Frank House (4)
- Anne Frank House (4)
- The Jordaan (4)
- Rijksmuseum (4)
- Amsterdam Canal Ring (3)
- English Brug (4)
- Night Light Boat (2)

More

**Specials**


- Special Offer (4)
- Likely to Sell Out (3)

Sort by: **Default**  Lowest Price  None

 **Amsterdam Canal Cruise in Luxury River Boat - Small Group - Fresh**  **from \$21.47**


Experience the beauty of Amsterdam's canals by getting on this scenic cruise. No need to worry about obstructed views—the open-air boat has no windows or walls, making it ideal for taking photos of landmarks. [View more](#)

**Popular Booked by 8,438 travelers**

 **Amsterdam Canal Cruise in Classic River Boat With Drinks & Dutch**  **from \$34.59**


Cruise down the famous canals of Amsterdam during this 75-minute boat tour. Also, observe the classic wooden wooden canal-side old buildings and enjoy the beautiful cityscape from the boat. [View more](#)

**Popular Booked by 2,438 travelers**

 **Open boat tour - All drinks included - 60 Minutes - Live guide**  **from \$27.20**

Amsterdam is best explored by water, and this affordable boat cruise gives you an ideal view of canal-side attractions like the Anne Frank House and the Herengracht Museum. The boat is small size and comfortable. [View more](#)

**Popular Booked by 5,287 travelers**

 **Luxury Canal Tour with Local Live Guide - Sightseeing - Free**  **from \$21.47**

We depart from the Rijksmuseum and sail through the beautiful area that includes the Church in the Boat, central churches and will discover the Amsterdam and the beauty of the canals. The view from our boat. [View more](#)

**Popular Booked by 5,287 travelers**

# A good product sort is essential for a smooth user experience

### Amsterdam: Tours and Tickets

When are you traveling?  Start Date  End Date

All Things to Do **Tours** Day Trips Outdoor Activities Attractions Concerts & Shows Food & Drink More

#### Tours and Tickets

- Traditional Tours and Tickets (52)
- Local Experiences (22)
- Recommended Experiences (78)
- Cruises, Boating & Water Tours (76)
- Outdoor & Theme Tours (76)
- Day Trips & Excursions (22)
- Food, Wine & Spirits (4)

More

#### Price

- \$0 - \$25 (48)
- \$25 - \$50 (29)
- \$50 - \$100 (25)
- \$100 - \$200 (16)
- \$200 - \$400

#### Language

- English (74)
- French (24)
- Spanish (16)
- Italian (10)
- Japanese (8)
- Swedish (3)

#### Popular Attractions

- Canal Boats (3)
- Anne Frank House (4)
- The Jordaan (3)
- Rijksmuseum (1)
- Amsterdam Canal Ring (1)
- English Bridge (1)
- Night Light Boat (1)

More

#### Species

- Based On (1)
- Likely to Sell Out (1)

Sort by: **Default**  Lowest Price  None

#### Amsterdam Canal Cruise in Luxury River Boat - Small Group - Fresh

★★★★★ 1,586 reviews

Experience the beauty of Amsterdam's canals by getting on this scenic cruise. No need to worry about obstructed views—the open-air boat has windows on both sides, making it ideal for taking photos of top landmarks.

Price: small size

- Booking safety insurance
- Free
- Private Amsterdam

**Popular Booked by 8,438 travelers**

from **\$21.47**

#### Amsterdam Canal Cruise in Classic River Boat With Drinks & Dutch

★★★★★ 171 reviews

Cruise down the famous canals of Amsterdam during this 75 minute boat tour that offers the classic wooden canal boat and all deck and enjoy the beautiful city on your way. See landmarks while relaxing on Dutch mead.

- Booking safety insurance
- 12 hours
- Private Amsterdam

**Popular Booked by 1,418 travelers**

from **\$34.59**

#### Open boat tour - All drinks included - 60 Minutes - Live guide

★★★★★ 17 reviews

Amsterdam is best explored by water, and this affordable boat cruise gives you an ideal view of popular attractions like the Anne Frank House and Heritage Museum. The boat is small size and waterborne.

- Booking safety insurance
- Free
- Private Amsterdam

**Popular Booked by 1,877 travelers**

from **\$27.20**

#### Luxury Canal Tour with Local Live Guide - Sightseeing - Free

★★★★★ 128 reviews

We depart from the Rijksmuseum and sail through a beautiful area that includes the Church in the Blue, creating a fantastic view of the city of Amsterdam and the beauty of the canals. The view from our boat.

