Команда

Долинний Денис (КМ-01)
Ганушке́вич Євгеній (КМ-02)
Рижко́ва Дар'я (КМ-02)
Гри́нів Юрій (КМ-02)
Голи́нський Денис (КМ-02)

Про датасет

Датасет створений шляхом SQL-запитів до бази даних Hotel property management systems. Дані відображають готелі у Португалії за 2015-2017 роки. Датасет має 119390 спостережень і 32 змінні.

Змінні

Факторні змінні
Числові змінні

hotel (chr) - тип готелю (H1 = Resort Hotel, H2 = City Hotel).
is_canceled (int) - показує чи була бронь скасована (1), чи ні (0).
lead_time (int) - кількість днів між бронюванням та прибуттям до готелю.
arrival_date_year - рік прибуття.
arrival_date_month(chr) - місяць прибуття.
arrival_date_week_number (int) - номер тижня прибуття .
arrival_date_day_of_month (int) - день прибуття.
stays_in_weekend_nights (int) - кількість вихідних (субота й неділя), які гість перебував або забронював у готелі.
stays_in_week_nights (int) - кількість будніх днів (понеділок - п'ятниця), які гість перебував або забронював у готелі.
adults (int) - кількість дорослих.
children (int) - кількість дітей.
babies (int)- кількість немовлят.
meal (chr) - тип замовленого харчування. Категорії: Undefined/SC – не подають харчування; BB – лише сніданок; HB – сніданок і ще один прийом їжі (зазвичай вечеря); FB – повне харчування (сніданок, обід і вечеря).
country (chr) - країна замовника.Закодовано в ISO 3155–3:2013 форматі.
market_segment (chr) - сегмент ринку. Категорії: "Aviation", "Complementary", "Corporate", "Direct", "Groups", "Offline TA/TO", "Online TA", "Undefined" ("TA" – "Travel Agents", "TO" – "Tour Operators").
distribution_channel (chr)- розподіл бронювання. Категорії: "TA" – "Travel Agents"/"TO" means "Tour Operators", "Corporate", "Direct", "GDS" – Global Distribution System.
is_repeated_guest (int) - бронювання від "старого" гостя (1) чи ні (0).
previous_cancellations (int) - кількість бронювань, які клієнт скасував до поточного бронювання.

previous_bookings_not_canceled (int) - кількість бронювань, які клієнт не скасував до поточного бронювання.
reserved_room_type (chr)- код номеру, який забронювали. Замість позначення наводиться код з міркувань анонімності.
assigned_room_type (chr)- код номеру, призначеного для бронювання. Іноді призначений тип номера відрізняється від типу зарезервованого номера через причини роботи готелю (наприклад, надмірне бронювання) або за запитом клієнта. Замість позначення наводиться код з міркувань анонімності.
booking_changes (int) - Кількість змін/доповнень, внесених до бронювання з моменту введення бронювання до моменту заселення або скасування замовлення.
deposit_type (chr) -зазначає чи вніс клієнт депозит, і якщо так – то який. Категорії: No Deposit, Non Refund, Refundable.
agent (chr)- ID туристичної агенції, що зробила замовлення.
company (chr)- ID компанії, що зробила замовлення або відповідальна за оплату.
days_in_waiting_list (int) - кількість днів, які бронювання було в списку очікування, перш ніж його було підтверджено
customer_type (chr) - тип бронювання. Категорії: Contract, Group, Transient, Transient-party.
adr (num) - середня добова ставка (статистична одиниця, яка показує дохід за номер за окремий період часу). Визначається як сума всіх операцій поділена на кількість ночей.
required_car_parking_spaces (int) - кількість паркувальних місць, які забронював клієнт.
total_of_special_requests (int)- кількість спеціальних запитів клієнта (наприклад двоспальне ліжко або високий поверх)
reservation_status (chr) - останній статус бронювання. Категорії: Canceled, Check-Ou, No-Show.
reservation_status_date (chr) - дата останнього оновлення статусу.

