**PREDIKSI PEMBATALAN PEMESANAN KAMAR HOTEL**

Oleh :

\*   Irfan Maulana Yusuf

\*   M. Amin Zakaria

\*   Siti Andriana Rahmayanti

**### \*\*A. Stakeholder\*\***

\* Pihak manajemen Hotel

**### \*\*B. Context\*\***

Industri pelayanan dan pariwisata merupakan sektor penting yang menunjang ekonomi global. Industri ini memberikan pendapatan, membuka lapangan kerja dan menstimulus perkembangan ekonomi di berbagai wilayah. Berbagai perusahaan pada industri ini memiliki karakteristik tertentu seperti pada makanan yang disajikan, akomodasi yang ada, hiburan sampai rekreasi yang diberikan. Karakteristik tersebut menjadikan tiap wilayah atau tempat memiliki daya tarik tersendiri bagi pada turis (Sampaio dkk, 2024. https://doi.org/10.3390/).

Industri pariwisata sangat sensitif terhadap perubahan sedangkan turis cenderung untuk menghindari risiko. Turis akan lebih memilih datang ke wilayah yang aman, bersih, tidak ada wabah tertentu dan memiliki iklim cuaca yang mudah diterima para turis. Di sisi lain, turis juga mempertimbangkan biaya yang diperlukan dari segi akomodasi, seperti harga tiket pesawat, hotel hingga perbandingan kurs mata uang dan juga fasilitas di tiap wilayah yang akan mereka kunjungi. Hal ini membuat para pelaku bisnis pariwisata melakukan berbagai strategi untuk memenuhi ekspektasi para turis untuk datang ke tempat mereka.

Negara-negara Eropa masih menjadi tempat destinasi yang banyak dikunjungi para turis, salah satunya Portugal. Berdasarkan 2018 Annual Research yang dilakukan oleh World Travel & Tourism Council, tercatat bahwa sektor pariwisata memberikan kontribusi GDP hingga 17% pada 2017 (https://wttc.org/research/economic-impact) yang menandakan bahwa Portugal menjadi salah satu negara favorit yang dikunjungi para turis. Menariknya, Portugal banyak melakukan pengembangan infrastruktur untuk pariwisata mulai dari transportasi, kebijakan politik yang mempengaruhi pariwisata negara, hingga pengembangan akomodasi seperti villa dan hotel.

Hotel merupakan salah satu daftar penting dari para turis sebagai syarat untuk melakukan wisata. Pilihan hotel di suatu negara akan mempengaruhi minat para turis untuk berkunjung, seperti menentukan hotel di tengah kota atau di pinggir laut. Pelaku bisnis perhotelan tentu akan memberikan banyak opsi menarik kepada para turis. Tantangannya adalah bagaimana pendekatan yang dilakukan pelaku bisnis ini agar sesuai dengan minat para turis. Pasalnya, tidak sedikit hotel di wilayah yang banyak disinggahi turis mengalami pembatalan yang membuat pendapatan turun hingga 20% dari forecast. (https://wttc.org/research/economic-impact)

Masalah pembatalan yang dihadapi oleh pelaku bisnis perhotelan tentu perlu dilakukan pendekatan seperti survey, kuisioner, observasi, atau melakukan analisa dengan membuat pemodelan. Pemodelan yang dimaksud adalah membangun model *\_Machine Learning\_* menggunakan data yang ada, yang nanti nya mampu memprediksi kemungkinan pembatalan atas setiap transaksi yang dilakukan oleh para pelanggan. Pendekatan ini sudah banyak dilakukan di berbagai sektor bisnis dan menunjukkan hasil yang signifikan dapat membantu sesuai dengan masalah yang dihadapi. Maka dari itu, pendekatan dengan Machine Learning diharapkan mampu menyelesaikan masalah  terkait pembatalan transaksi yang dihadapi pelaku bisnis perhotelan.

**### \*\*C. Problem Statement\*\***

Potensi terjadinya pembatalan pemesanan kamar menimbulkan dampak negatif bagi industri perhotelan, terutama dari segi pendapatan dan biaya. Hal ini dikarenakan pihak hotel harus menjamin ketersediaan kamar yang sesuai dengan ekspektasi calon penghuni kamar dan harus menanggung biaya atas kekosongan kamar tersebut. Pihak manajemen hotel berupaya untuk memperkecil peluang adanya pembatalan pemesanan sehingga dapat memaksimalkan resources yang ada.

Dalam penyelesaiannya, penggunakan teknologi Machine Learning menjadi solusi yang potensial untuk memprediksi pembatalan pemesanan para pelanggan. Teknologi ini mmampu menganalisa pola data yang kompleks hingga dapat membantu pihak manajemen membangun model yang secara prediksi lebih akurat dibanding pendekatan konvesional (survey, kuisioner, observasi lapangan, etc.). Apalagi, model ini mampu mengolah data secara besar sehingga dapat efisien dalam mengidentifikasi pola serta faktor utama yang mempengaruhi pembatalan.

Maka dari itu, proyek ini harus mampu menjawab pertanyaan berikut.

1. Apa saja faktor utama yang mempengarhui kemungkinan pemesanan dibatalkan?

2. Bagaimana customer behaviour para pelanggan dalam menentukan melakukan pembatalan atau tidaknya pemesanan hotel?

3. Seberapa jauh model yang dibangun dapat memprediksi pembatalan pemesanan secara akurasi?

4. Apakah pendekatan dengan Machine Learning mampu meningkatkan akurasi dalam memprediksi pembatalan pesanan pada industri perhotelan dibanding pendekatan konvensional?