- Booking safety insurance
- Free
- Private Amsterdam

from **\$21.47**

### Amsterdam Hotels and Places to Stay

Search for

1,218 properties in Amsterdam

Sort by: **Best Value**

#### COVID-19

- Properties taking safety measures

#### Deals

- Free cancellation
- Reserve now, pay at stay
- Properties with special offers

#### Price

\$0 - \$100

Price per night

#### Popular

- Breakfast included 63
- 4 stars 214
- 4+ stars 116
- Free WiFi 437

#### Property types

- Hotels 132
- B&Bs & Inns 10
- Inns 10
- Cottages 21

#### Room types

- New Year Deals
- 12 hours 407
- Breakfast included 10
- Free 10
- Free parking 10
- Stay all 10

#### Distance from

25 mi

#### Live saving money!

the search up to 20% sites to help you save up to 20%

#### Embassade Hotel

★★★★★ 1,172 reviews

Best Value of 1,218 places to stay in Amsterdam

from **\$141**

#### NH Collection Amsterdam Grand Hotel Krasnapolsky

★★★★★ 2,776 reviews

Best Value of 1,218 places to stay in Amsterdam

from **\$164**

#### Hotel Jakarta Amsterdam

★★★★★ 1,100 reviews

Best Value of 1,218 places to stay in Amsterdam

from **\$164**

#### Van der Valk Hotel Amsterdam-Anstel

★★★★★ 707 reviews

Best Value of 1,218 places to stay in Amsterdam

from **\$86**

# A good product sort is essential for a smooth user experience

### Amsterdam: Tours and Tickets

When are you traveling?  Start Date  End Date  Search

All Things to Do  Tours  Day Trips  Outdoor Activities  Adventures  Concerts & Shows  Food & Drink  More   Viewmap

#### Tours and Tickets

- Traditional Tours and Tickets (52)
- Local Experiences (22)
- Classes, Cooking & Water Tours (9)
- Outdoor & Theme Tours (76)
- Day Trips & Excursions (23)
- Tours, Walks & Sightseeing (34)

More

#### Price

- \$0 - \$25 (148)
- \$25 - \$50 (226)
- \$50 - \$100 (228)
- \$100 - \$200 (162)
- \$200 - \$400

#### Language

- English (247)
- Hindi (24)
- French (60)
- Italian (30)
- Spanish (28)
- German (20)

#### Popular Attractions

- Anne Frank House (14)
- The Jordaan (24)
- Rijksmuseum (14)
- Amsterdam Canal Ring (11)
- Singel Brug (6)
- The Light Tower (2)

More

#### Species

- Based On (1)
- Likely to See (0)

Sort by: **Default**  Lowest Price  Name

**Amsterdam Canal Cruise in Luxury River Boat - Small Group - 90min**  **from \$21.47**

Experience the beauty of Amsterdam's canals by getting on this scenic cruise. No need to worry about obstructed views - the open-air boat has no windows or walls, making it clear for taking photos of top landmarks.

Free, small size

- Taking safety measures
- Free cancellation
- Popular Booking by 6,436 travelers

**Amsterdam Canal Cruise in Classic River Boat With Drinks & Dutch**  **from \$34.59**

Cruise over the famous canals of Amsterdam during this 75-minute canal tour. This classic river cruise includes a complimentary soft drink and enjoy the beautiful city from your pier. See landmarks while relaxing on the water.

- Taking safety measures
- Free cancellation
- Popular Booking by 2,438 travelers

**Open boat tour - All drinks included - 60 Minutes - Live guide**  **from \$27.20**

Amsterdam is best explored by water, and this affordable boat cruise gives you a central view of canal-side attractions like the Anne Frank House and Rembrandt Museum. The boat's small size and walkability is ideal.

- Taking safety measures
- Free cancellation
- Popular Booking by 3,877 travelers

**Luxury Canal Tour with Local Live Guide - Sightseeing - 90min**  **from \$21.47**

We depart from the Rijksmuseum and sail through the most scenic stretch of Amsterdam's canals in the Blue canal. You'll see the view from the boat, including:

- Taking safety measures
- Free cancellation
- Popular Booking by 3,877 travelers

### Amsterdam Hotels and Places to Stay

Check in  Check out  Rooms  Adults  Children

1,218 properties in Amsterdam

Live using money? The search up to 200 cities to help you start up to 20%

**Amsterdam**

**COVID-19**

- Properties taking safety measures

**Deals**

- Free cancellation
- Reserve now, pay at stay
- Properties with special offers

**Price**

\$0 - \$100

Price per night

**Popularity**

- Not reviewed included 60
- 5 stars 216
- 4 stars 216
- 3 stars 401

**Property types**

- Hotels 132
- B&Bs & Inns 60
- Hostels 21
- Clinics 21

**Stays more**

- New location hotels 0

**Amenities**

- Free WiFi 401
- Breakfast included 60
- Pool 16
- Free parking 16

**View all**

**Distance from**

**Amsterdam**

**Amboosde Hotel**  **from \$141**

4.5 (1,172 reviews)

4.5 Best Value of 128 properties in Amsterdam

- Free WiFi
- Breakfast
- Airport shuttle
- Pet friendly welcome

View all deals from \$141

**NH Collection Amsterdam Grand Hotel Krasnapolsky**  **from \$164**

4.5 (1,776 reviews)

4.5 Best Value of 128 properties in Amsterdam

- Free WiFi
- Breakfast
- Airport shuttle
- Taking safety measures
- Pet friendly welcome

