PREDICT HOTEL BOOKING CANCELLATIONS IN PORTUGAL

Final Project :

Zalfa Maitsa NR - Batch 17

Table of Contents

DESCRIPTION

DESCRIPTION

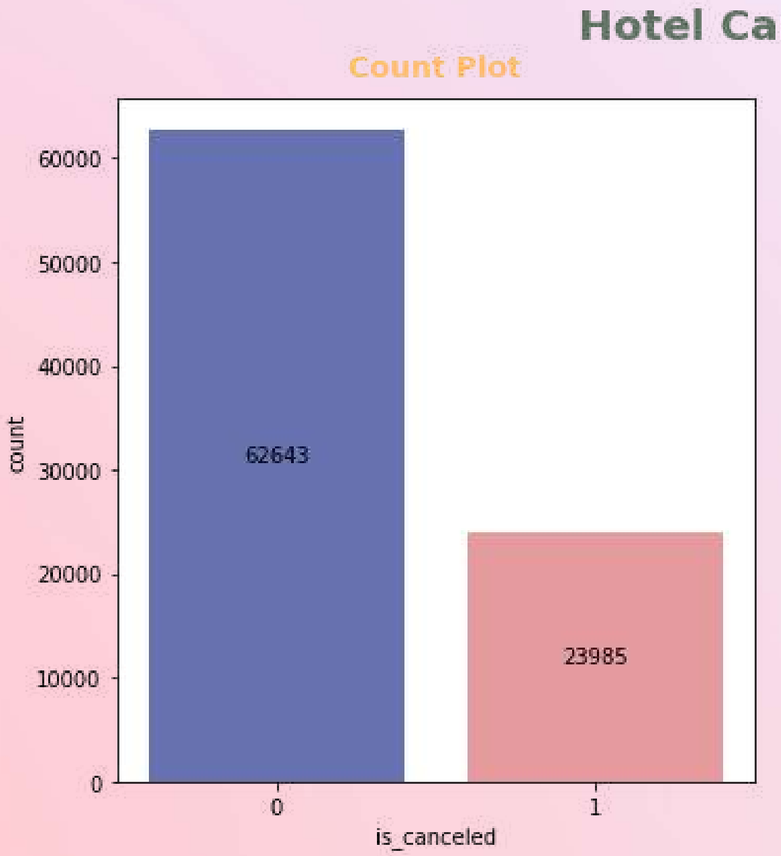
Dataset :

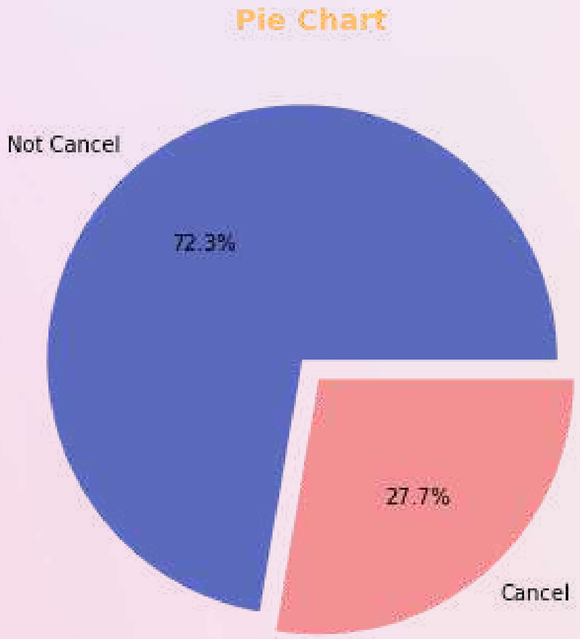
Proyek ini menggunakan dua kumpulan data demand hotel di Portugal. Kedua hotel tersebut adalah **Resort Hotel** dan **City Hotel**. Adapun jumlah data yang akan diolah adalah **32 variabel** dan **119390 pemesanan hotel.** Pemesanan yang terdaftar yaitu dari **1 Juli 2015 s/d 31 Agustus 2017** termasuk pelanggan yang membatalkan pemesanan.

Goals :

- Memprediksi model terbaik untuk pelanggan yang akan membatalkan pemesanan.
- Memberikan rekomendasi bisnis terkait hal tersebut.
- Memberikan bussiness insight terkait data tersebut.

Mengapa mengetahui model cancellation itu penting?





27,7 %

Canceled Booking

- Memaksimalkan penjualan
- Apa yang tidak terjual di hari ini tidak dapat digantikan di hari berikutnya
- Antisipasi
- Memaksimalkan promo kepada target yang tepat

DATA UNDERSTANDING

Feature

Cancellation
is_cancelled
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
booking_changes

booked
hotel
meal
reserved_room_type
required_car_parking_spaces
total_of_special_requests

Arrival Date
lead_time
arrival_date_year
arrival_date_month
arrival_date_week_number
arrival_date_day_of_month

customer
country
market_segment
distribution_channel
deposit_type
agent
company
customer_type

Guest
adults
children
babies

Check-In
assigned_room_type
adr

Long Stays
stays_in_weekend_nights
stay_in_week_nights

Status
days_in_waiting_list
reservation_status
reservation_status_date

Data Cleaning

MISSING VALUES

- Children : 4 rows
- Country : 488 rows
- Agent : 16340 rows
- Company : 112593 rows

- Diasumsikan tidak punya anak
- Dilabeli "Undefined"
- Diisi dengan mean
- Drop column

OTHER TREATMENT

- ADR < 0 : 1 rows
- Total Pengunjung < 0 : 180 rows
- Total Stays < 0 :
- 645 rows
- 29 February : 0 row