**### \*\*C. Goals\*\***

1. Mengetahui dan memahami faktor utama yang mempengaruhi pembatalan pesanan sehingga mampu menurunkan potensi pembatalan dari pemesanan hotel

2. Memahami customer behaviour para pelanggan sehingga mampu membuat strategi yang sesuai dengan minat para pelanggan

3. Model ini mampu memiliki akurasi prediksi di atas 50% dari pembangunan model dengan data yang tersedia. Faktor-faktor yang memicu adanya pembatalan nantinya akan dapat dihindari atau diminimalisir sehingga dapat mengoptimalkan resources dan revenue yang didapatkan.

4. Model ini diharapakan mampu melakukan prediksi lebih baik dibanding pendekatan secara konvensional. Harapannya model ini juga dapat digunakan dengan data-data ditahun yang akan datang sehingga mampu memprediksi lebih akurat.

**### \*\*D. Analytical Approach\*\***

Analisa minat pelanggan terhadap perminataan hotel membutuhkan pendekatan analitis untuk mengetahui tren, forecast, dan memahami faktor apa saja yuang mempengaruhi customer behavior para pelanggan. Berikut pendekatan yang akan dilakukan pada proyek ini.

1. Descriptive Analysis: Merangkum data historik guna mengidentifikasi pola minat yang ada.

2. Diagnostic Analysis: Melakukan investigasi terhadap fluktuasi permintaan hotel dengan melakukan visualisasi dan korelasi antar variabel.

3. Predictive Analytics: Melakukan pemodelan untuk memberikan forecast permintaan hotel kedepannya dengan data yang ada menggunakan model klasifikasi. Penggunaan model klasifikasi karena kita akan memprediksi output berupa data kategorikal secara kualitatif (Batal atau Tidak)

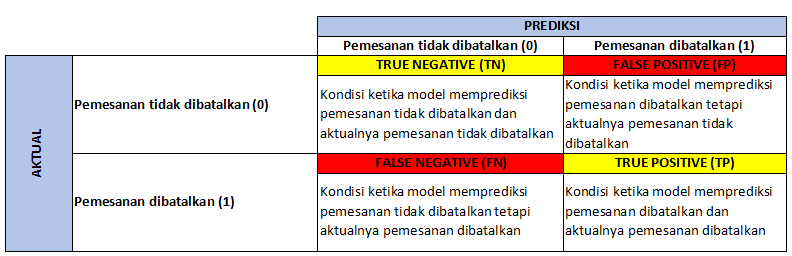
**### \*\*E. Target\*\***

Kolom yang menjadi target adalah kolom 'is\_canceled':

\* 0 : pemesanan tidak dibatalkan

\* 1 : pemesanan dibatalkan

**\*\*Metric Evaluation\*\***



Analisis Kesalahan Prediksi:

Tipe 1 : *\*False Positive\** (FP)

Jika yang terjadi adalah FP, dimana model memprediksi pemesanan dibatalkan padahal aktualnya pemesanan tidak dibatalkan, maka risiko yang diterima oleh pihak hotel adalah:

\* marketing yang tidak tepat sasaran, sehingga mengeluarkan biaya marketing yang sia-sia

Tipe 2 : *\*False Negative\** (FN)

Jika yang terjadi adalah FN, dimana model memprediksi pemesanan tidak dibatalkan padahal aktualnya pemesanan dibatalkan, maka risiko yang diterima oleh pihak hotel adalah:

\* kehilangan pendapatan

\* sia-sianya biaya, waktu, dan sumber daya untuk merealisasikan ekspektasi calon penghuni

Analisis *\*Metric\** yang Akan Digunakan:

Dari analisis risiko atas terjadinya kesalahan prediksi, diperoleh kerugian yang lebih besar akan diterima oleh pihak hotel jika terjadi kesalahan prediksi tipe 2, yaitu *\*False Negative\** (FN), dimana model memprediksi pemesanan tidak dibatalkan padahal aktualnya pemesanan dibatalkan. Oleh karena itu, untuk mereduksi terjadinya kesalahan prediksi tersebut, digunakan *\*metric F2 Score\**.

*\*F2 Score\** adalah salah satu metric yang menitikberatkan pada nilai recall daripada precision. *\*Metric\** ini digunakan jika mereduksi FN menjadi prioritas tanpa meninggalkan FP. Rumus *\*F2 Score\** :

![image.png](attachment:image.png)

**2. Data Understanding**

Pada bagian ini., kita melakukan import libraries yang diperlukan untuk melakukan analisis dan membangun model.

A. Data Information

Selanjutnya, kita melakukan eksplor informasi umum mengenai dataset yang kita punya, yaitu sebagai berikut.

**\*\*Dataset source:\*\***

[https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data](https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data)

**\*\*Notes:\*\***

Dari sumber : [https://www.sciencedirect.com/science/article/pii/S2352340918315191](https://www.sciencedirect.com/science/article/pii/S2352340918315191), diperoleh:

\* Dataset ini diambil dari data asli, sehingga identitas pribadi dihapus dari dataset untuk menjaga privasi

\* Dataset ini diambil dari data pemesanan hotel yang diterima antara tanggal 1 Juli 2015 sampai dengan 31 Agustus 2017, termasuk pemesanan yang berhasil dan dibatalkan

\* Dataset ini memiliki 32 kolom dan 119.390 baris dengan sebanyak 79.330 data dari pemesanan di City Hotel dan 40.060 data dari pemesanan di Resort Hotel

\* Dataset ini diambil dari hotel yang berada di Portugal dengan Resort Hotel yang berada di wilayah Algarve dan City Hotel yang berada di Kota Lisbon.

