ATE estimations from real observational data

This notebook examines the use of Bayesian Networks for estimating Average Treatment Effects (ATE) in Observational Studies within the Neyman-Rubin potential outcome framework from real data: N. Antonio et al. (2019)

Dataset

The data used in this notbook come from "Hotel booking demand datasets" by N. Antonio et al. The data contains 31 variables describing the 104,641 observations. Each observation represents a hotel booking.

We aim to study the impact of assigning a different room to a customer on its likelihood to cancel the reservation. Here, some data preprocessing is needed to match our objectives.

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	deposit_type	agent	company	c
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	No Deposit	NaN	NaN	
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	No Deposit	NaN	NaN	
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	No Deposit	NaN	NaN	
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	No Deposit	304.0	NaN	
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	No Deposit	240.0	NaN	

5 rows × 32 columns

```
Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
       'arrival_date_month', 'arrival_date_week_number',
       'arrival_date_day_of_month', 'meal', 'country', 'market_segment',
       'distribution_channel', 'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'booking_changes', 'deposit_type',
       'agent', 'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date', 'total_stay', 'guests',
       'different_room_assigned'],
      dtype='object')

Number of Null entries: hotel                                  0
is_canceled                          0
lead_time                            0
arrival_date_year                    0
arrival_date_month                   0
arrival_date_week_number             0
arrival_date_day_of_month            0
meal                                 0
country                            488
market_segment                       0
distribution_channel                 0
is_repeated_guest                    0
previous_cancellations               0
previous_bookings_not_canceled       0
booking_changes                      0
deposit_type                         0
agent                            16340
company                         112593
days_in_waiting_list                 0
customer_type                        0
adr                                  0
required_car_parking_spaces          0
total_of_special_requests            0
reservation_status                   0
reservation_status_date              0
total_stay                           0
guests                               4
different_room_assigned              0
dtype: int64

Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_month',
       'arrival_date_week_number', 'meal', 'country', 'market_segment',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'booking_changes', 'deposit_type',
       'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'total_stay', 'guests', 'different_room_assigned'],
      dtype='object')
```

	meal	country	market_segment	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	deposit_type	days_in_waiting_list	customer_type	adr	required_car_parking_spaces	tota
0	BB	PRT	Direct	0	0	0	3	No Deposit	0	Transient	0.0	0	
1	BB	PRT	Direct	0	0	0	4	No Deposit	0	Transient	0.0	0	
2	BB	GBR	Direct	0	0	0	0	No Deposit	0	Transient	75.0	0	
3	BB	GBR	Corporate	0	0	0	0	No Deposit	0	Transient	75.0	0	
4	BB	GBR	Online TA	0	0	0	0	No Deposit	0	Transient	98.0	0	

		hotel	lead_time	arrival_date_month	arrival_date_week_number	meal	country	market_segment	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes
deposit_type	**is_canceled**											
No Deposit	**False**	74947	74947	74947	74947	74947	74947	74947	74947	74947	74947	74947
	True	29694	29694	29694	29694	29694	29694	29694	29694	29694	29694	29694

```
Percentage of customers with different room assignment and cancelation : 0.588294

Percentage of customers with different room assignment and cancelation when there are no booking changes : 0.572001

Percentage of customers with different room assignment and cancelation when there are booking changes : 0.665402
```

We observe that changes in a customer's booking may influence the probability of different room assignments and booking cancellations. We will now investigate whether a causal relationship exists between these factors.

Bayesian Network Preparation

We use `skbn.BNDiscretizer` to discretize the continous variables found in the dataset. The structure of the network will also be provided, `gum.BNLearner` will be used for parameter learning.