Очистка даних

Створимо нову змінну (all_guests), що включає всіх гостей.
Видалимо 180 рядків, де змінні adults, children і babies одночасно рівні 0 (тобто залишаться спостереження, де all_guests != 0).
Видалемо змінні company та agent, бо вони не несуть ніякої корисної інформації.
Додамо змінну stays_in_nights, яка показуватиме загальну кількість ночей. Бачимо 645 рядків, де stays_in_nights = 0 (тобто stays_in_weekend_nights і stays_in_week_nights одночасно дорівнюють 0). Видаляємо ці рядки.
Додамо змінну all_children, яка показуватиме загальну кількість дітей.
У змінній required_car_parking_spaces є 5 значень, які більше 2. У них бачимо, що максимальна кількість людей 2. Тобто очевидна помилка. Замінимо їх на 2.
У змінній adr є 1 значення, що менше 0, та 1 значення, що більше 550 (5400). Видалемо їх, оскільки їх мало. Також видалимо ті, що більше 400 (їх 7).
У змінних babies та children сумарно 3 викиди, які ми видаляємо.Також видаляємо 16 рядків, де adults > 4.
З lead_time видаляємо значення, які більше 700.
Видаляємо спостереження, де country має значення NULL.

Усі інші змінні мають не такі явні викиди (тобто ми не можемо сказати це помилки чи ні). Отже, доцільно розглядати випадки з ними і без них, і подивитися, який вони мали вплив.
Після очистки дескриптивні характеристики мають вигляд:

	Min.	1st Qu.	Median	Mean	Mean 3rd Qu.	Max.
lead_time	0	19	71	105,4	162	629
stays_in_weekend_nights	0	1	1	0,9372	2	16
stays_in_week_nights	0	1	2	2,522	3	40
adults	0	2	2	1,863	2	4
children	0	0	0	0,1047	0	3
babies	0	0	0	0,007706	0	2
previous_cancellations	0	0	0	0,08698	0	26
previous_bookings_not_canceled	0	0	0	0,01197	0	72
booking_changes	0	0	0	0,2158	0	18
days_in_waiting_list	0	0	0	2,348	0	391
adr	0,26	71	95	103,64	126	397,38
required_car_parking_spaces	0	0	0	0,06185	0	2
total_of_special_requests	0	0	0	0,5712	1	5

Отже, зміни в датасеті: 119390 → 116920 спостережень (видалили 2470 ~ 2,1%)

Додаткова робота з даними

Була створена зміна season, яка показує, у який сезон прибули відвідувачі (теплий (1) чи холодний (0)).
Була створена зміна with_children, яка показує, чи були з відвідувачами діти.
Була створена зміна market_segment_b, яка показує сегмент ринку, до якого належить гість (онлайн – 1, інше – 0).
Була створена зміна distribution_channel_b, як було зроблено бронювання (через TA/TO – 1, інше – 0).
Була створена зміна deposit_type_b, яка показує, чи був зроблений депозит (так – 1, ні – 0).
Була створена зміна with_meal, яка показує, чи була замовлена їжа (так – 1, ні – 0).

Гіпотези

У якості залежної змінної було взято adr (прибуток готелю за 1 номер за 1 ніч).
Висуваємо такі гіпотези щодо коефіцієнтів:

1) $coef_{lead-time} < 0$
2) $coef_{all-nights} < 0$
3) $coef_{booking-changes} > 0$
4) $coef_{all-guests} < 0$
5) $coef_{total-of-special-request} > 0$
6) $coef_{market-segment-b} < 0$

7) $coef_{distribution-channel-b} < 0$
8) $coef_{deposit-type-b} < 0$
9) $coef_{with-meal} > 0$

Також ми очікуємо, що місяць прибуття (arrival_date_month) буде суттєво впливати на adr.