\* Property Management System (PMS) memastikan tidak ada data yang hilang (missing) pada dataset ini.

|Kolom | Deskripsi |

|---- | ---- |

| Hotel  | Tipe hotel ( "City Hotel" atau "Resort Hotel")|

| is\_canceled| Status pembatalan pemesanan (0 = tidak dibatalkan, 1 = dibatalkan)|

| lead\_time|Jumlah hari antara waktu pemesanan dan tanggal check-in|

| arrival\_date\_year  | Tahun dari tanggal kedatangan |

| arrival\_date\_month| Bulan dari tanggal kedatangan|

| arrival\_date\_week\_number | pada minggu ke berapa tamu akan tiba di hotel|

| arrival\_date\_day\_of\_month| Hari dalam bulan untuk tanggal kedatangan|

| stays\_in\_weekend\_nights | Jumlah malam menginap pada akhir pekan (Sabtu-Minggu)|

| stays\_in\_week\_nights | Jumlah malam menginap pada hari kerja (Senin-Jumat)|

| adults | Jumlah orang dewasa yang terdaftar dalam pemesanan|

| children| Jumlah anak-anak yang terdaftar dalam pemesanan|

| babies|Jumlah bayi yang terdaftar dalam pemesanan|

| meal  | Tipe paket makanan yang dipesan (misalnya: "BB" = Bed & Breakfast, "FB" = Full Board)|

| country| Kode negara asal pelanggan berdasarkan standar ISO 3166-1 Alpha-3|

| market\_segment |Segmen pasar pemesanan (misalnya: "Direct", "Corporate", "Online TA")|

| distribution\_channel| Saluran distribusi pemesanan (misalnya: "Direct", "Corporate", "TA/TO")|

| is\_repeated\_guest | Indikator apakah tamu adalah pelanggan yang menginap berulang (0 = baru, 1 = berulang)|

| previous\_cancellations | Jumlah pembatalan sebelumnya yang dilakukan oleh pelanggan|

| previous\_bookings\_not\_canceled | Jumlah pemesanan sebelumnya yang tidak dibatalkan oleh pelanggan|

| reserved\_room\_type| Tipe kamar yang awalnya dipesan oleh pelanggan|

| assigned\_room\_type| Tipe kamar yang sebenarnya diberikan kepada pelanggan|

| booking\_changes  | Jumlah perubahan pemesanan yang dilakukan oleh pelanggan |

| deposit\_type| Jenis deposit yang dilakukan (misalnya: "No Deposit", "Refundable", "Non Refund")|

| agent |ID agen yang mengatur pemesanan (nilai "NULL" jika tidak ada agen)|

| company|ID perusahaan yang mengatur pemesanan (nilai "NULL" jika tidak ada perusahaan)|

| days\_in\_waiting\_list | Jumlah hari pemesanan berada dalam daftar tunggu sebelum dikonfirmasi|

| customer\_type | Tipe pelanggan berdasarkan perilaku pemesanan (misalnya: "Transient", "Contract")|

| adr | *\*Average Daily Rate\** (ADR) atau rata-rata pendapatan per kamar per malam yang dipesan|

| required\_car\_parking\_spaces| Jumlah tempat parkir mobil yang diperlukan oleh pelanggan|

| total\_of\_special\_requests|Jumlah permintaan khusus yang dibuat oleh pelanggan|

| reservation\_status| Status pemesanan terakhir (misalnya: "Check-Out", "Canceled")|

| reservation\_status\_date| Tanggal dari status terakhir pemesanan|

Dari eksplorasi tersebut, kita mendapatkan beberapa keterangan yang menunjukkan anomali data, seperti adanya nilai negatif pada kolom ADR, kemudian adanya missing values, tipe data yang tidak sesuai, yang bisa dihandle pada bagian data cleaning.

1. Data Cleaning
2. Handle Missing Values

Pada bagian data cleaning, kita akan menghandle missing values dengan mengisi kolomnya ataupun dengan menghapus data yang ada berdasarkan informasi yang diterima.

- Missing values pada kolom children akan didrop karena tidak ada keterangan mengenai missing values tersebut dan missing valuesnya hanya 0,003% dari keseluruhan data.

- Missing values pada kolom country juga akan didrop karena tidak ada keterangan mengenai data yang missing tersebut, sehingga apabila diisi dikhawatirkan akan menimbulkan bias pada data.

- Missing values pada kolom agent kita isi dengan 0, yang artinya orang tersebut memesan kamar tidak melalui agen. Hal ini didasari oleh informasi pada sumber [https://www.sciencedirect.com/science/article/pii/S2352340918315191](https://www.sciencedirect.com/science/article/pii/S2352340918315191).

- Berdasarkan sumber yang sama dengan yang di atas, missing values pada kolom company kita isi juga missing valuesnya dengan 0, yang artinya tamu tersebut tidak berkaitan dengan suatu perusahaan tertentu

1. Data Correction

Pada bagian ini, kita memperbaiki tipe data yang tidak sesuai dan mengubah isi kolom yang tidak sesuai berdasarkan informasi atau pengetahuan yang dimiliki.

- Mengubah tipe kolom reservation\_status\_date dari object menjadi datetime

- Mengubah tipe kolom children dari float menjadi integer karena kolom tersebut menunjukkan banyak anak yang akan menginap, sehingga tidak mungkin nilainya berupa desimal.

- Mengubah isi ‘Undefined’ pada kolom meal dengan SC, karena berdasarkan sumber [https://www.sciencedirect.com/science/article/pii/S2352340918315191](https://www.sciencedirect.com/science/article/pii/S2352340918315191), Undefined di sini sama dengan tamu tersebut tidak memesan paket makan (SC)

- Menghapus data dengan isi ‘Undefined’ pada kolom distribution\_channel karena tidak ada sumber yang menerangkan maksud dari Undefined pada kolom ini. Di samping itu, persentase jumlah data yang memiliki isi Undefined juga sangat kecil, sehingga data tersebut kita drop.

1. Duplicate

Pada bagian ini, kita akan mengecek apakah terdapat data duplikat di dataset kita. Ternyata setelah dicek, terdapat data duplikat, sehingga data duplikat tersebut akan kita drop.

1. Anomaly Data

Pada bagian ini, kita akan meng-handle data-data yang tidak logis atau anomali.