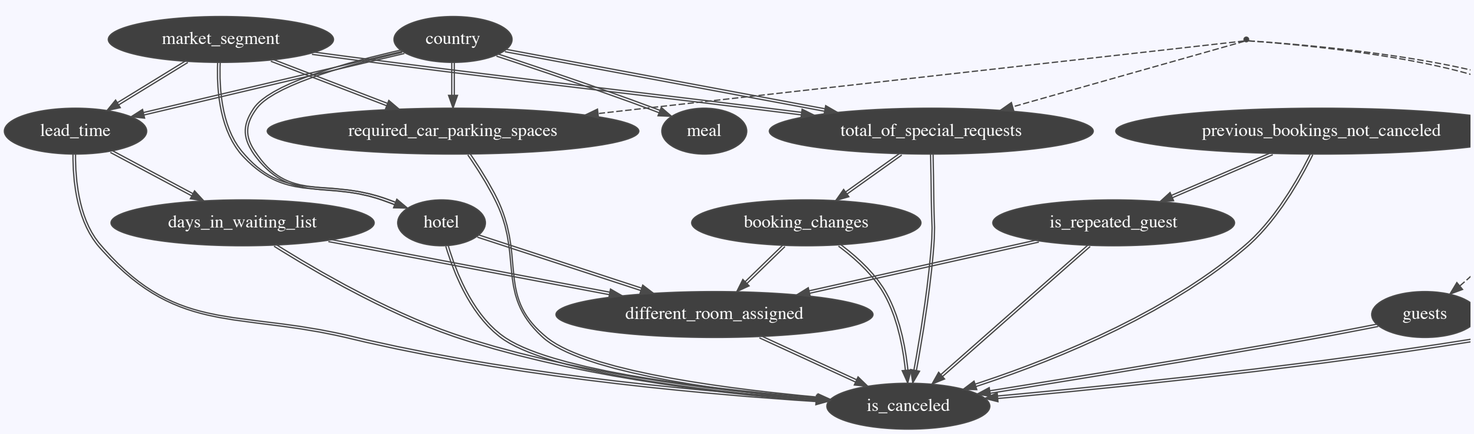
```
lead_time 431
is_repeated_guest 2
previous_bookings_not_canceled 73
booking_changes 21
days_in_waiting_list 99
required_car_parking_spaces 5
total_of_special_requests 6
total_stay 45
guests 15
```

```
/home/thierry/.local/lib/python3.10/site-packages/sklearn/preprocessing/_discretization.py:307: UserWarning: Bins whose width are too small (i.e., <= 1e-8) in feature 0 are removed. Consider decr
easing the number of bins.
  warnings.warn(
```

```
(pyAgrum.BNLearner<double>@0x609ec7fc08e0) Filename       : /tmp/tmp8v165az6.csv
Size           : (104641,15)
Variables      : hotel[2], is_canceled[2], lead_time[5], meal[5], country[177], market_segment[8], is_repeated_guest[2], previous_bookings_not_canceled[5], booking_changes[5], days_in_waiting_lis
t[5], required_car_parking_spaces[5], total_of_special_requests[6], total_stay[4], guests[5], different_room_assigned[2]
Induced types  : False
Missing values : False
Algorithm      : MIIC
Score          : BDeu  (Not used for constraint-based algorithms)
Correction     : NML  (Not used for score-based algorithms)
Prior          : Smoothing
Prior weight   : 0.000000
```

Causal Model

A causal Baysian Network is then created using `csl.CausalModel`, a latent variable being the cause of multiple covariates is also added.



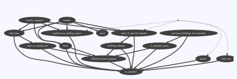
Causal Model

$$P(is_{c}anceled \mid \mathrm{do}(different_{r}oom_{a}ssigned)) = \sum_{booking_{c}hanges,days_{i}n_{w}aiting_{l}ist,hotel,is_{r}epeated_{g}uest} P\left(is_{c}anceled \mid booking_{c}hanges, days_{i}n_{w}aiting_{l}ist, different_{r}oom_{a}ssigned, hotel, is_{r}epeated_{g}uest\right) \cdot P\left(booking_{c}hanges, days_{i}n_{w}aiting_{l}ist, hotel, is_{r}epeated_{g}uest\right)$$

Explanation : backdoor ['hotel', 'is_repeated_guest', 'booking_changes', 'days_in_waiting_list'] found.