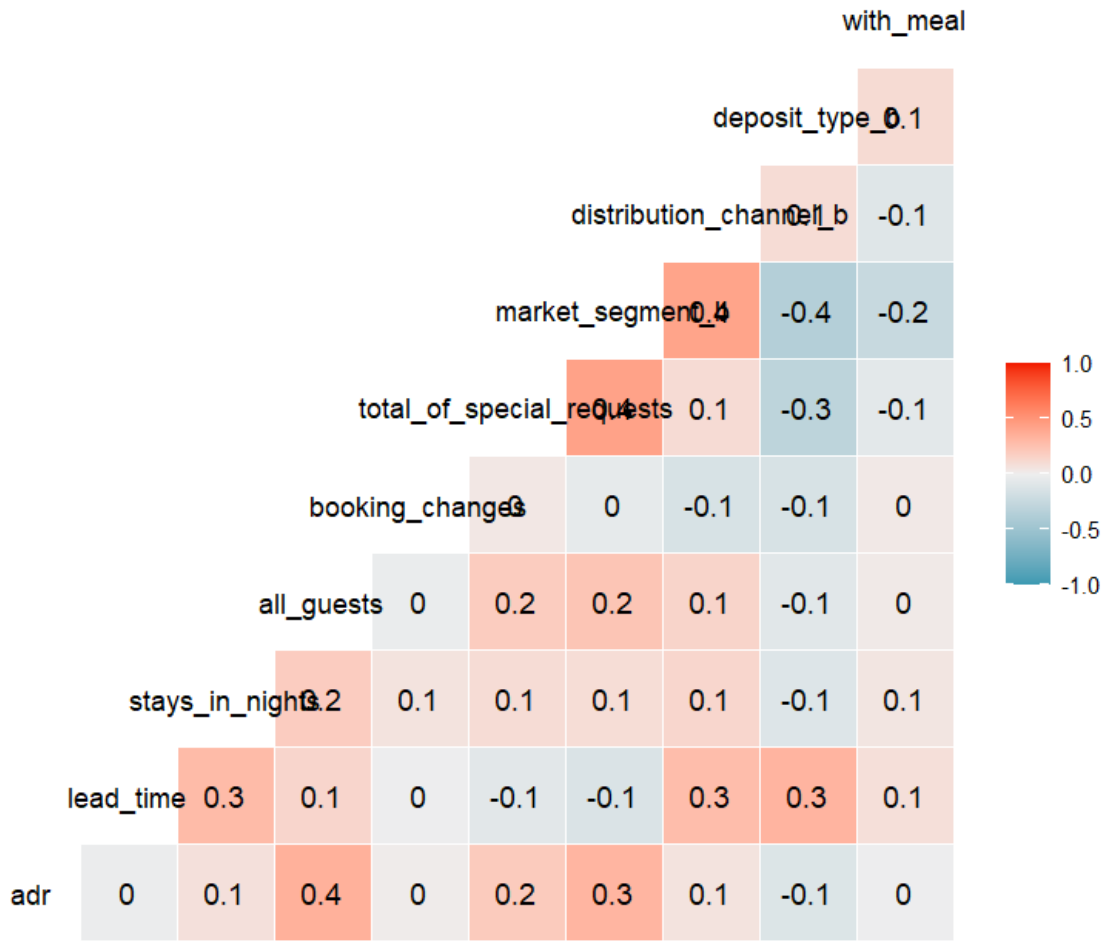
Побудова моделей

Наша основна модель:

Table 1: Regression

	Dependent variable:
	Середнє adr
lead_time	-0.041^{***}
	$(-0.043, -0.039)$
Constant	107.979^{***}
	$(107.591, 108.367)$
Observations	116,920
Adjusted R^2	0.009
Note:	*p<0.1; **p<0.05; ***p<0.01