- Terdapat nilai negatif pada kolom ADR, dimana ADR adalah rata-rata harga per kamar per hari yang tidak mungkin negatif. Dengan demikian, data dengan nilai negatif pada kolom ADR kita drop.

- Nilai 0 pada kolom adults artinya tidak ada orang dewasa yang ikut menginap, dimana menurut regulasi secara umum, yang boleh reservasi ataupun check-in hanyalah orang dewasa yang berumur 17 atau 18 tahun, sehingga kasus ini tidak mungkin terjadi. Jadi, data yang memiliki nilai 0 pada kolom adults kita drop.

- Nilai yang lebih besar dari 365 hari pada kolom lead\_time, yang artinya jarak antara waktu booking dengan waktu kedatangan lebih dari 1 tahun. Secara umum, hotel memperbolehkan pelanggan untuk membooking hotel maksimal dari 1 tahun sebelumnya. Dengan demikian, nilai lead\_time yang lebih besar dari 365 hari akan didrop.

- Nilai ADR 0 pada segmen pasar selain complementary artinya ada segmen pasar lainnya selain orang-orang yang ditunjuk sebagai promotor yang tidak membayar kamar hotel, dimana ini adalah anomali, sehingga data yang memiliki ADR 0 pada segmen pasar selain complementary akan kita drop.

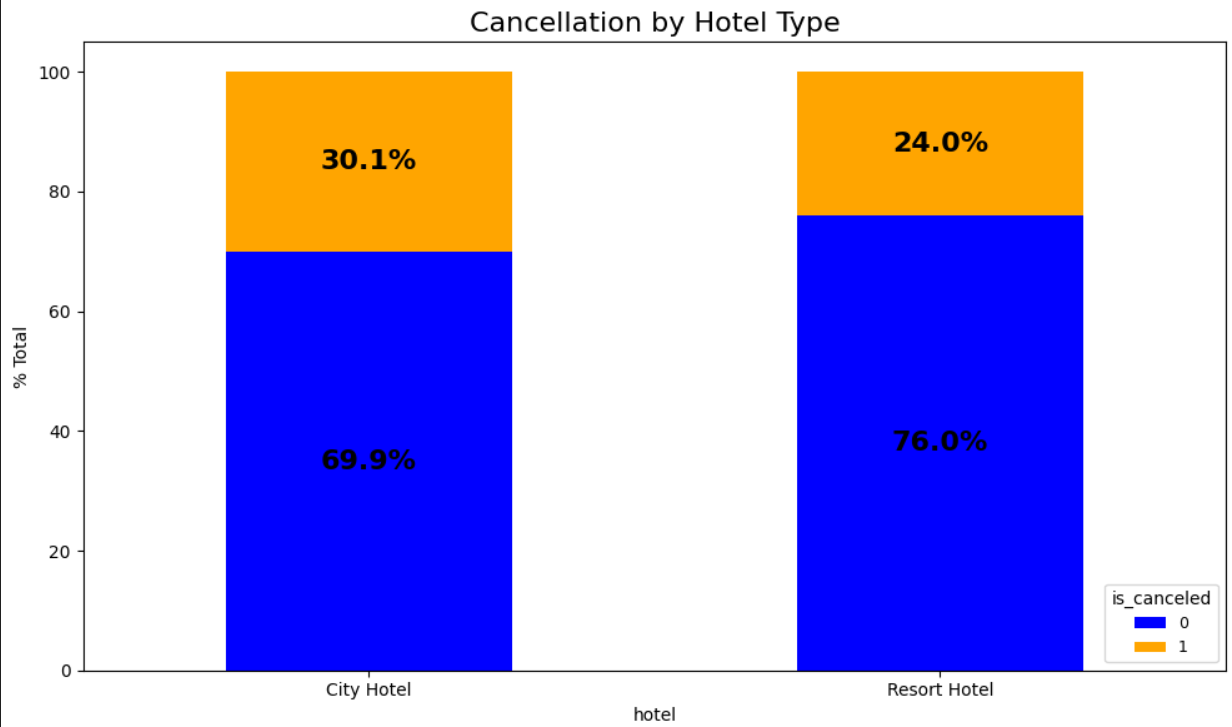
1. Outliers

Pengecekan outliers dilakukan untuk kolom numerik. Pada bagian ini, kita akan drop data yang merupakan outliers ekstrem, seperti pada kolom ADR yang bernilai lebih dari 5000, kolom children dan babies yang nilainya lebih dari 7, dan kolom requires\_car\_parking\_spaces yang lebih dari 7.