- Drop all rows

DUPLICATED

26,9% are duplicated

- Drop all rows

OUTLIERS

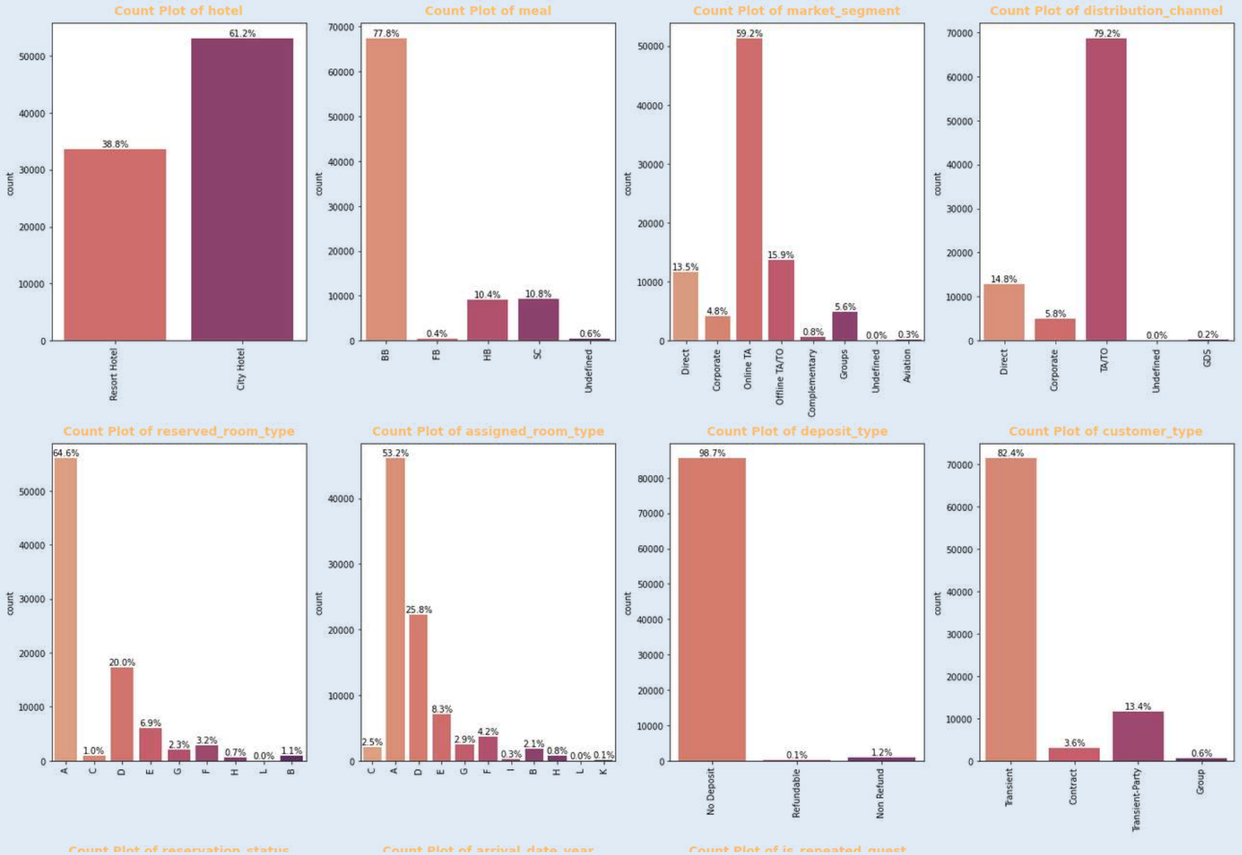
Too many outliers

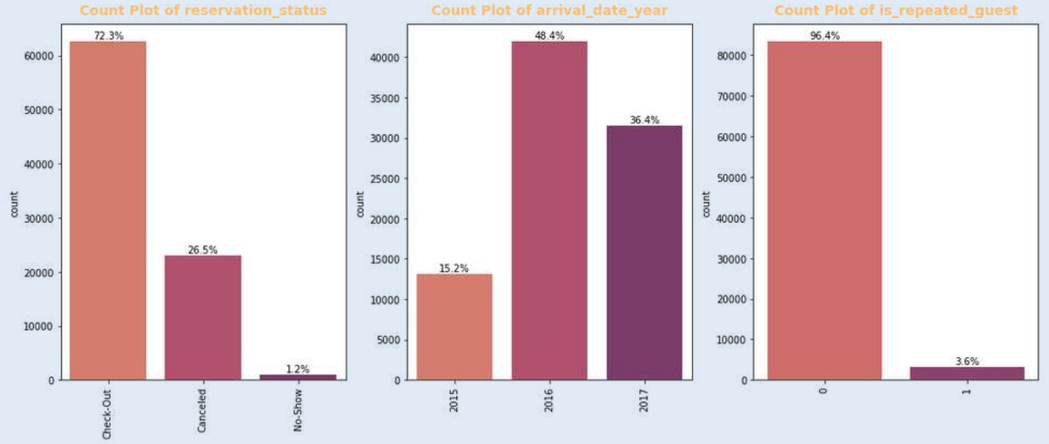
- Drop babies > 4
- Drop adr > 1000

Information

- Rata-rata pengunjung melakukan pemesanan hotel jauh-jauh hari yaitu sekitar 50-80 hari. Tetapi ada juga yang memesan dalam waktu dekat dan ada yang lebih dari 1 tahun. Karena banyak yang tidak melakukan booking, perlu adanya antisipasi terkait fasilitas yang ada.
- Rata-rata pengunjung menginap di weekday 2-3 hari.
- Rata-rata jumlah orang dewasa yang menginap adalah 1-2 orang dan kebanyakan mereka tidak membawa anak dan bayi.
- Ada pengunjung yang seringkali melakukan pembatalan pemesanan sampai 26 kali.
- Ada pengunjung yang loyal sampai melakukan pemesanan sampai 72 kali tanpa cancel.
- Rata-rata pengunjung tidak masuk ke dalam waiting list. It's good.
- Rata-rata ADR 107 dan cukup bervariatif.
- Rata-rata pengunjung tidak memerlukan parkir mobil. Tetapi tetap ada yang memerlukan. Jadi perlu adanya perhitungan yang efisien untuk menentukan jumlah ruang parkir.
- Pemasaran cukup baik karena dapat menjangkau 177 negara lainnya