View all deals from \$164

**Hotel Jakarta Amsterdam**  **from \$164**

4.5 (1,135 reviews)

4.5 Best Value of 128 properties in Amsterdam

- Free WiFi
- Breakfast
- Airport shuttle
- Taking safety measures

View all deals from \$164

**Van der Valk Hotel Amsterdam-Amstel**  **from \$66**

4.5 (707 reviews)

4.5 Best Value of 128 properties in Amsterdam

- Free WiFi
- Pool

View all deals from \$66

### Amsterdam Apartment Rentals

Check in  Check out  Rooms  Adults  Children

648 rentals in Amsterdam

Sort by: **Popularity**  Sort

**Linnaeus Suite: award winning apartment, with roof terrace**  **from \$141**

4.5 (12 reviews)

- Free WiFi
- Breakfast
- Sleeps 2
- Multiple stories

Payment Protection

18 other travelers have booked this property.

**Amsterdam Boutique Apartments Private design suite**  **from \$141**

4.5 (10 reviews)

- Free WiFi
- Sleeps 2
- Multiple stories

Payment Protection

11 other travelers have booked this property.

**DE'Y B&B in the heart of the Jordaan, Amsterdam**  **from \$141**

4.5 (16 reviews)

- Free WiFi
- Sleeps 2
- Multiple stories

Payment Protection

10 other travelers have booked this property.

**Oske Garden View Room**  **from \$141**

4.5 (10 reviews)

- Free WiFi
- Sleeps 2
- Multiple stories

Payment Protection

12 other travelers have booked this property.



# Today's story





## Today's story

- **Task:** Sorting lists of products in Tripadvisor's Experiences business

## Today's story

- **Task:** Sorting lists of products in Tripadvisor's Experiences business
- **Features:** Mix of different types

## Today's story

- **Task:** Sorting lists of products in Tripadvisor's Experiences business
- **Features:** Mix of different types
  - ▶ Product specific (price, star rating, CTR, CVR, etc)

# Today's story

- **Task:** Sorting lists of products in Tripadvisor's Experiences business
- **Features:** Mix of different types
  - ▶ Product specific (price, star rating, CTR, CVR, etc)
  - ▶ User specific (prior browsing history, location, etc)

## Today's story

- **Task:** Sorting lists of products in Tripadvisor's Experiences business
- **Features:** Mix of different types
  - ▶ Product specific (price, star rating, CTR, CVR, etc)
  - ▶ User specific (prior browsing history, location, etc)
  - ▶ Seasonality/time-of-year, etc

## Today's story

- **Task:** Sorting lists of products in Tripadvisor's Experiences business
- **Features:** Mix of different types
  - ▶ Product specific (price, star rating, CTR, CVR, etc)
  - ▶ User specific (prior browsing history, location, etc)
  - ▶ Seasonality/time-of-year, etc
- **Model:** Gradient-boosted trees (XGBoost/LightGBM) that are great at capturing interactions between such diverse features

## Today's story

- **Task:** Sorting lists of products in Tripadvisor's Experiences business
- **Features:** Mix of different types
  - ▶ Product specific (price, star rating, CTR, CVR, etc)
  - ▶ User specific (prior browsing history, location, etc)
  - ▶ Seasonality/time-of-year, etc
- **Model:** Gradient-boosted trees (XGBoost/LightGBM) that are great at capturing interactions between such diverse features
- **Problem:** Missing values in model features

# Problem: Missing values in model features





## Problem: Missing values in model features

- **How** do missing values arise in a sorting task?

## Problem: Missing values in model features

- **How** do missing values arise in a sorting task?
- **Why** do they hurt performance in a sorting task?

## How do missing values arise in a sorting task?

Let's sort some products!

Products	Views	Bookings	CVR	Price (USD)	# Reviews	Rating
Old Faithful	1000	200	0.2	20	500	4.5
Expensive 'n Dazzling	1000	50	0.05	200	600	4.5
New Kid on the Block	100	10	0.1	50	0	???
Out of Season	0	0	???	40	500	4.5

## How do missing values arise in a sorting task?

Let's sort some products!

Products	Views	Bookings	CVR	Price (USD)	# Reviews	Rating
Old Faithful	1000	200	0.2	20	500	4.5
Expensive 'n Dazzling	1000	50	0.05	200	600	4.5
New Kid on the Block	100	10	0.1	50	0	???
Out of Season	0	0	???	40	500	4.5

# How do missing values arise in a sorting task? (contd)



**4-in-1 Cancun Snorkeling Tour: Swim with turtles, reef, statues and...**  
from \$65.00  
307 reviews  
[More Info](#)


Discover four of the Yucatan's top underwater sights in just one morning during a snorkeling tour that's designed with anxious swimmers and first-time snorkelers in mind. A guide is on hand to ensure you feel...[read more](#)

 Taking safety measures  
 3-4 hours  
By: [Total Snorkel](#)  
**Popular: Booked by 971 travelers!**