1. EDA

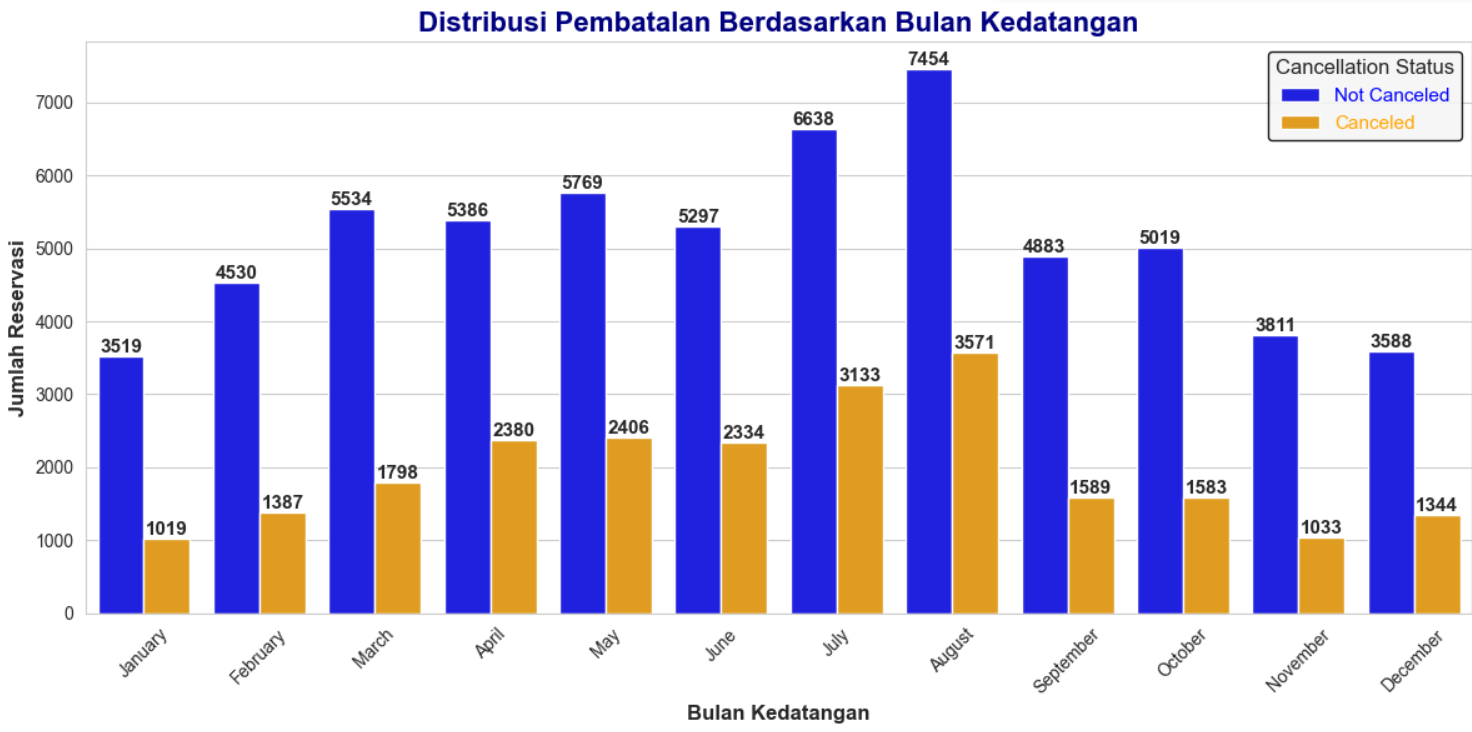
EDA pada proyek ini difokuskan untuk melakukan Diagnostic Analysis dimana kita akan menganalisa faktor-faktor, behaviour pelanggan, serta pola yang mempengaruhi pembatalan pemesanan hotel terhadap kolom `is\_canceled` sebagai variabel target. Analisis ini membantu kita dalam mengidentifikasi insight penting yang nantinya menjadi dasar untuk melakukan pembangunan model Machine Learning menjadi lebih akurat. Berikut pendekatan analisa yang akan kita lakukan.

1. Kita cek persentase jumlah kolom target. Diperoleh bahwa data target imbalanced karena persentase jumlah data target atas tamu yang tidak membatalkan pesanan sebesar 72,3% dan yang membatalkan pesanan sebesar 27,7%.
2. Kita analisis behaviour tamu

- Keputusan pembatalan pelanggan berdasarkan tipe hotel  


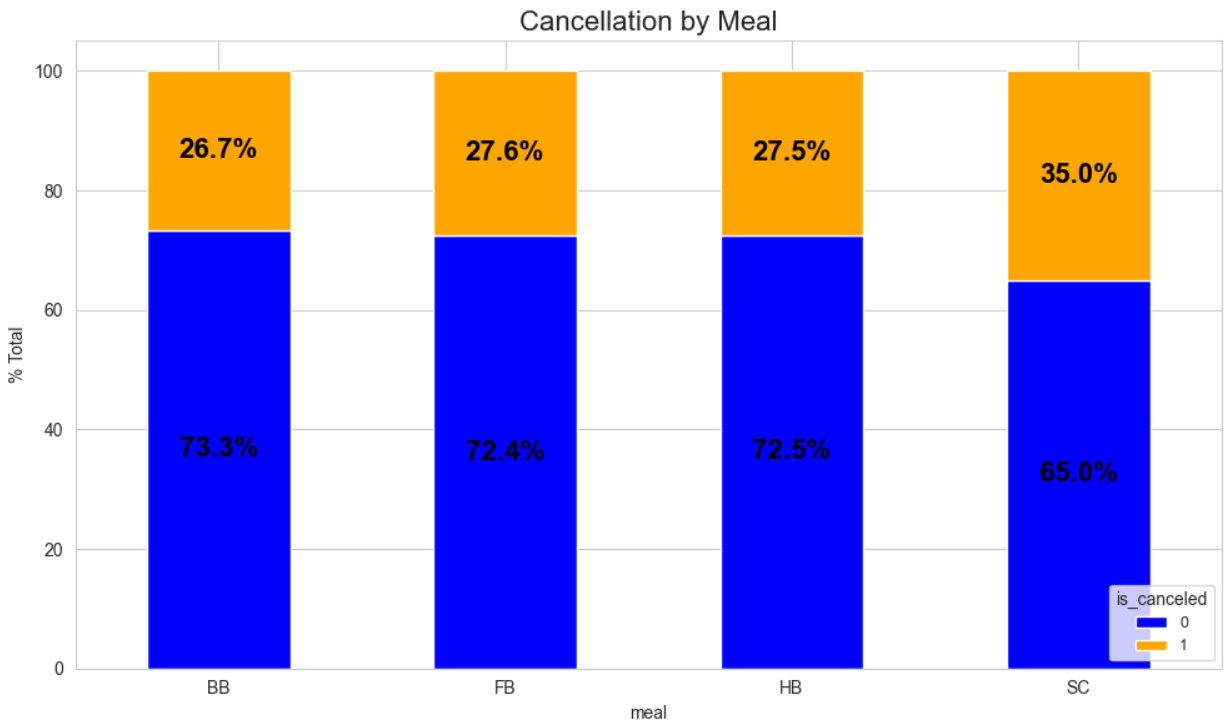
Berdasarkan grafik di atas, pada City Hotel presentase pembatalan yang dilakukan oleh tamu lebih tinggi dibandingkan dengan presentase pembatalan pada Resort Hotel