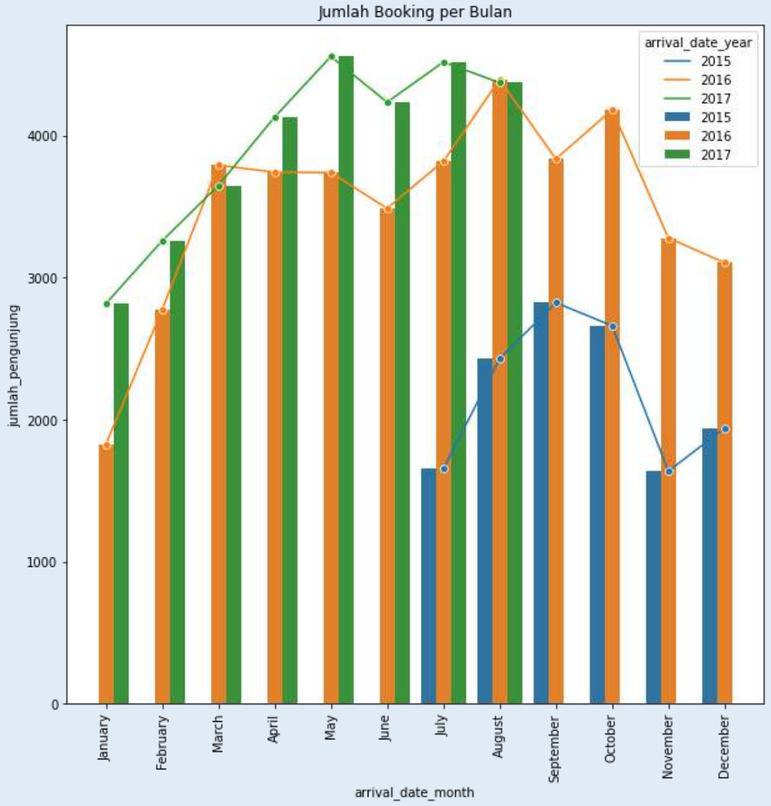
Most Popular

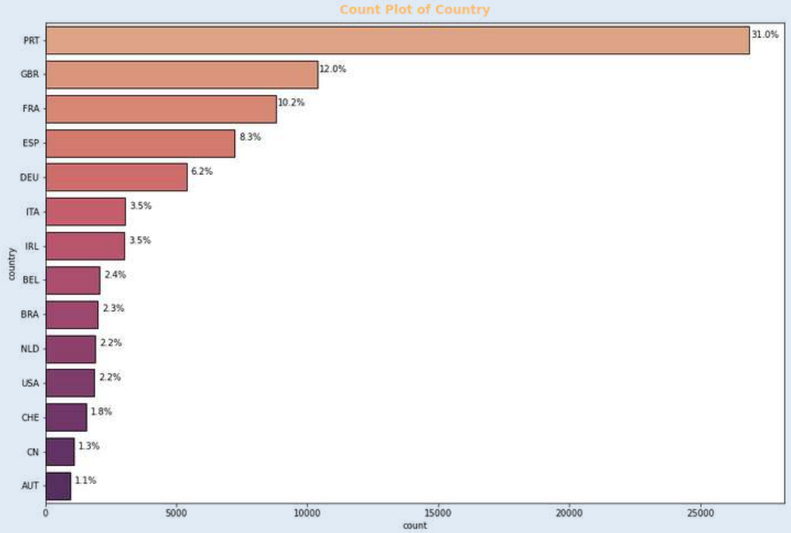


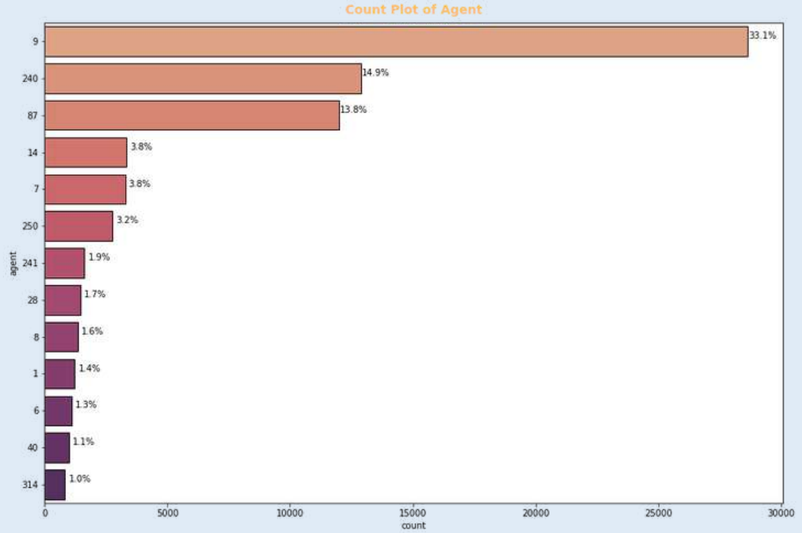


- City Hotel lebih populer ketimbang Resort Hotel.
- Mayoritas pengunjung hanya breakfast saja.
- Market segment dan distributin channel didominasi oleh Travel Agent.
- Tipe kamar paling populer adalah tipe A.
- Kebanyakan pelanggan tidak melakukan deposit.
- Tipe pelanggan transient sangat mendominasi. Perlu adanya promo untuk pelanggan lainnya.
- Jumlah repeated guest sangat minim.

Most Popular







- Warga lokal mendominasi sebagai pelanggan.
- Pemasaran cukup baik karena dapat menjangkau 177 negara lainnya.
- Agent 9 merupakan agen yang paling sering membawa pelanggan.
- Summer Holiday menjadi bulan-bulan tertinggi dalam reservasi pada dua tahun terakhir.

MODELLING & RECCOMENDATION

Filtering Features

DROP COLUMN

- Tanggal kedatangan
- Meal
- Reservation Status
- Tipe Kamar

22

Columns

ENGINEERING

- Country di kelompokkan menjadi lokal dan tidak lokal

22

Columns

ENCODING

- Label Encoding
- One Hot Encoding

38

Columns

Baseline Model

Model Performance Baseline

Model	Recall	AUC	F1 Score	precision	accuracy
RandomForestClassifier	62.83%	76.60%	66.79%	71.27%	82.79%
DecisionTreeClassifier	61.57%	73.34%	61.33%	61.09%	78.62%
LogisticRegression	40.73%	66.05%	49.85%	64.21%	77.43%
XGBClassifier	63.82%	77.65%	68.54%	74.03%	83.87%
GradientBoostingClassifier	56.67%	74.51%	64.11%	73.81%	82.53%
LGBMClassifier	62.33%	76.93%	67.52%	73.66%	83.49%
ExtraTreesClassifier	61.74%	75.55%	65.08%	68.79%	81.76%
HistGradientBoostingClassifier	61.89%	76.83%	67.43%	74.06%	83.54%

Akurasi pada baseline model terbilang cukup besar. Tentu karena data imbalance. Jawaban akan lebih mengarah kepada modeling yang tepat untuk nilai negatif. Sedangkan model yang kita perlukan adalah model untuk nilai positif.