**Views:** 5000



**Books:** 971

**CVR:** 0.19



**Private Yacht Rental Sea Ray 46ft Cancun 23P3**  
from \$560.00  
5 reviews  
[More Info](#)

We are the one the first Luxury yachts companies in cancon we have a Large fleet of yachts. Our crew will take you to the best places in the area. All our fleet is equipud with twin engines and well maintained...[read more](#)

 Taking safety measures  
 2 hours  
By: [Cancun Yacht Rentals](#)

**Views:** 5000

**Books:** 50

**CVR:** 0.01

# How do missing values arise in a sorting task? (contd)

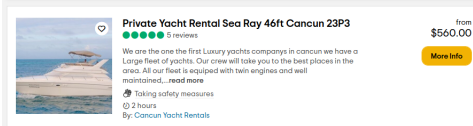


**4-in-1 Cancun Snorkeling Tour: Swim with turtles, reef, statues and...**  
from \$65.00  
307 reviews  
More Info

Discover four of the Yucatan's top underwater sights in just one morning during a snorkeling tour that's designed with anxious swimmers and first-time snorkelers in mind. A guide is on hand to ensure you feel...[read more](#)

🛡️ Taking safety measures  
🕒 3-4 hours  
By: [Total Snorkel](#)  
**Popular: Booked by 971 travelers!**

**Views:** 5000  
**Books:** 971  
**CVR:** 0.19



**Private Yacht Rental Sea Ray 46ft Cancun 23P3**  
from \$560.00  
5 reviews  
More Info

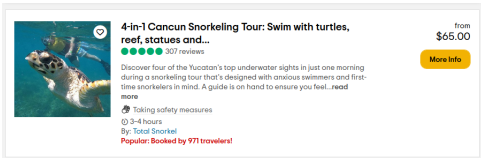
We are the one the first Luxury yachts companies in cancon we have a Large fleet of yachts. Our crew will take you to the best places in the area. All our fleet is equipud with twin engines and well maintained...[read more](#)

🛡️ Taking safety measures  
🕒 2 hours  
By: [Cancun Yacht Rentals](#)

**Views:** 5000  
**Books:** 50  
**CVR:** 0.01

① What if only 100 of the 5000 views were from users who selected a **Yacht Rentals** filter?

# How do missing values arise in a sorting task? (contd)



**4-in-1 Cancun Snorkeling Tour: Swim with turtles, reef, statues and...**  
from \$65.00  
307 reviews  
More Info

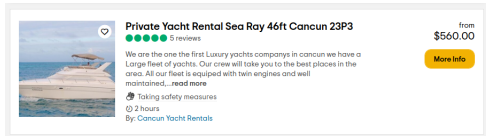
Discover four of the Yucatan's top underwater sights in just one morning during a snorkeling tour that's designed with anxious swimmers and first-time snorkelers in mind. A guide is on hand to ensure you feel...[read more](#)

🛡️ Taking safety measures  
🕒 3-4 hours  
By: [Total Snorkel](#)  
Popular: **Booked by 971 travelers!**

**Views:** 5000

**Books:** 971

**CVR:** 0.19



**Private Yacht Rental Sea Ray 46ft Cancun 23P3**  
from \$560.00  
5 reviews  
More Info

We are the one the first Luxury yachts companies in cancen we have a Large fleet of yachts. Our crew will take you to the best places in the area. All our fleet is equipud with twin engines and well maintained...[read more](#)

🛡️ Taking safety measures  
🕒 2 hours  
By: [Cancun Yacht Rentals](#)

**Views:** 5000

**Books:** 50

**CVR:** 0.01

- 1 What if only 100 of the 5000 views were from users who selected a **Yacht Rentals** filter?
- 2  $CVR = \frac{50}{5000} = 0.01$  or  $CVR = \frac{50}{100} = 0.5$  if the list is being served to a user who has selected the **Yacht Rentals** filter?

# How do missing values arise in a sorting task? (contd)



**4-in-1 Cancun Snorkeling Tour: Swim with turtles, reef, statues and...**  
from \$65.00  
307 reviews  
More Info

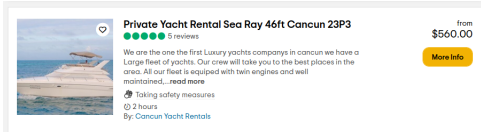
Discover four of the Yucatan's top underwater sights in just one morning during a snorkeling tour that's designed with anxious swimmers and first-time snorkelers in mind. A guide is on hand to ensure you feel...[read more](#)

🛡️ Taking safety measures  
🕒 3-4 hours  
By: Total Snorkel  
Popular: Booked by 971 travelers!

**Views:** 5000

**Books:** 971

**CVR:** 0.19



**Private Yacht Rental Sea Ray 46ft Cancun 23P3**  
from \$560.00  
5 reviews  
More Info

We are the one the first Luxury yachts companies in cancon we have a Large fleet of yachts. Our crew will take you to the best places in the area. All our fleet is equipped with twin engines and well maintained...[read more](#)

🛡️ Taking safety measures  
🕒 2 hours  
By: Cancun Yacht Rentals

**Views:** 5000

**Books:** 50

**CVR:** 0.01

- 1 What if only 100 of the 5000 views were from users who selected a **Yacht Rentals** filter?
- 2  $CVR = \frac{50}{5000} = 0.01$  or  $CVR = \frac{50}{100} = 0.5$  if the list is being served to a user who has selected the **Yacht Rentals** filter?
- 3 CVR is missing for product  $\times$  filter combinations that aren't viewed by users.



