Rakamin Academy

Laporan Project

Stage 1

1. Apa saja attributes dan target output dari dataset yang dipilih?

Attributes :

Berikut asumsi fitur yang berdampak terhadap “is_canceled”

- High: hotel, lead_time, arrival_date_year, arrival_date_month, is_repeated_guest, previous_cancellations, previous_bookings_not_canceled, reserved_room_type, assigned_room_type, deposit_type, days_in_waiting_list, customer_type, stays_in_weekend_nights, stays_in_week_nights
- Medium: children, babies, distribution_channel, booking_changes, adr
- Low: arrival_date_week_number, arrival_date_day_of_month,country, meal, adults, market_segment, agent, company, required_car_parking_spaces, total_of_special_requests, reservation_status, reservation_status_date

Target : is_canceled

2. Untuk setiap feature yang disiapkan, apakah sudah dicek distribusinya terhadap variabel target?

Kami sudah melakukan cek distribusi terhadap variabel target ketika pada tahapan univariate analysis. Didapatkan hasil sebagai berikut :

1. Univariate Analysis :

- Dari Hasil Pengamatan didapat :

A) Box Plot (Numerical)

1. Terlihat adanya outlier pada banyak kolom, yaitu `lead_time`, `stays_in_weekend_nights`,`stays_in_week_nights`, `adults`, `children`, `babies`, `previous_cancellations`, `previous_bookings_not_canceled`, `booking_changes`,`days_in_waiting_list`, `adr`, `required_car_parking_spaces`,`total_of_special_requests`.
2. Pada boxplot terlihat ada terdapat banyak kolom yang kemungkinan skewed yaitu semua kolom kecuali `is_canceled`,`arrival_date_week_number`, `arrival_date_day_of_month`,`is_repeated_guest`

B) Distribution Plot (Numerical)

1. Seperti dugaan ketika melihat boxplot, hampir semua kemungkinan skewed kecuali `is_canceled`,`arrival_date_week_number`, `arrival_date_day_of_month`,`is_repeated_guest`. Berarti ada kemungkinan kita perlu melakukan sesuatu pada kolom-kolom tersebut nantinya.
2. Ada 2 kolom yang terlihat mendekati normal yaitu `arrival_date_week_number` dan `arrival_date_day_of_month`.

• Selain melalui visualisasi data, dicek juga nilai skew untuk memastikan dugaan

a) Dari hasil di atas terlihat bahwa tidak ada yang berdistribusi normal (distribusi normal : skew = 0).

b) Hampir semua skew, baik positively maupun negatively skew, kecuali 'arrival date week number' dan 'arrival date day of month’.

c) Ada 2 kolom yang mendekati distribusi normal yaitu 'arrival date week number' dan 'arrival date day of month’.

d) Sehingga dugaan pada pengamatan boxplot dan kde plot terkait skew sudah terbukti.

C) Individual Count Plot (Categorical)

Hasil dari pengamatan di atas di dapat kesimpulan antara lain hampir semua feature categorical terjadi ketimpangan yang cukup tinggi antara lain

1. Hotel = Perbedaan city hotel dan resort hotel mencapai 50%
2. Arrival date month = Pada bulan November - Januari mengalami penurunan yg signifikan dibanding bulan sebelumnya
3. Meal = BB (Bed and Breakfast) sangat berbeda jauh dengan HB (Breakfast and Dinner, SC (Self Catering) dan FB (Breakfast, lunch and dinner)
4. Market_Segment = Market Segment berbeda jauh dengan offline
5. Distribution_Channel = Via TA/TO (Travel Agent, Tour Operator) sangat jauh perbedaannya dengan distribution channel lainnya

6. reserved_room_type = type A sangat berbeda jauh dengan type lain (terdekat perbedaan lebih dari 4 kali
7. assigned_room_type = sama seperti no 6 akan tetapi gap nya semakin kecil karena beberapa type mengalami peningkatan (G, E, D, F, B)
8. Deposit type = No deposit masih tinggi penggunaannya dibanding dengan non refund atau refundable
9. customer type = Transient dengan customer type lain berbeda sangat jauh
10. Untuk features country juga memiliki nilai timpang dimana teringgi adalah turis dari PRT(Portugal) jauh lebih banyak dari country lainnya