Undersampling Model

Model Performance Undersampling

Model	Recall	AUC	F1 Score	precision	accuracy
RandomForestClassifier	83.90%	79.76%	67.64%	56.66%	77.90%
DecisionTreeClassifier	76.67%	74.63%	61.63%	51.53%	73.72%
LogisticRegression	76.14%	73.87%	60.69%	50.45%	72.84%
XGBClassifier	86.90%	81.12%	69.01%	57.24%	78.52%
GradientBoostingClassifier	86.60%	79.34%	66.59%	54.08%	76.07%
LGBMClassifier	86.60%	80.91%	68.77%	57.03%	78.34%
ExtraTreesClassifier	84.30%	78.90%	66.36%	54.72%	76.47%
HistGradientBoostingClassifier	86.16%	81.26%	69.38%	58.07%	79.06%

Oversampling Model

Model Performance Oversampling

Model	Recall	AUC	F1 Score	precision	accuracy
RandomForestClassifier	69.27%	78.09%	68.01%	66.80%	82.06%
DecisionTreeClassifier	61.49%	73.04%	60.86%	60.25%	78.23%
LogisticRegression	76.73%	73.61%	60.32%	49.69%	72.21%
XGBClassifier	85.79%	81.43%	69.71%	58.71%	79.48%
GradientBoostingClassifier	86.04%	79.44%	66.82%	54.62%	76.47%
LGBMClassifier	86.12%	81.24%	69.36%	58.06%	79.05%
ExtraTreesClassifier	62.73%	75.92%	65.55%	68.64%	81.85%
HistGradientBoostingClassifier	86.39%	81.39%	69.51%	58.15%	79.14%

Best Model

Dalam hal ini, kita perlu memerhatikan beberapa metric evaluation :

- Precision : Untuk mengetahui keakuratan model dalam menentukan booking yang akan cancel.
- Recall : Untuk mengetahui banyaknya kesalahan dalam menentukan cancel (dilihat recall kecil)

Selain itu model yang akan dipilih adalah undersampling model karena merupakan model dengan balanced data yang di treatment tanpa adanya data sintesis.

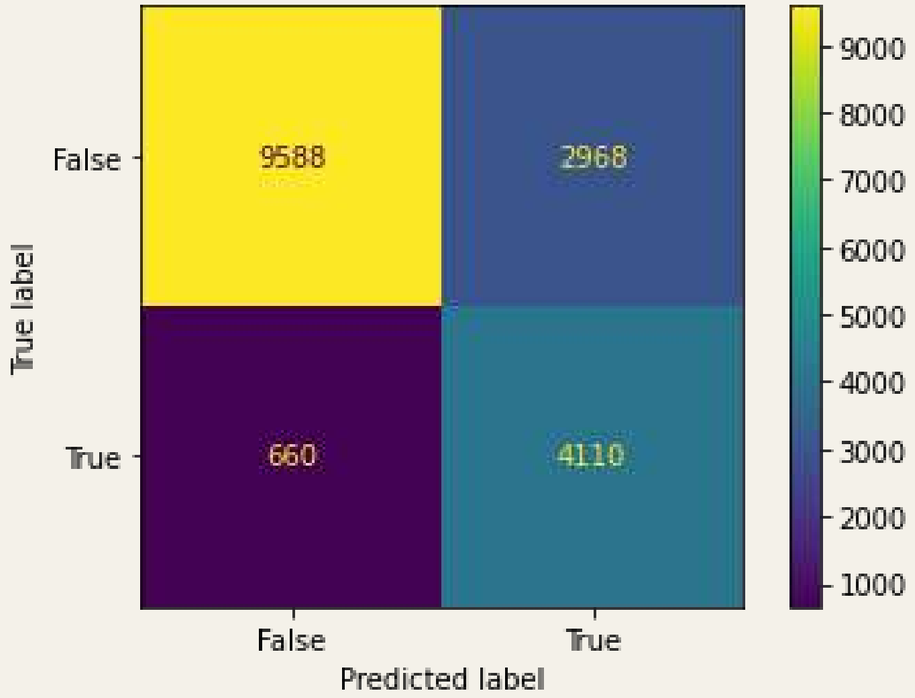
Model Performance Undersampling

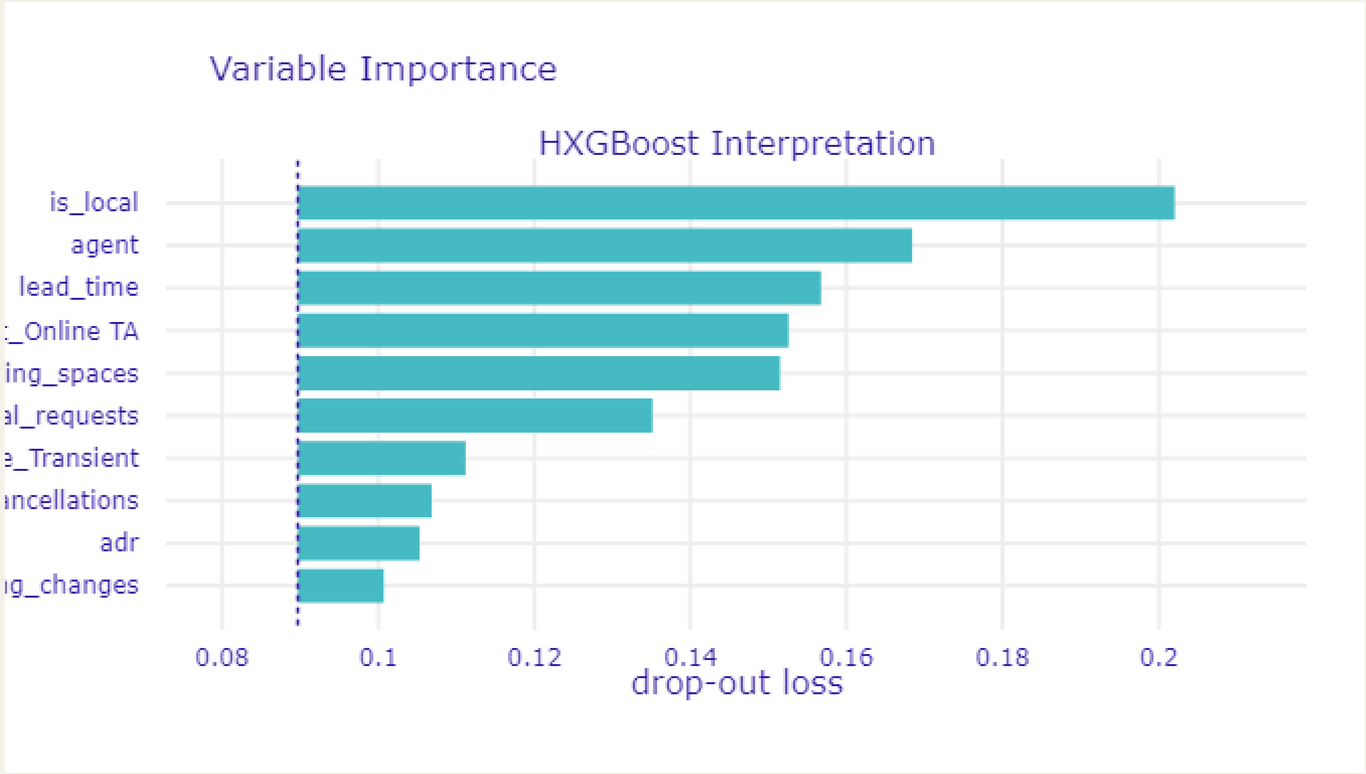
Model	Recall	AUC	F1 Score	precision	accuracy
RandomForestClassifier	83.90%	79.76%	67.64%	56.66%	77.90%
DecisionTreeClassifier	76.67%	74.63%	61.63%	51.53%	73.72%
LogisticRegression	76.14%	73.87%	60.69%	50.45%	72.84%
XGBClassifier	86.90%	81.12%	69.01%	57.24%	78.52%
GradientBoostingClassifier	86.60%	79.34%	66.59%	54.08%	76.07%
LGBMClassifier	86.60%	80.91%	68.77%	57.03%	78.34%
ExtraTreesClassifier	84.30%	78.90%	66.36%	54.72%	76.47%
HistGradientBoostingClassifier	86.16%	81.26%	69.38%	58.07%	79.06%

Maka dari itu, terpilih lah Hist Gradient Boosting Clasifier.