# Why do missing values hurt performance in sorting task?

## Why do missing values hurt performance in sorting task?

- 1 Training data is built using user impressions/views  $\Rightarrow$  many products/user features (like filters) are missing from the training data

## Why do missing values hurt performance in sorting task?

- ① Training data is built using user impressions/views  $\Rightarrow$  many products/user features (like filters) are missing from the training data
- ② Model has to serve predictions for all combinations of product  $\times$  user features  $\Rightarrow$  lots of missing features at prediction time and model has to be extrapolate beyond training data

## Why do missing values hurt performance in sorting task?

- 1 Training data is built using user impressions/views  $\Rightarrow$  many products/user features (like filters) are missing from the training data
- 2 Model has to serve predictions for all combinations of product  $\times$  user features  $\Rightarrow$  lots of missing features at prediction time and model has to be extrapolate beyond training data
- 3 Tree-based models *cannot* extrapolate beyond their training data - they usually make constant predictions outside the regime of what they've seen during training

## Why do missing values hurt performance in sorting task?

- 1 Training data is built using user impressions/views  $\Rightarrow$  many products/user features (like filters) are missing from the training data
- 2 Model has to serve predictions for all combinations of product  $\times$  user features  $\Rightarrow$  lots of missing features at prediction time and model has to be extrapolate beyond training data
- 3 Tree-based models *cannot* extrapolate beyond their training data - they usually make constant predictions outside the regime of what they've seen during training
- 4 Bad predictions are *extremely visible* in a list of sorted recommendations - it is not enough to be doing well "on average" as in other ML tasks

So, what can we do with the missing values?

## So, what can we do with the missing values?

- **Drop all rows with missing values**

## So, what can we do with the missing values?

- **Drop all rows with missing values**

- ▶ Sorting Tripadvisor Experiences with large number of features drops significant fraction of rows



## So, what can we do with the missing values?

- **Drop all rows with missing values**

- ▶ Sorting Tripadvisor Experiences with large number of features drops significant fraction of rows

- **Replace missing values with each feature's mean**

## So, what can we do with the missing values?

- **Drop all rows with missing values**

- ▶ Sorting Tripadvisor Experiences with large number of features drops significant fraction of rows

- **Replace missing values with each feature's mean**

- ▶ The mean of a feature is dominated by extreme values, esp when % of missing values is high

## So, what can we do with the missing values?

- **Drop all rows with missing values**

- ▶ Sorting Tripadvisor Experiences with large number of features drops significant fraction of rows

- **Replace missing values with each feature's mean**

- ▶ The mean of a feature is dominated by extreme values, esp when % of missing values is high

- **Let XGBoost/LightGBM natively handle missing values**

## So, what can we do with the missing values?

- **Drop all rows with missing values**

- ▶ Sorting Tripadvisor Experiences with large number of features drops significant fraction of rows

- **Replace missing values with each feature's mean**

- ▶ The mean of a feature is dominated by extreme values, esp when % of missing values is high

- **Let XGBoost/LightGBM natively handle missing values**

- ▶ GB algorithms make best split, "on average", for features with missing values, but disregard previous splits while doing so. Bad when almost all features have missing values. [HousingAnywhere](#) blog post

## So, what can we do with the missing values?

- **Drop all rows with missing values**

- ▶ Sorting Tripadvisor Experiences with large number of features drops significant fraction of rows

- **Replace missing values with each feature's mean**

- ▶ The mean of a feature is dominated by extreme values, esp when % of missing values is high

- **Let XGBoost/LightGBM natively handle missing values**

- ▶ GB algorithms make best split, "on average", for features with missing values, but disregard previous splits while doing so. Bad when almost all features have missing values. [HousingAnywhere](#) [blog post](#)

- **K-nearest-neighbors(KNN) imputation of missing values**

## So, what can we do with the missing values?

- **Drop all rows with missing values**

- ▶ Sorting Tripadvisor Experiences with large number of features drops significant fraction of rows

- **Replace missing values with each feature's mean**

- ▶ The mean of a feature is dominated by extreme values, esp when % of missing values is high

- **Let XGBoost/LightGBM natively handle missing values**

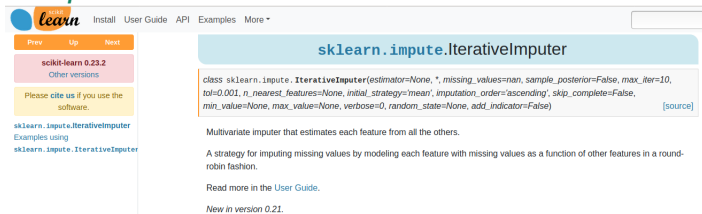
- ▶ GB algorithms make best split, "on average", for features with missing values, but disregard previous splits while doing so. Bad when almost all features have missing values. [HousingAnywhere blog post](#)