- Keputusan pembatalan pelanggan berdasarkan bulan kedatangan



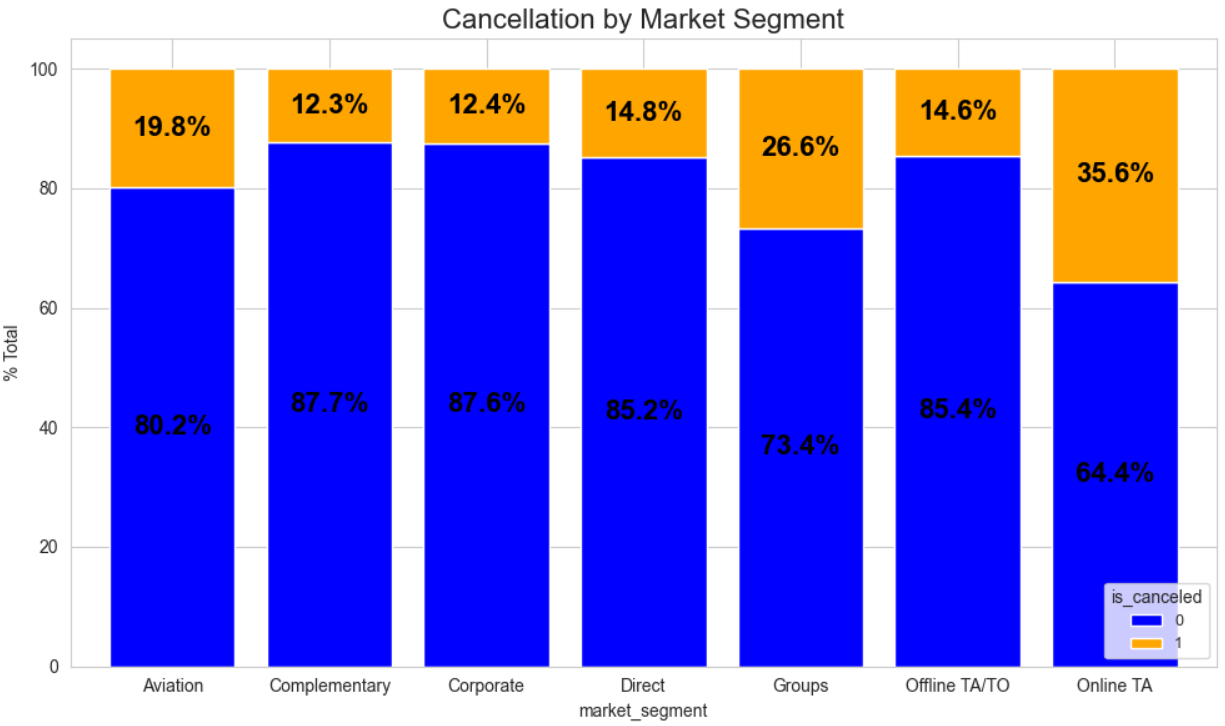
Selama periode Juli 2015- Agustus 2017 Total reservasi tertinggi terjadi pada bulan Juli dan Agustus, yang mana pada bulan-bulan tersebut masuk pada musim panas

- Keputusan pembatalan pelanggan berdasarkan pilihan paket makan



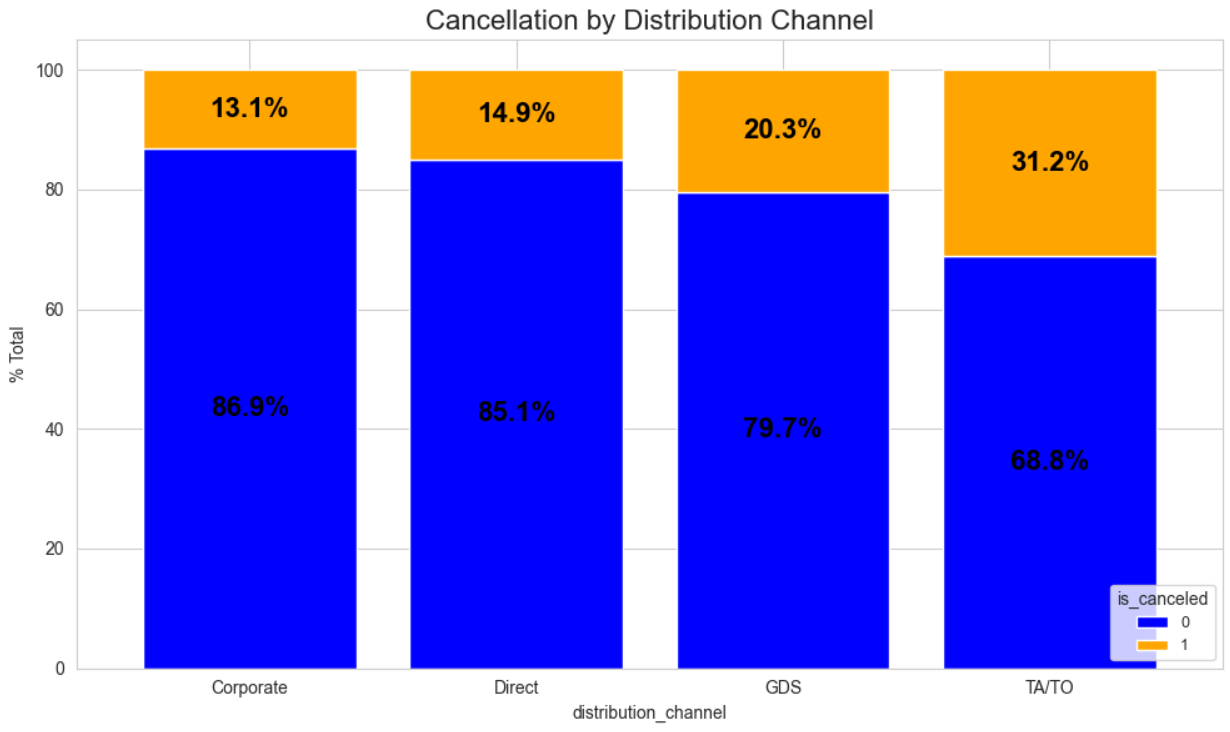
Paket makan SC memiliki tingkat pembatalan tertinggi yaitu sebesar 35%. Paket makan BB, FB, HB cenderung lebih stabil dengan pembatalan di bawah 30%. Tamu yang tidak memilih paket makan (SC) kemungkinan mengalami ketidakpastian dalam perencanaan.