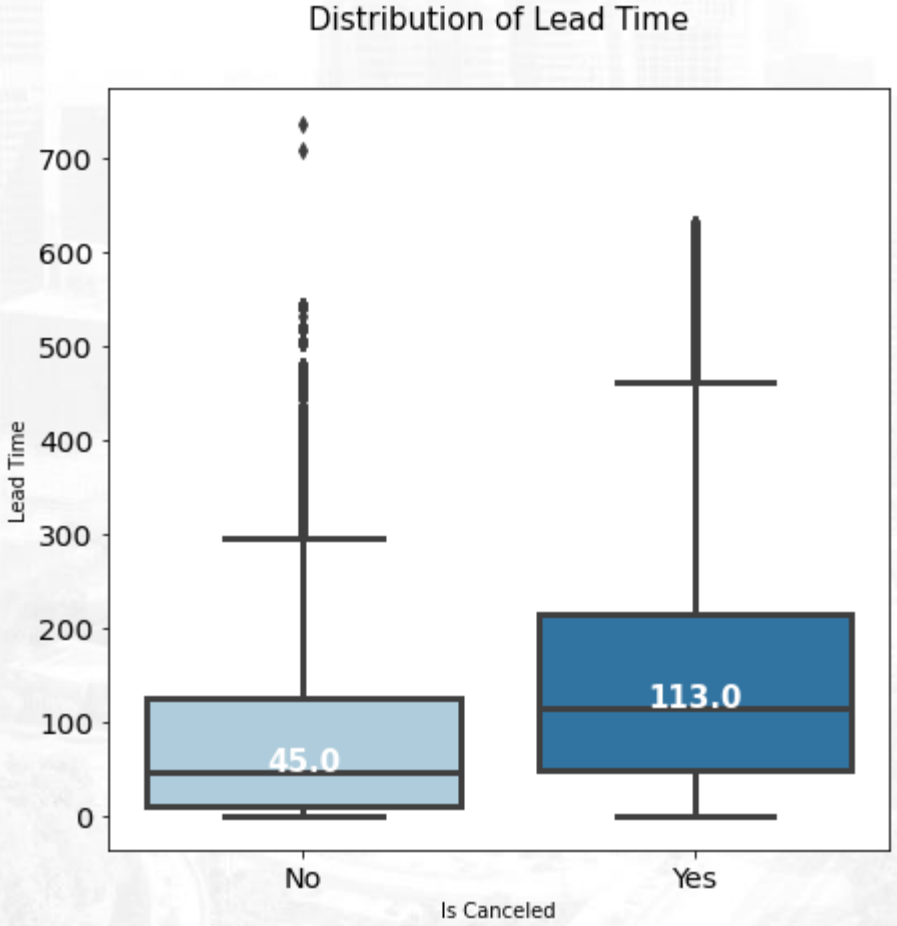
3. Apakah sudah menemukan beberapa insight menarik dari dataset tersebut, termasuk visualisasi yang mendukung serta kaitan dari problem utama yang ingin diselesaikan?

1. Lead Time and Cancelation :

Insight:

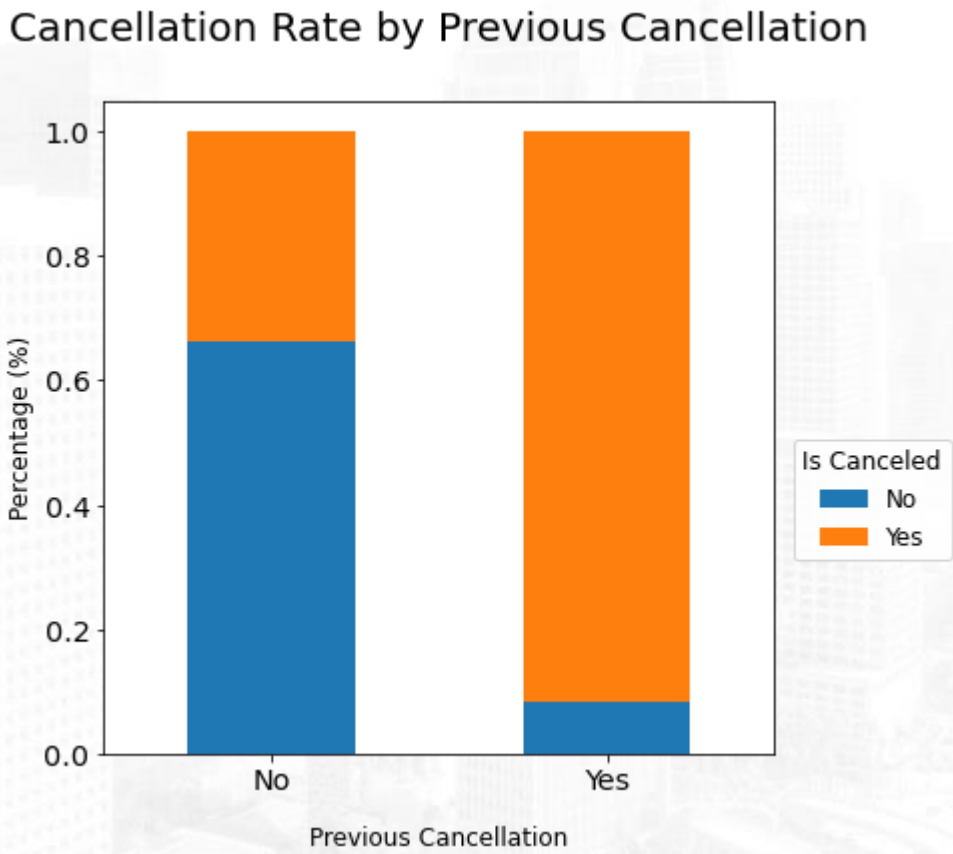
•Semakin lama lead time, semakin besar kemungkinan pelanggan akan membatalkan pemesanan hotel

•Jika jumlah hari antara saat pemesanan dibuat dan perkiraan tanggal kedatangan meningkat, kemungkinan pelanggan memiliki lebih banyak waktu untuk membatalkan reservasi dan lebih banyak waktu untuk keadaan tak terduga yang dapat membatalkan reservasi



2. Previous Cancelation and Cancelation :

Insight:
Pelanggan yang telah membatalkan pemesanan hotel sebelumnya kemungkinan akan membatalkan lagi



3. Repeated Guest and Cancelation :

Insight:
“Repeated Guest" diidentifikasi memiliki peluang pembatalan yang lebih rendah.