- Memiliki nilai akurasi paling tinggi dan cukup baik.
- Memiliki nilai precision paling tinggi. Artinya dari 100 yang di prediksi akan cancel, 58 benar cancel.
- Memiliki nilai recall 86%. Artinya, dari 100 orang yang cancel, yang tidak terprediksi cancel hanya 14 orang.

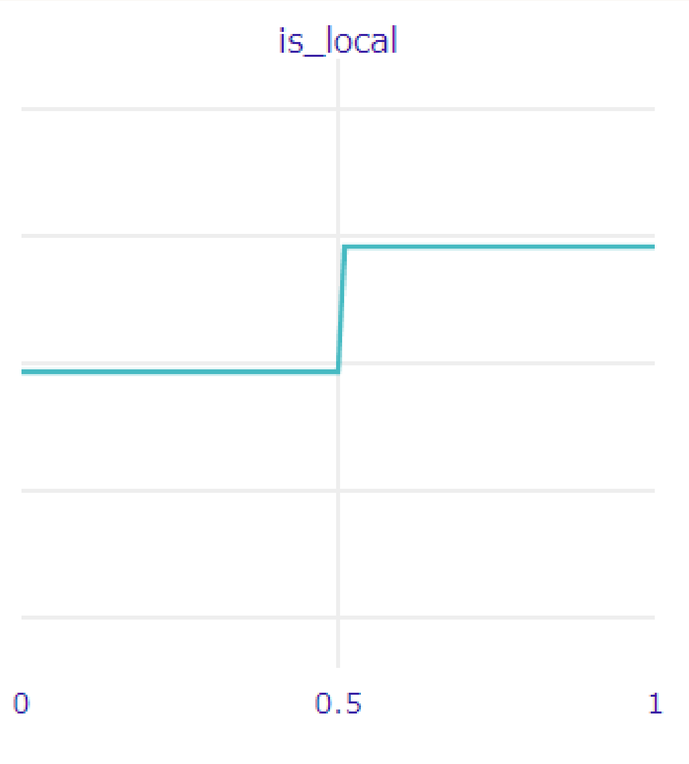
Hist Gradient Boosting Clasifier

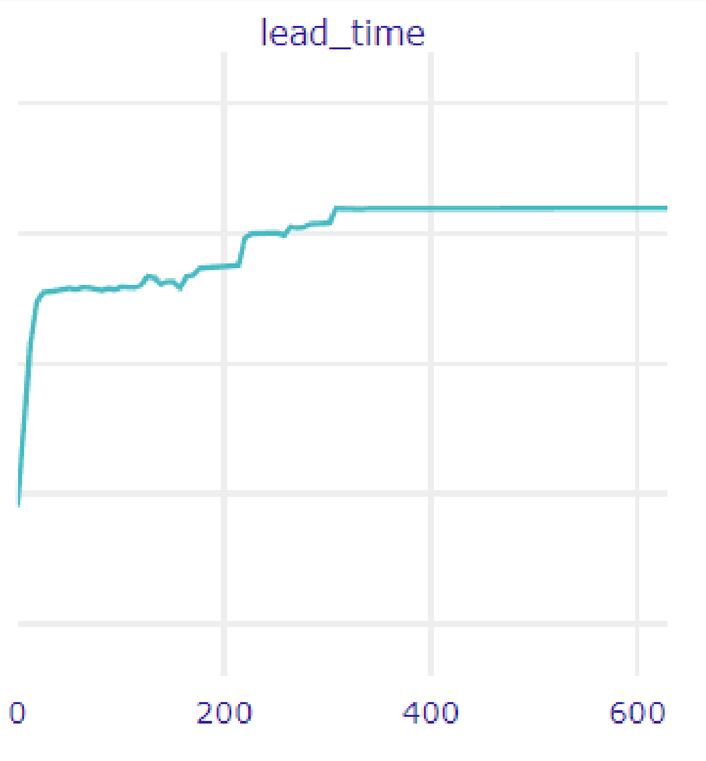




Dapat dilihat bahwa domisili pengunjung dan rentang pemesanan memberikan pengaruh yang cukup besar. Pengunjung lokal lebih dominan untuk cancel. Selain itu jika waktu tunggu lebih lama juga lebih dominan cancel. Sesuai fakta bahwa orang yang memesan di hari H lebih jelas untuk menginap. Maka dari itu saya merekomendasikan untuk memberikan email notification terkait confirmation booking ketika waktu sudah dekat. Sehingga jika cancel, dapat segera dibuka untuk penawar lain.