- **K-nearest-neighbors(KNN) imputation of missing values**

- ▶ KNN finding is impractical given the strict latency requirements of a recommender system. [AirBnB blog post](#)



# Sklearn *IterativeImputer*



The screenshot shows the sklearn documentation for `sklearn.impute.IterativeImputer`. The page title is `sklearn.impute.IterativeImputer`. The class signature is: `class sklearn.impute.IterativeImputer(estimator=None, *, missing_values=nan, sample_posterior=False, max_iter=10, tol=0.001, n_nearest_features=None, initial_strategy='mean', imputation_order='ascending', skip_complete=False, min_value=None, max_value=None, verbose=0, random_state=None, add_indicator=False)`. The description states: "Multivariate imputer that estimates each feature from all the others. A strategy for imputing missing values by modeling each feature with missing values as a function of other features in a round-robin fashion. Read more in the User Guide. New in version 0.21."

**Require:** Initial imputation of missing values in {model features} by, say, their means  
**repeat**

**for** feature<sub>*i*</sub> ∈ {model features} **do**

feature<sub>*i*</sub><sup>Not Missing</sup>  $\xrightarrow{\text{Train}}$   $f(\{\text{model features}\}_{-i})$

feature<sub>*i*</sub><sup>Missing</sup>  $\xleftarrow{\text{Predict}}$   $f(\{\text{model features}\}_{-i})$

**end for**

**until**  $|\Delta \text{feature}_i^{\text{Missing}}| \leq \text{Tolerance}$

$f$  and Tolerance are defined by the user

## Sklearn *IterativeImputer*

**Require:** Initial imputation of missing values in {model features} by, say, their means

**repeat**

**for**  $\text{feature}_i \in \{\text{model features}\}$  **do**

$\text{feature}_i^{\text{Not Missing}} \xrightarrow{\text{Train}} f(\{\text{model features}\}_{-i})$

$\text{feature}_i^{\text{Missing}} \xleftarrow{\text{Predict}} f(\{\text{model features}\}_{-i})$

**end for**

**until**  $|\Delta \text{feature}_i^{\text{Missing}}| \leq \text{Tolerance}$



## Sklearn *IterativeImputer*

**Require:** Initial imputation of missing values in  $\{\text{model features}\}$  by, say, their means

**repeat**

**for**  $\text{feature}_i \in \{\text{model features}\}$  **do**

$\text{feature}_i^{\text{Not Missing}} \xrightarrow{\text{Train}} f(\{\text{model features}\}_{-i})$

$\text{feature}_i^{\text{Missing}} \xleftarrow{\text{Predict}} f(\{\text{model features}\}_{-i})$

**end for**

**until**  $|\Delta \text{feature}_i^{\text{Missing}}| \leq \text{Tolerance}$

- Takes into account the relationships between the different features (only KNN is able to do this)

## Sklearn *IterativeImputer*

**Require:** Initial imputation of missing values in {model features} by, say, their means  
**repeat**

**for**  $feature_i \in \{\text{model features}\}$  **do**

$feature_i^{\text{Not Missing}} \xrightarrow{\text{Train}} f(\{\text{model features}\}_{-i})$

$feature_i^{\text{Missing}} \xleftarrow{\text{Predict}} f(\{\text{model features}\}_{-i})$

**end for**

**until**  $|\Delta feature_i^{\text{Missing}}| \leq \text{Tolerance}$

- Takes into account the relationships between the different features (only KNN is able to do this)
- Provides a model ( $f$ ) that can scale up the imputation process

## Sklearn *IterativeImputer*

**Require:** Initial imputation of missing values in {model features} by, say, their means

**repeat**

**for**  $\text{feature}_i \in \{\text{model features}\}$  **do**

$\text{feature}_i^{\text{Not Missing}} \xrightarrow{\text{Train}} f(\{\text{model features}\}_{-i})$

$\text{feature}_i^{\text{Missing}} \xleftarrow{\text{Predict}} f(\{\text{model features}\}_{-i})$

**end for**

**until**  $|\Delta \text{feature}_i^{\text{Missing}}| \leq \text{Tolerance}$

- Takes into account the relationships between the different features (only KNN is able to do this)
- Provides a model ( $f$ ) that can scale up the imputation process
- User can modify  $f$  and Tolerance to tradeoff between imputation speed and quality

# Examples from Tripadvisor's Experiences sort



## Examples from Tripadvisor's Experiences sort

**Naive approach:** Replace missing values for CTR/CVR of product  $\times$  filter combinations by -1 (lower than the minimum possible value of 0). Happens when  $\# \text{ Views} = 0$

## Examples from Tripadvisor's Experiences sort

**Naive approach:** Replace missing values for CTR/CVR of product  $\times$  filter combinations by -1 (lower than the minimum possible value of 0). Happens when  $\# \text{ Views} = 0$