is_canceled	
False	True
0.9224	0.0776

Impact



Causal Model

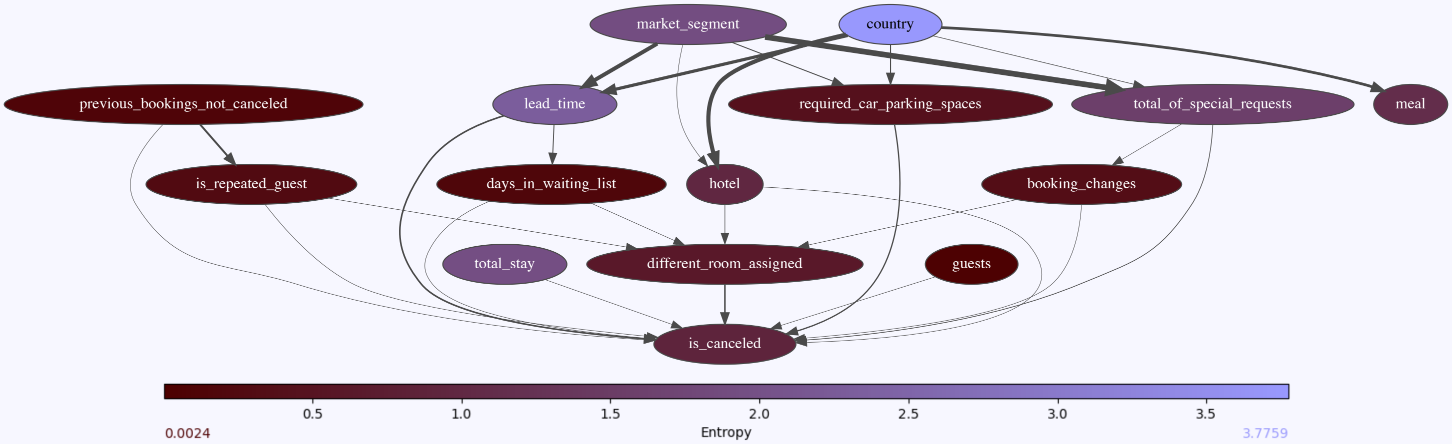
$$P(is_{c}anceled \mid \mathrm{do}(different_{r}oom_{a}ssigned)) = \sum_{booking_{c}hanges,days_{i}n_{w}aiting_{l}ist,hotel,is_{r}epeated_{g}uest} P\left(is_{c}anceled \mid booking_{c}hanges, days_{i}n_{w}aiting_{l}ist, different_{r}oom_{a}ssigned, hotel, is_{r}epeated_{g}uest\right) \cdot P\left(booking_{c}hanges, days_{i}n_{w}aiting_{l}ist, hotel, is_{r}epeated_{g}uest\right)$$

Explanation : backdoor ['hotel', 'is_repeated_guest', 'booking_changes', 'days_in_waiting_list'] found.

is_canceled	
False	True
0.6693	0.3307

Impact

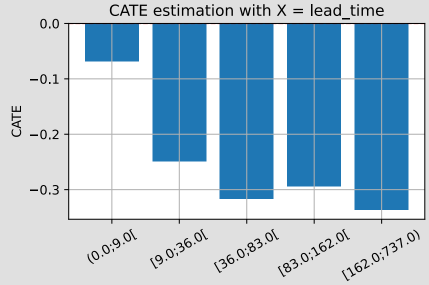
```
[pyAgrum] pyAgrum.lib.notebook.showInformation is deprecated since 0.20.2. Please use pyAgrum.lib.explain.showInfomation instead.
```

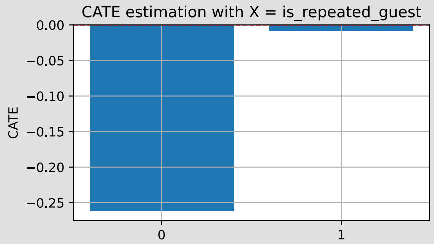


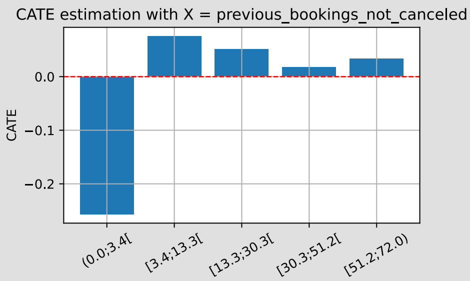

```
ate = -0.2531058345752799
```

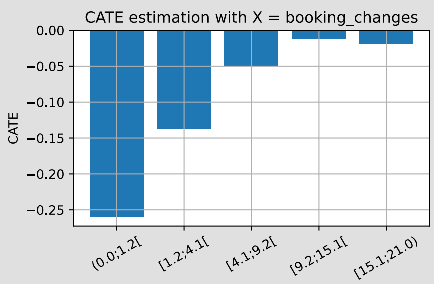
We observe a negative Average Treatment Effect (ATE), which is counterintuitive, as it suggests that assigning a different room reduces the likelihood of a reservation cancellation. To explore this further, we will examine the Conditional Average Treatment Effect (CATE) by conditioning on the covariates to provide additional insights.