Можемо побачити, що збільшення lead_time на 1 змінюється adr на 0,041 (що дуже мало) та що коефіцієнт статистично значущий.
Проаналізувавши відповідні середньоквадратичні відхилення, ми побачимо, що збільшення середньоквадратичного відхилення lead_time на 1 зменшує середньоквадратичне відхилення adr на 9%.
Також, дивлячись на R можна сказати, що частка lead_time у варіації adr **дуже** мала.
Додамо контрольні змінні.
Проаналізуємо кореляційну матрицю:



У нас немає змінних, що сильно корелюють між собою тому ми можемо включити усіх їх у нашу модель. Побудуємо 2 нові моделі. Перша включає в себе усі фактори + поліном другого порядку all_guests, що розглядаються, друга - ті, у яких кореляція ≠ 0. Отримані результати:

Table 1: Multiple regression

	Середнє adr		
	(1)	(2)	(3)
lead_time	−0.041***	−0.022***	
	(−0.043, −0.039)	(−0.024, −0.020)	
stays_in_nights		0.237***	−0.005
		(0.133, 0.341)	(−0.107, 0.096)
all_guests		−13.319***	27.927***
		(−14.972, −11.666)	(27.449, 28.405)
I(all_guests^2)		9.077***	
		(8.685, 9.468)	
booking_changes		1.346***	
		(0.911, 1.781)	
total_of_special_requests		3.029***	2.440***
		(2.667, 3.391)	(2.076, 2.805)
market_segment_b		23.852***	24.236***
		(23.256, 24.447)	(23.676, 24.797)
distribution_channel_b		−11.326***	−16.316***
		(−12.098, −10.553)	(−17.063, −15.568)
deposit_type_b		9.033***	5.611***
		(8.360, 9.706)	(4.930, 6.292)
with_meal		8.019***	
		(7.362, 8.676)	
Constant	107.979***	79.863***	48.278***
	(107.591, 108.367)	(79.475, 80.251)	(47.890, 48.667)
Observations	116,920	116,920	116,920
Adjusted R^2	0.009	0.274	0.244
Note:			*p<0.1; **p<0.05; ***p<0.01

Можемо побачити, що для моделі(2) усі коефіцієнти виявилися статистично значущими, але прибравши змінні, кореляція яких ≈ 0, коефіцієнт R^2зменшиться всього на 0,03, тобто вилучення незначущих регресорів було виправдано. Також ми бачимо, що змінна stays_in_nights стала статистично незначущою. Протестувавши це явно (гіпотезу, що коефіцієнт при stays_in_nights = 0), отримаємо:

```
  Res.Df Df      F Pr(>F)
1 116914
2 116913  1 0.011 0.9164
```

p-value дуже велике, тому немає підстав стверджувати, що stays_in_nights справді є статистично значущою. Змінюємо відповідно модель(3). Отримаємо:

Table 1: Multiple regression

	Середнє adr		
	(1)	(2)	(3)
lead_time	−0.041***	−0.022***	
	(−0.043, −0.039)	(−0.024, −0.020)	
stays_in_nights		0.237***	
		(0.133, 0.341)	
all_guests		−13.319***	27.925***
		(−14.972, −11.666)	(27.449, 28.400)
I(all_guests^2)		9.077***	
		(8.685, 9.468)	
booking_changes		1.346***	
		(0.911, 1.781)	
total_of_special_requests		3.029***	2.440***
		(2.667, 3.391)	(2.075, 2.804)
market_segment_b		23.852***	24.239***
		(23.256, 24.447)	(23.681, 24.797)
distribution_channel_b		−11.326***	−16.320***
		(−12.098, −10.553)	(−17.062, −15.578)
deposit_type_b		9.033***	5.616***
		(8.360, 9.706)	(4.944, 6.289)
with_meal		8.019***	
		(7.362, 8.676)	
Constant	107.979***	79.863***	48.267***
	(107.591, 108.367)	(79.475, 80.251)	(47.878, 48.655)
Observations	116,920	116,920	116,920
Adjusted R^2	0.009	0.274	0.244
Note:			*p<0.1; **p<0.05; ***p<0.01

Бачимо, що вилучення stays_in_nights було виправдано.
Побудуємо нові моделі, у яких враховується місяць (базовий квітень).