Top 5 products for one of the Experiences' pages in London

Products	Views	Bookings	CVR
Tower of London Entrance Ticket Including Crown Jewels and Beefeater Tour	$\gg 0$	$\gg 0$	$> 0$
Christmas Lights Tour of London	$\gg 0$	$\gg 0$	$> 0$
Stonehenge, Windsor Castle, and Bath from London	$\gg 0$	$\gg 0$	$> 0$
Private Walking Tour of London with Brazilian Portuguese Speaking Guide	0	0	-1
Big Bus London Hop-On Hop-Off Tour	$\gg 0$	$\gg 0$	$> 0$

SHAP analysis shows the product's predicted score is driven up *precisely because*  $\text{Views}=0$  and  $\text{CVR}=-1$



## Examples from Tripadvisor's Experiences sort

**Principled approach:** Replace missing values using *IterativeImputer*

Top 5 products for one of the Experiences' pages in London

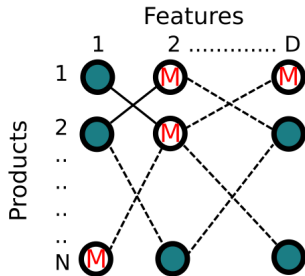
Products	Views	Bookings	CVR
Tower of London Entrance Ticket Including Crown Jewels and Beefeater Tour	» 0	» 0	> 0
Christmas Lights Tour of London	» 0	» 0	> 0
Stonehenge, Windsor Castle, and Bath from London	» 0	» 0	> 0
Big Bus London Hop-On Hop-Off Tour	» 0	» 0	> 0
Warner Bros. Studio: The Making of Harry Potter with Luxury Round-Trip Transport from London	» 0	» 0	> 0

# How/why does it work? Enter Bayes!



# Probabilistic understanding of imputation

Think of a probabilistic graphical model with missing values (M) as latent (hidden) variables



Two extreme views of imputation:

- 1 Forget about the relationships between the features, and model them as independent Gaussians  $\Rightarrow$  replace missing values by MLE, aka, sample mean
- 2 Explicitly model the entire joint distribution of all features (Joint Modelling in statistics). Can only be done under restrictive assumptions



## Can we do better and strike a middle ground?

Instead of the full joint distribution:

$$P(\text{Missing, Observed, Features}) \quad (1)$$

Model the *posterior* distribution of what we don't know (i.e, latent variables=missing values) *given* what we do know (i.e, the observed):

$$P(\text{Missing}|\text{Observed, Features}) \sim P(\text{Observed}|\text{Missing, Features}) \times P(\text{Missing, Features}) \quad (2)$$

Full machinery of Bayesian inference can be applied to this problem!

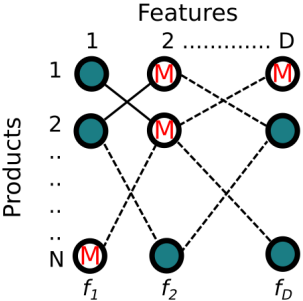
# Posterior inference using MCMC

PyMC3, by default, performs posterior inference over any missing values in the data  
Upcoming tutorial on this at PyMCon: [Missing value tutorial](#)



Task becomes even easier if we can sample from the conditional distributions of the features, i.e,  $P(\text{Feature}_i | \{\text{Features}\}_{-i}) \rightarrow$  Gibbs sampling

# Approximate Bayesian inference

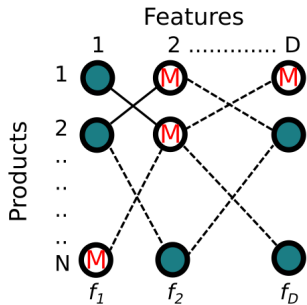


$$f_i \sim P(\text{Feature}_i | \{\text{Features}\}_{-i})$$

Ideal for expectation-maximization (EM): cycle between optimizing the "parameters" of conditional distributions,  $f_i$ , and latent variables (missing values,  $M$ ) iteratively



# Approximate Bayesian inference



$$f_i \sim P(\text{Feature}_i | \{\text{Features}\}_{-i})$$

Ideal for expectation-maximization (EM): cycle between optimizing the "parameters" of conditional distributions,  $f_i$ , and latent variables (missing values,  $M$ ) iteratively  
Guaranteed to find a local maximum of the posterior  $P(\text{Missing} | \text{Observed}, \text{Features})$

## Approximate Bayesian inference: different flavors of EM

- ① **Variational inference ("fully" Bayesian):** Maintain the conditional distributions over the parameters of  $f_i$  and for the latent variables  $M$  at every iteration - all estimates are averaged wrt these conditionals

## Approximate Bayesian inference: different flavors of EM

- ① **Variational inference ("fully" Bayesian):** Maintain the conditional distributions over the parameters of  $f_i$  and for the latent variables  $M$  at every iteration - all estimates are averaged wrt these conditionals
- ② **"Standard" EM ("partially" Bayesian):** Maintains conditional distributions only for the latent variables, not for the parameters of  $f_i$ . eg: Gaussian Mixture Modelling (GMM)