Hist Gradient Boosting Clasifier





Best Model

Model Performance Undersampling

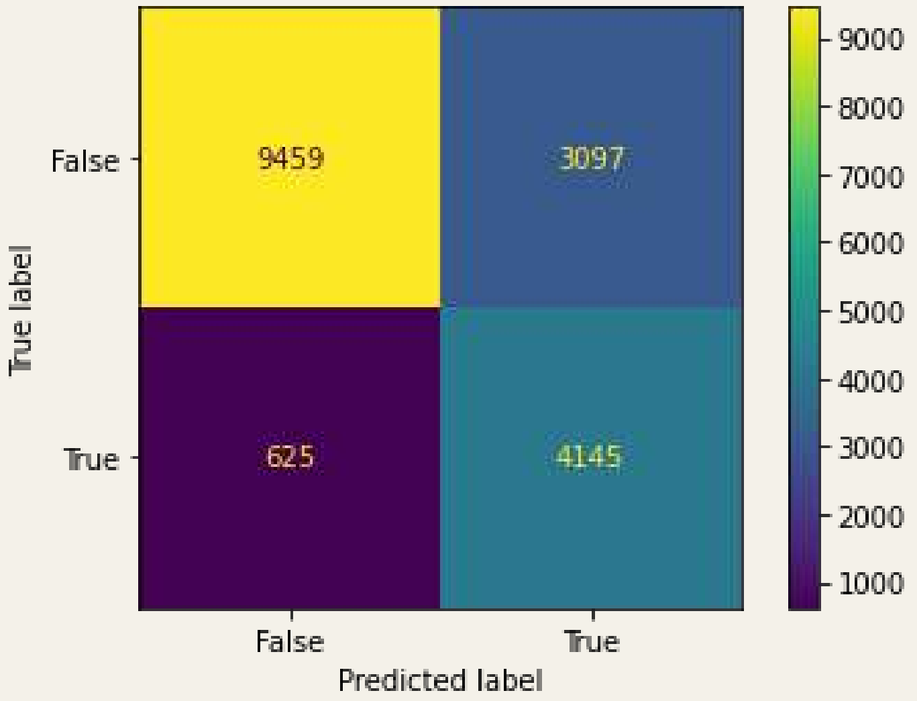
Model	Recall	AUC	F1 Score	precision	accuracy
RandomForestClassifier	83.90%	79.76%	67.64%	56.66%	77.90%
DecisionTreeClassifier	76.67%	74.63%	61.63%	51.53%	73.72%
LogisticRegression	76.14%	73.87%	60.69%	50.45%	72.84%
XGBClassifier	86.90%	81.12%	69.01%	57.24%	78.52%
GradientBoostingClassifier	86.60%	79.34%	66.59%	54.08%	76.07%
LGBMClassifier	86.60%	80.91%	68.77%	57.03%	78.34%
ExtraTreesClassifier	84.30%	78.90%	66.36%	54.72%	76.47%
HistGradientBoostingClassifier	86.16%	81.26%	69.38%	58.07%	79.06%

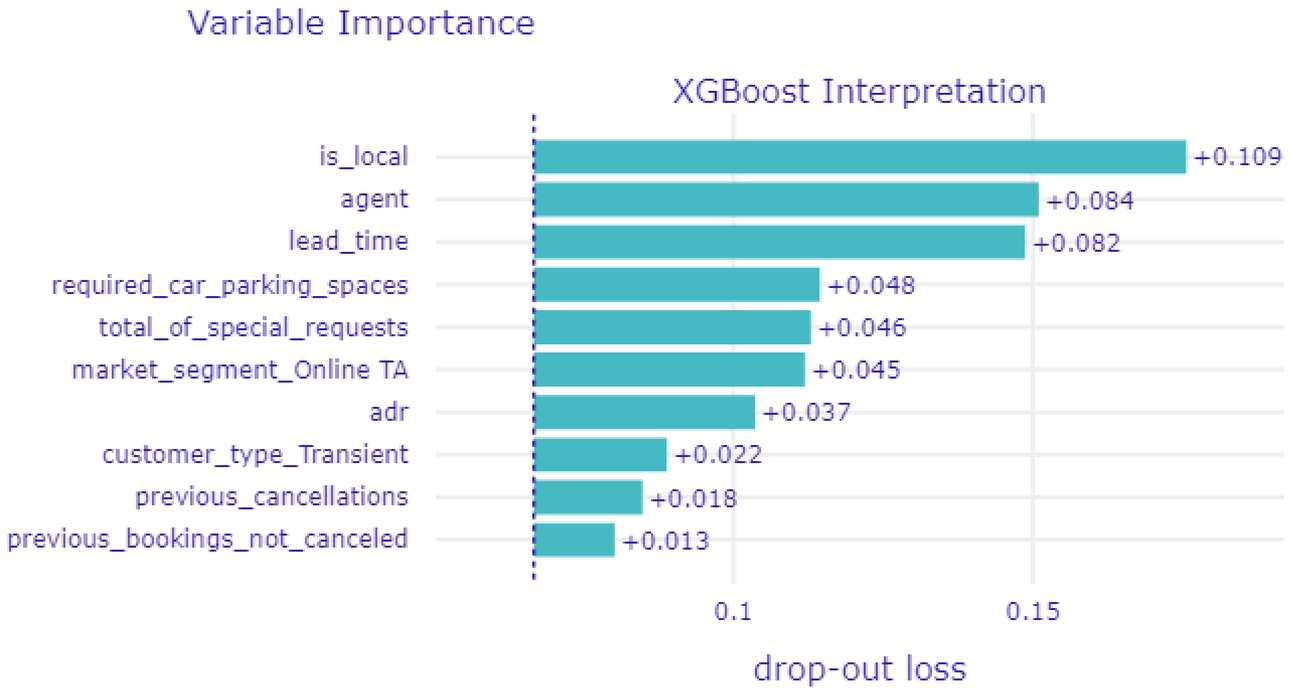
Selain itu, jika kita memerhatikan nilai recallnya, XGB Classifier adalah yang terbaik. Lalu jika kita perhatikan, nilai kedua model ini tidak jauh berbeda.

- Dari 100 yang di prediksi akan cancel, 57 benar cancel.
- Dari 100 orang yang cancel, yang tidak terprediksi cancel hanya 13 orang.

Kedua model dapat digunakan dengan baik.

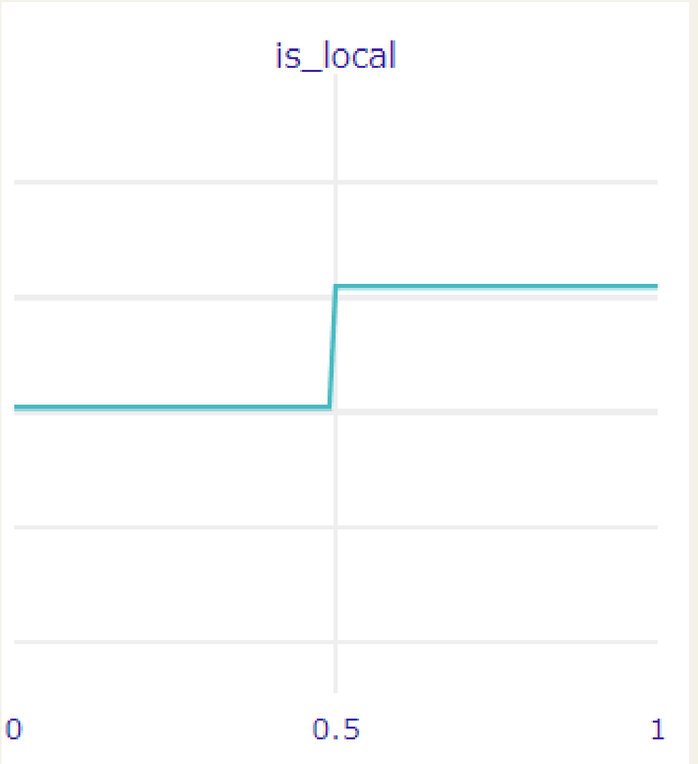
XGB Classifier





Dapat dilihat bahwa domisili pengunjung dan rentang pemesanan memberikan pengaruh yang cukup besar. Pengunjung lokal lebih dominan untuk cancel. Selain itu jika waktu tunggu lebih lama juga lebih dominan cancel. Sesuai fakta bahwa orang yang memesan di hari H lebih jelas untuk menginap. Maka dari itu saya merekomendasikan untuk memberikan email notification terkait confirmation booking ketika waktu sudah dekat. Sehingga jika cancel, dapat segera dibuka untuk penawar lain.