Table 1: Multiple regression

	Середнє adr		
	(1)	(2)	(3)
factor(arrival_date_month)August	39.924	32.841	35.723
factor(arrival_date_month)December	−17.876	−17.630	−18.606
factor(arrival_date_month)February	−26.624	−24.786	−27.834
factor(arrival_date_month)January	−29.598	−26.345	−29.476
factor(arrival_date_month)July	26.733	21.094	25.528
factor(arrival_date_month)June	16.214	17.117	19.969
factor(arrival_date_month)March	−20.158	−18.117	−19.525
factor(arrival_date_month)May	8.602	11.044	13.035
factor(arrival_date_month)November	−26.125	−20.186	−21.176
factor(arrival_date_month)October	−11.827	−6.874	−4.958
factor(arrival_date_month)September	4.916	9.992	12.576
lead_time			−0.087***
			(−0.089, −0.085)

stays_in_nights			−0.129**
			(−0.233, −0.025)
all_guests		23.082***	23.740***
		(21.430, 24.735)	(22.088, 25.393)
booking_changes			2.701***
			(2.266, 3.136)
total_of_special_requests		1.633***	1.709***
		(1.271, 1.996)	(1.347, 2.071)
market_segment_b		25.248***	22.487***
		(24.652, 25.843)	(21.891, 23.083)
distribution_channel_b		−19.592***	−12.839***
		(−20.364, −18.820)	(−13.611, −12.066)
deposit_type_b		6.736***	14.860***
		(6.063, 7.408)	(14.187, 15.533)
with_meal			9.342***
			(8.684, 9.999)
Constant	101.798***	57.933***	50.802***
	(101.111, 102.486)	(57.245, 58.621)	(50.114, 51.489)
Observations	116,920	116,920	116,920
Adjusted R^2	0.224	0.406	0.438
Note:			*p<0.1; **p<0.05; ***p<0.01

З моделі(1) бачимо, що літні місяці суттєво збільшують adr, а зимові + березень і листопад суттєво зменшують, хоча ці коефіцієнти можуть не бути статистично значущими. Протестувавши гіпотези, що коефіцієнти для всіх місяців з factor(arrival_date_month) = 0 для всіх 3-х моделей, ми отримали, що p-value < 2.2e-16, тобто ці коефіцієнти все ж статистично значущі.

Також ми можемо побачити, що використання всіх коефіцієнтів не є виправданим, R^2 для моделі(2) і моделі(3) відрізняється всього лише на 0,03.

Висновки

1) $coef_{lead-time} < 0$ – підтверджено, хоча коефіцієнт впливає не сильно.
2) $coef_{all-nights} < 0$ – підтверджено, хоча коефіцієнт впливає не сильно.
3) $coef_{booking-changes} > 0$ – підтверджено, хоча статистично незначущий.
4) $coef_{all-guests} < 0$ – спростовано, коефіцієнт суттєво впливає на adr.
5) $coef_{total-of-special-request} > 0$ – підтверджено, хоча коефіцієнт впливає не сильно.
6) $coef_{market-segment-b} < 0$ – спростовано, коефіцієнт суттєво впливає на adr.
7) $coef_{distribution-channel-b} < 0$ – підтверджено, коефіцієнт суттєво впливає на adr.
8) $coef_{deposit-type-b} < 0$ – спростовано, коефіцієнт суттєво впливає на adr.
9) $coef_{with-meal} > 0$ підтверджено, хоча статистично незначущий.

Також, можемо підтвердити, що місяць сильно впливає на adr, причому літні суттєвого його збільшують, а зимові + березень і листопад суттєво зменшують.