- Keputusan pembatalan pelanggan berdasarkan Segmen Pasar



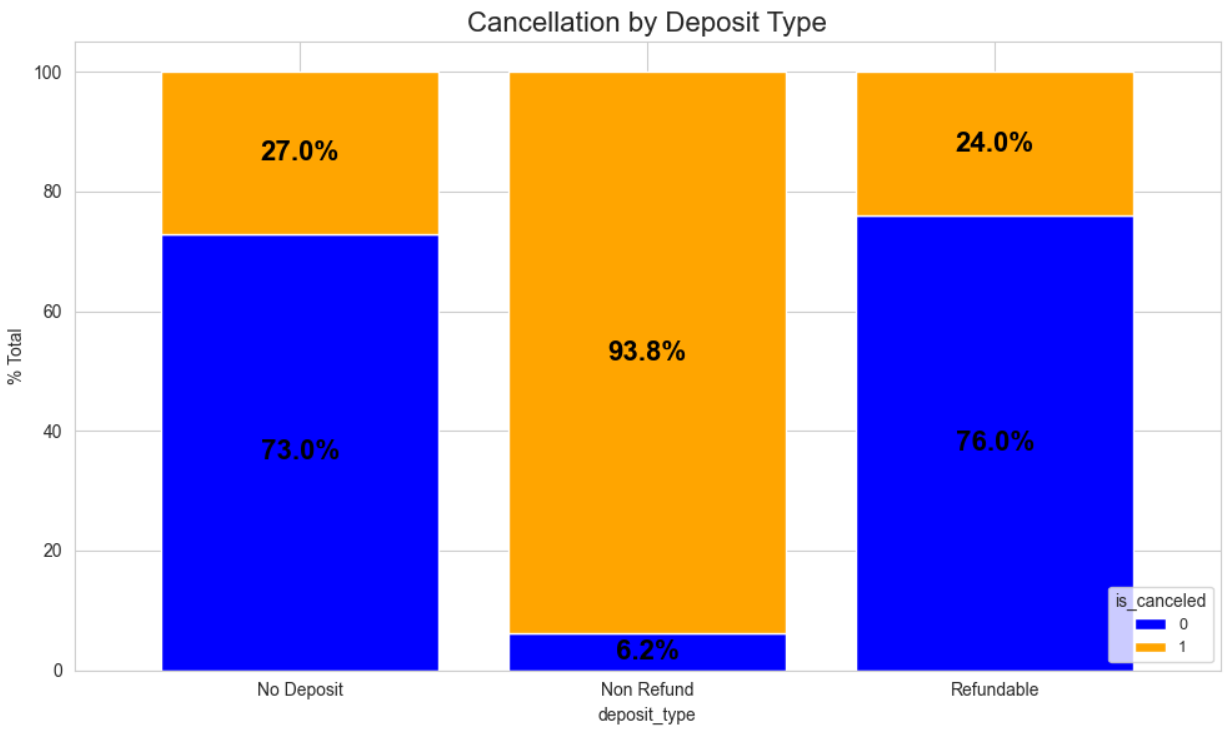
Market segment Online TA memiliki tingkat pembatalan tertinggi sebesar 35.6%, diikuti oleh Groups (26.6%) dan Aviation (19.8%). Hal ini menunjukkan bahwa tamu yang memesan melalui agen perjalanan online lebih cenderung sering melakukan pembatalan dibandingkan segmen lainnya. Kemungkinan penyebabnya adalah fleksibilitas pembatalan, kemudian pada segmentasi Groups mungkin dipengaruhi oleh perubahan jadwal acara, dan Aviation bisa terkait dengan perubahan jadwal penerbangan.