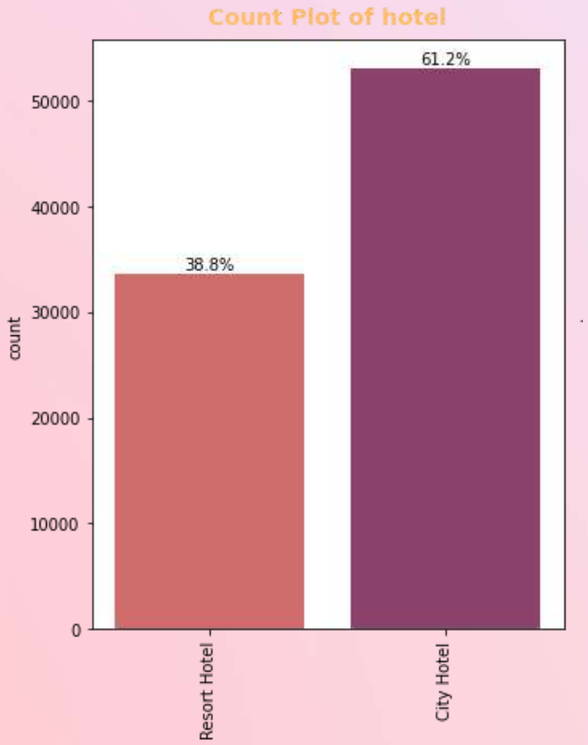
XGB Classifier

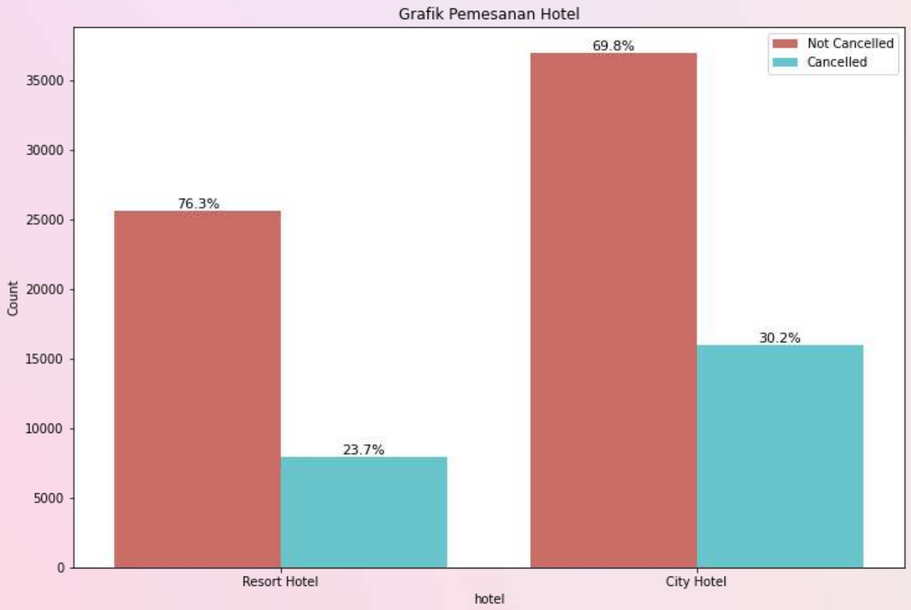


DEEP DIVE QUESTION & BUSSINESS INSIGHT

Deep Dive Question

Hotel mana yang persentase cancelnya lebih besar?

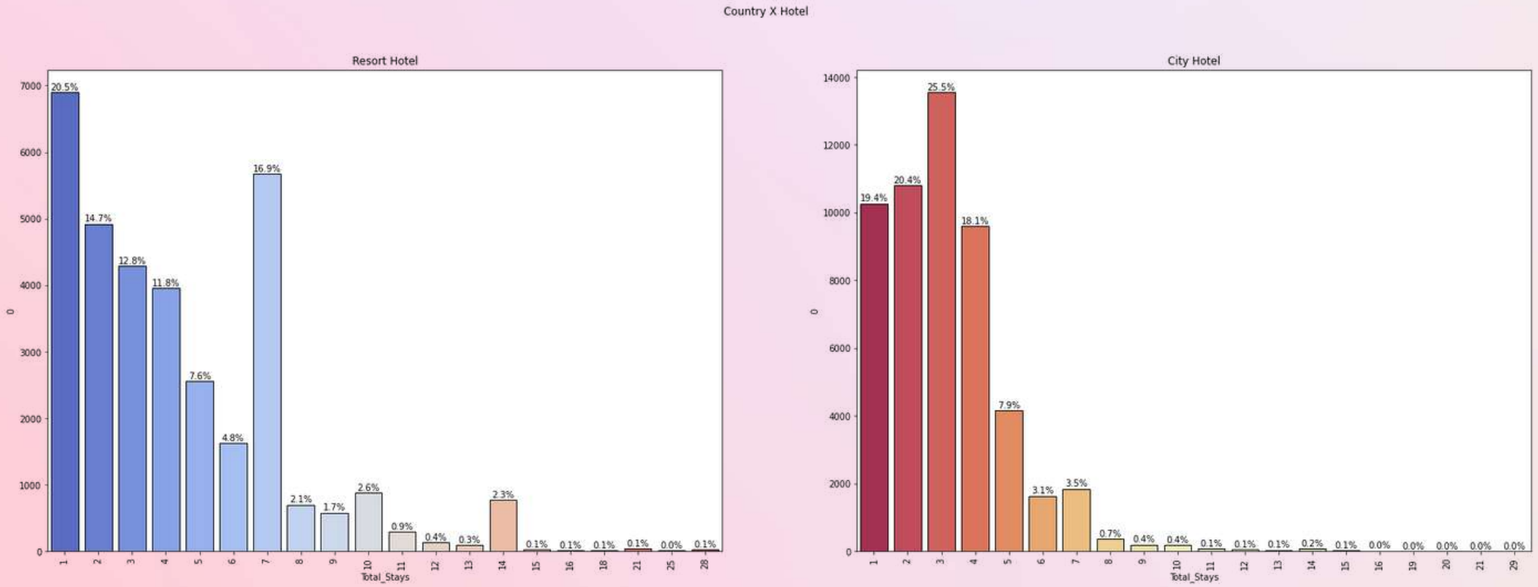




Walaupun City Hotel lebih banyak melakukan pemesanan dan checkout, tetapi cancelnya juga lebih besar daripada Resort Hotel. Jadi City Hotel belum tentu lebih stabil ketimbang Resort Hotel.