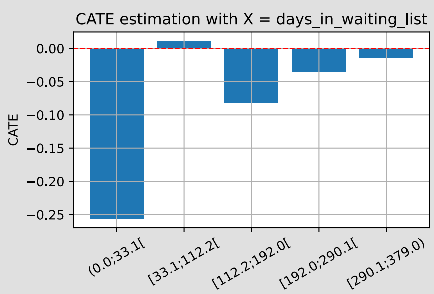
CATE estimations

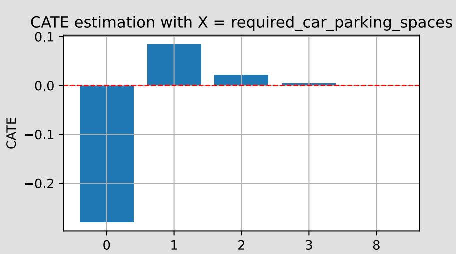

CATE estimation with X = lead_time
CATE
0.0
−0.1
−0.2
−0.3
(0.0;9.0[
[9.0;36.0[
[36.0;83.0[
[83.0;162.0[
[162.0;737.0)

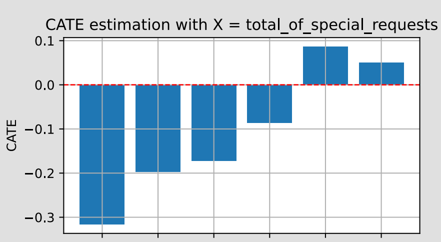

CATE estimation with X = is_repeated_guest
CATE
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
0
1

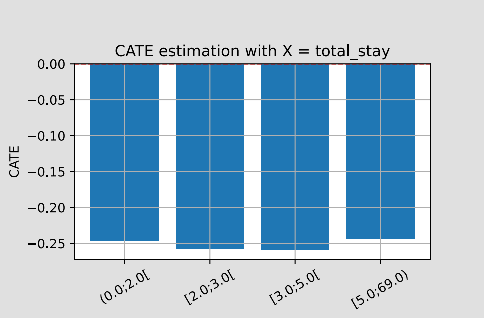

CATE estimation with X = previous_bookings_not_canceled
CATE
0.0
−0.1
−0.2
(0.0;3.4[
[3.4;13.3[
[13.3;30.3[
[30.3;51.2[
[51.2;72.0)

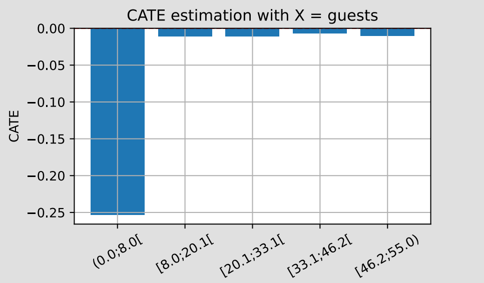

CATE estimation with X = booking_changes
CATE
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
(0.0;1.2[
[1.2;4.1[
[4.1;9.2[
[9.2;15.1[
[15.1;21.0)