## Approximate Bayesian inference: different flavors of EM

- 1 **Variational inference ("fully" Bayesian):** Maintain the conditional distributions over the parameters of  $f_i$  and for the latent variables  $M$  at every iteration - all estimates are averaged wrt these conditionals
- 2 **"Standard" EM ("partially" Bayesian):** Maintains conditional distributions only for the latent variables, not for the parameters of  $f_i$ . eg: Gaussian Mixture Modelling (GMM)
- 3 **Iterated Conditional Modes (ICM):** Set all random variables to the MAP/MLE estimate at each iteration, no distributions are maintained. eg: K-Means clustering



## IterativeImputer: ICM in the probabilistic graph of missing values

**Require:** Initial imputation of missing values in  $\{\text{model features}\}$  by, say, their means

**repeat**

**for**  $\text{feature}_i \in \{\text{model features}\}$  **do**

$\text{feature}_i^{\text{Not Missing}} \xrightarrow{\text{Train}} f(\{\text{model features}\}_{-i})$

$\text{feature}_i^{\text{Missing}} \xleftarrow{\text{Predict}} f(\{\text{model features}\}_{-i})$

**end for**

**until**  $|\Delta \text{feature}_i^{\text{Missing}}| \leq \text{Tolerance}$

## IterativeImputer: ICM in the probabilistic graph of missing values

**Require:** Initial imputation of missing values in  $\{\text{model features}\}$  by, say, their means (Setting a prior over the missing values)

**repeat**

**for**  $\text{feature}_i \in \{\text{model features}\}$  **do**

$\text{feature}_i^{\text{Not Missing}} \xrightarrow{\text{Train}} f(\{\text{model features}\}_{-i})$

$\text{feature}_i^{\text{Missing}} \xleftarrow{\text{Predict}} f(\{\text{model features}\}_{-i})$

**end for**

**until**  $|\Delta \text{feature}_i^{\text{Missing}}| \leq \text{Tolerance}$

## IterativeImputer: ICM in the probabilistic graph of missing values

**Require:** Initial imputation of missing values in  $\{\text{model features}\}$  by, say, their means (**Setting a prior over the missing values**)

**repeat**

**for**  $\text{feature}_i \in \{\text{model features}\}$  **do**

$\text{feature}_i^{\text{Not Missing}} \xrightarrow{\text{Train}} f(\{\text{model features}\}_{-i})$  (**M step: Set the parameters of the conditionals,  $f_i$  to their MLE/MAP**)

$\text{feature}_i^{\text{Missing}} \xleftarrow{\text{Predict}} f(\{\text{model features}\}_{-i})$

**end for**

**until**  $|\Delta \text{feature}_i^{\text{Missing}}| \leq \text{Tolerance}$

## IterativeImputer: ICM in the probabilistic graph of missing values

**Require:** Initial imputation of missing values in  $\{\text{model features}\}$  by, say, their means (Setting a prior over the missing values)

**repeat**

**for**  $\text{feature}_i \in \{\text{model features}\}$  **do**

$\text{feature}_i^{\text{Not Missing}} \xrightarrow{\text{Train}} f(\{\text{model features}\}_{-i})$  (M step: Set the parameters of the conditionals,  $f_i$  to their MLE/MAP)

$\text{feature}_i^{\text{Missing}} \xleftarrow{\text{Predict}} f(\{\text{model features}\}_{-i})$  ("Hard" E step: Set the latent variables to the MLE/MAP, given the parameters)

**end for**

**until**  $|\Delta \text{feature}_i^{\text{Missing}}| \leq \text{Tolerance}$

## IterativeImputer: ICM in the probabilistic graph of missing values

**Require:** Initial imputation of missing values in  $\{\text{model features}\}$  by, say, their means (**Setting a prior over the missing values**)

**repeat**

**for**  $\text{feature}_i \in \{\text{model features}\}$  **do**

$\text{feature}_i^{\text{Not Missing}} \xrightarrow{\text{Train}} f(\{\text{model features}\}_{-i})$  (**M step: Set the parameters of the conditionals,  $f_i$  to their MLE/MAP**)

$\text{feature}_i^{\text{Missing}} \xleftarrow{\text{Predict}} f(\{\text{model features}\}_{-i})$  (**"Hard" E step: Set the latent variables to the MLE/MAP, given the parameters**)

**end for**

**until**  $|\Delta \text{feature}_i^{\text{Missing}}| \leq \text{Tolerance}$

Known in the statistics community as Multiple Imputation with Chained Equations (MICE)

[StackOverflow](#), [MICE](#)

## A word of caution

- ICM can get stuck in local optima - multiple random restarts are needed
- ML: pick the best (by MLE etc) fit amongst the multiple runs. Statistics: use the restarts to get confidence intervals on the estimates ("multiple" in MICE)
- We got reasonably good results in the missing value imputation problem with even just 1 run

Thank you!!

Visit [narendramukherjee.github.io](https://narendramukherjee.github.io)