Deep Dive Question

Berapa lama pengunjung biasanya menginap?



	hotel	Total_Stays	0
0	City Hotel	3	13541
1	City Hotel	2	10812
2	City Hotel	1	10267
3	City Hotel	4	9610
4	Resort Hotel	1	6894
5	Resort Hotel	7	5675
6	Resort Hotel	2	4921
7	Resort Hotel	3	4285
8	City Hotel	5	4168
9	Resort Hotel	4	3955

City Hotel kebanyakan diinapi 3 hari atau rentang 1-4 hari . Sedangkan Resort Hotel kebanyakan diinapi 1 atau 7 hari, selebihnya 2-4 hari.

Deep Dive Question

Berapa jumlah paling banyak dan rata-rata pengunjung tiap booking?

	Total_Guests	
	mean	max
hotel		
City Hotel	2.022267	5
Resort Hotel	2.043555	55

Jumlah pengunjung di kedua hotel rata-rata adalah 2 orang. Adapun untuk maksimum pengunjung per pemesanan adalah 5 untuk City Hotel dan 55 untuk Resort Hotel. Artinya City Hotel tidak pernah menerima pelanggan dalam jumlah besar, kemungkinan hanya family dan bussiness trip saja. Sedangkan Resort Hotel kemungkinan lebih sering menerima group.

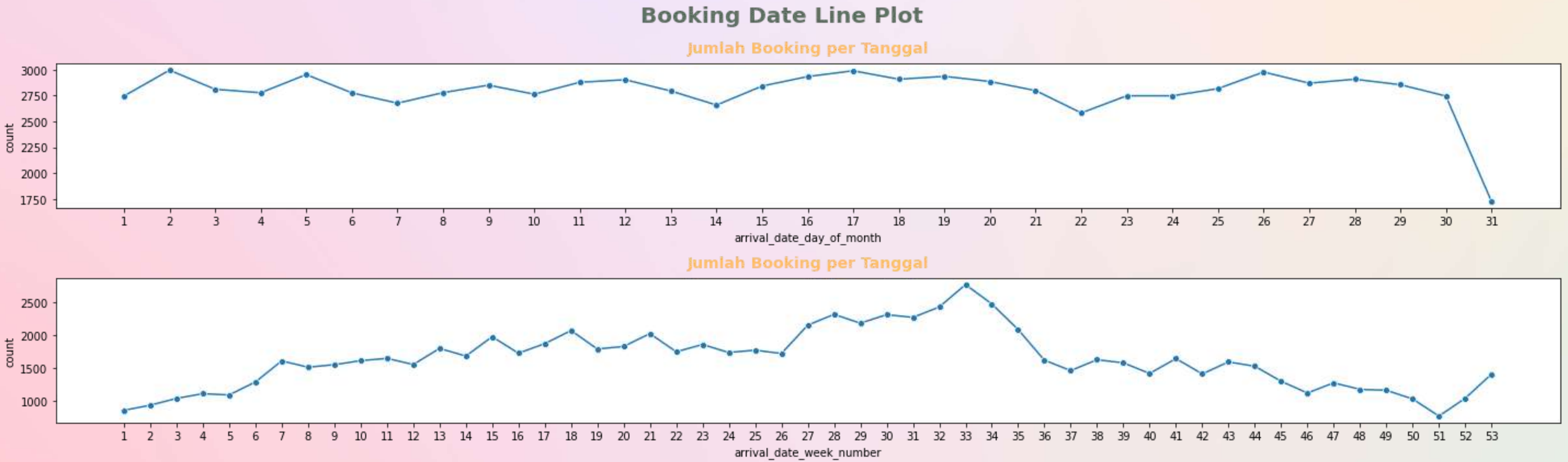
Deep Dive Question

Berapa jumlah parkiran yang diperlukan agar pihak hotel dapat selalu memenuhi keinginan customernya?

hotel	arrival_date	jumlah_mobil
Resort Hotel	2016-06-25	22
City Hotel	2016-03-24	20
Resort Hotel	2017-04-29	20

Untuk Resort Hotel yaitu sebanyak 22 mobil. Adapun City Hotel 20 mobil.

Time



Jumlah booking tiap tanggal hampir rata berkisar 2600 sampai 3000 kecuali tanggal 31 kurang dari 1750. Hal ini wajar karena kemunculan tanggal 31 dalam satu tahun setengahnya dari kemunculan tanggal lainnya.

Output Modelling

CASE STUDY

- ADR = 107 Euro (diambil dari rata-rata)
- Total stays per book : 2-3 hari (diambil dari data sebelumnya)
- Jumlah pengunjung / hari : 96 - 111
- Cancel : 27,7%

Cancel booking

15 - 18

/ hari

Income

3210 - 5778

/ hari

2 sampai 3 orang belum terprediksi cancel

428 - 963

Income Increase

17% - 18 %

Recommendation

- Banyak guest yang tidak melakukan book, jadi selalu perlu untuk antisipasi terkait fasilitas.
- Persebaran pemasaran sudah cukup baik.
- Tipe Kamar A sering kali overbook, sehingga perlu ditambah.
- Karena summer holiday seringkali ramai, baik jika mengadakan langkah preventif untuk menghindari cancel, seperti menawarkan booking dari jauh hari dengan deposit. Diluar daripada itu tidak diperkenankan memesan kecuali walk-in dan masih kosong.
- Memberikan email notification terkait confirmation booking ketika waktu sudah dekat.
- Menawarkan pengisian feedback untuk mengetahui pelanggan mana yang puas dengan pelayanan kita. Untuk yang memberi rating baik, bisa diberikan promo sehingga ia akan menjadi repeated guest melihat data repeated guest masih sedikit.
- Meningkatkan pelayanan supaya rating baik.
- City Hotel bisa meningkatkan pemasarannya keluar dari transient. Karena dominan guest adalah family.
- Maximizing parking.

THANK YOU!

Reach me out!

www.linkedin.com/in/zalfamaitsa

Link project : here