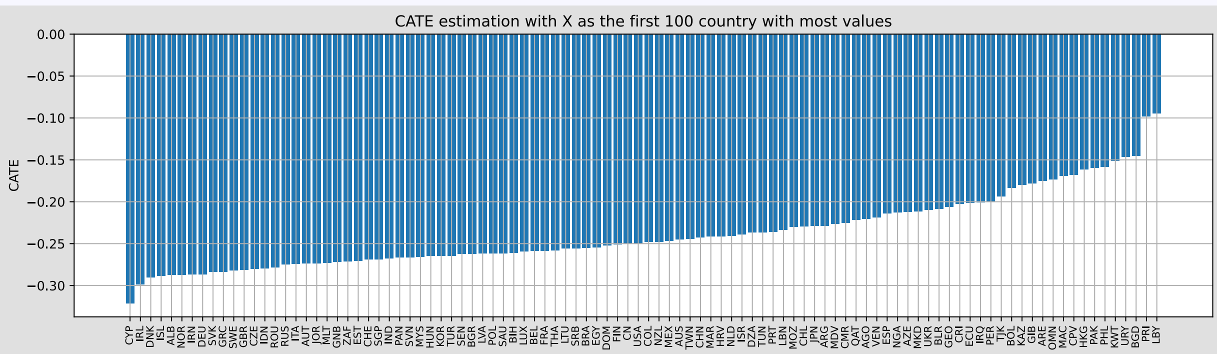

CATE estimation with X = days_in_waiting_list
CATE
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
(0.0;33.1[
[33.1;112.2[
[112.2;192.0[
[192.0;290.1[
[290.1;379.0)

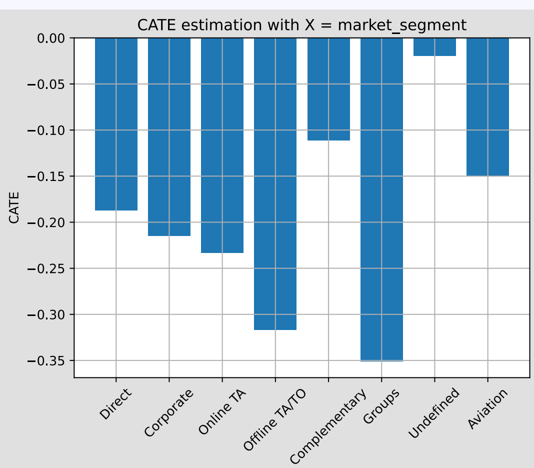

CATE estimation with X = required_car_parking_spaces
CATE
0.1
0.0
−0.1
−0.2
0
1
2
3
8

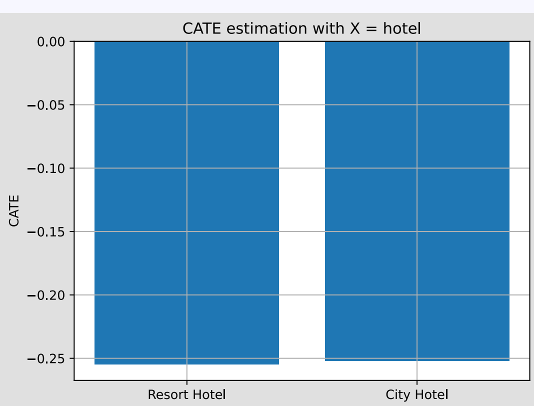

CATE estimation with X = total_of_special_requests
CATE
0.1
0.0
−0.1
−0.2
−0.3
0
1
2
3
4
5


CATE estimation with X = total_stay
CATE
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
(0.0;2.0[
[2.0;3.0[
[3.0;5.0[
[5.0;69.0)


CATE estimation with X = guests
CATE
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
(0.0;8.0[
[8.0;20.1[
[20.1;33.1[
[33.1;46.2[
[46.2;55.0)


CATE estimation with X as the first 100 country with most values
CATE
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.30
CYP IRL DNK ISL ALB NOR IRN DEU SVK GRC SWE GBR CZE IDN ROU RUS ITA AUT JOR MLT GNB ZAF EST CHE SGP IND PAN SVN MYS HUN KOR TUR SEN BGR LVA POL SAU BIH LUX BEL FRA THA LTU SRB BRA EGY DOM FIN CN USA COL NZL MEX AUS TWN CHN MAR HRV NLD ISR DZA TUN PRT LBN MOZ CHL JPN ARG MDV CMR QAT AGO VEN ESP NGA AZE MKD UKR BLR GEO CRI ECU IRQ PER TJK BOL KAZ GIB ARE OMN MAC CPV HKG PAK PHL KWT URY BGD PRI LBY


CATE estimation with X = market_segment
CATE
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
-0.35
Direct
Corporate
Online TA
Offline TA/TO
Complementary
Groups
Undefined
Aviation


CATE estimation with X = hotel
CATE
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
Resort Hotel
City Hotel