- Keputusan pembatalan pelanggan berdasarkan saluran pemesanan



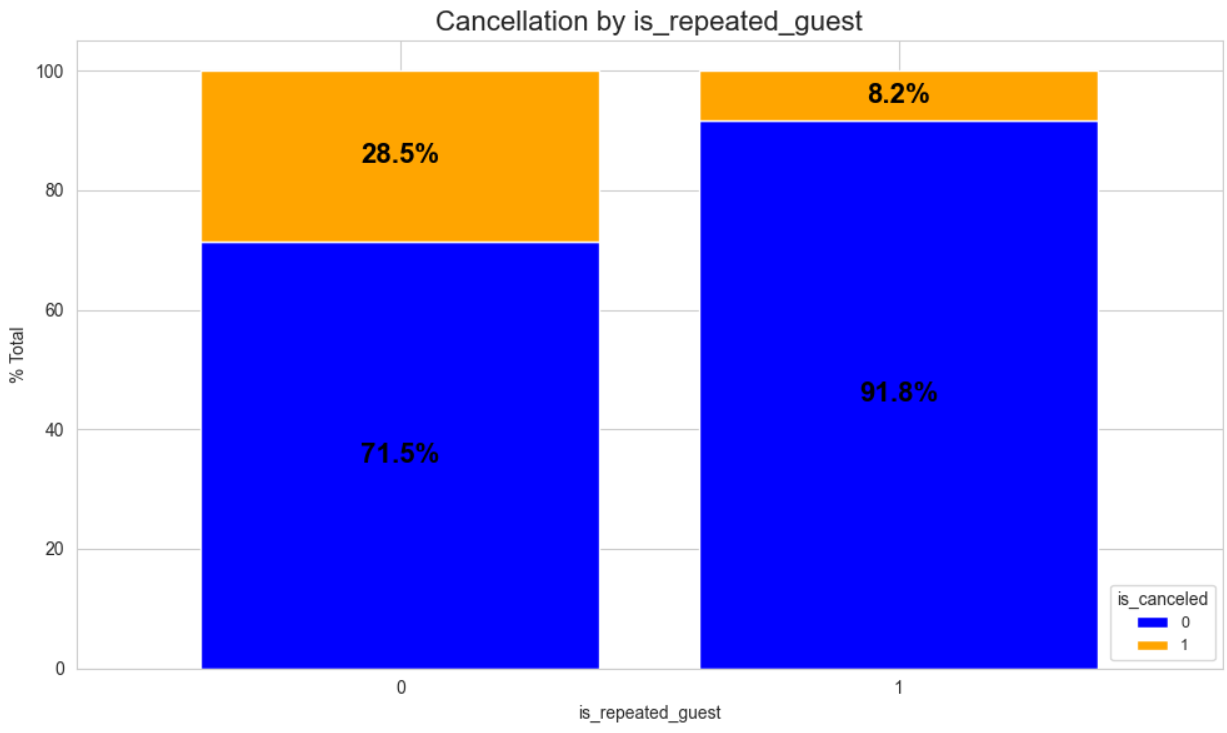
Tamu yang melakukan pemesanan melalui agen perjalanan (TA/TO) memiliki tingkat pembatalan tertinggi (31.2%) dibandingkan dengan saluran distribusi lainnya. kemungkinan hal ini disebabkan oleh kebiasaan pelanggan yang lebih cenderung membandingkan opsi lain sebelum mengonfirmasi pemesanan. sedangkan, pemesanan langsung (14.9%), corporate (13.1%), dan GDS (20.3%) memiliki tingkat pembatalan lebih rendah.

- Keputusan pembatalan pelanggan berdasarkan tipe pembayaran deposit



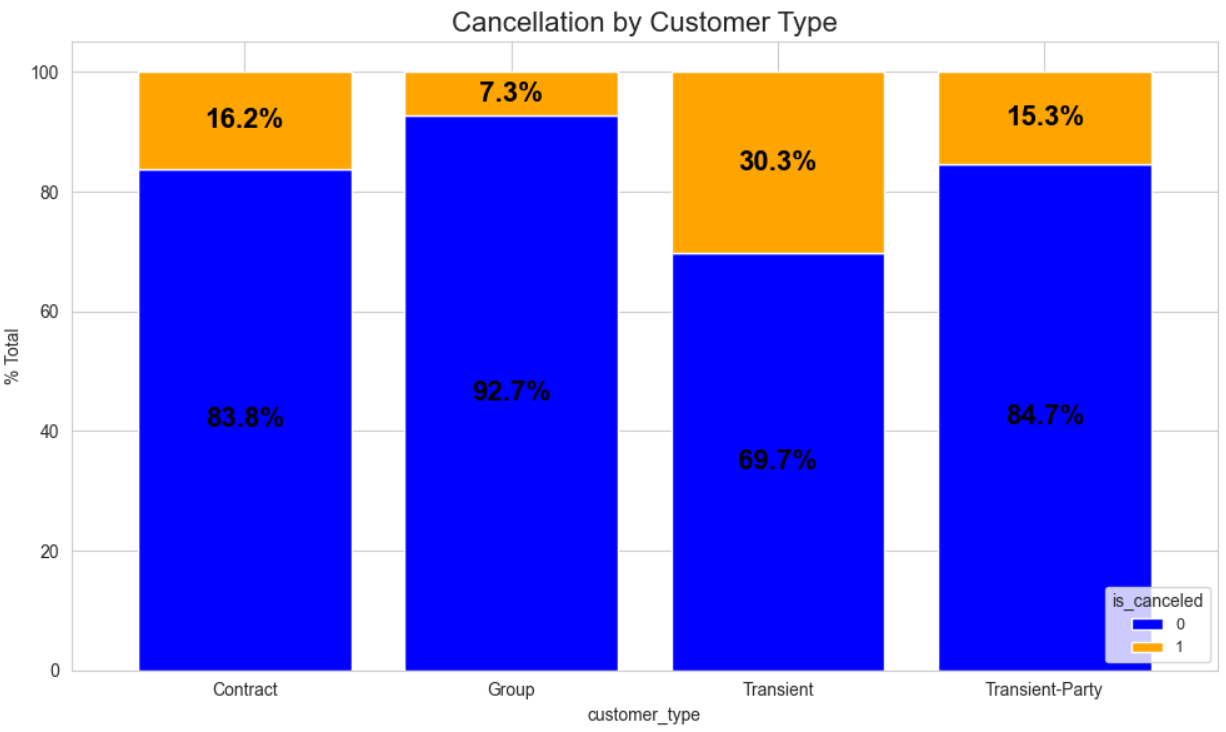
Tamu yang memilih tipe No Deposit memiliki presentase pembatalan sebesar 27%, dan tipe Refundalbe memiliki presentase pembatalan sebesar 24%. tipe Non-Refundable seharusnya mengurangi pembatalan karena tamu sudah membayar di muka, justru data menunjukkan bahwa 93.8% dari pemesan dengan metode ini tetap melakukan pembatalan.

- Keputusan pembatalan pelanggan berdasarkan status tamu tersebut adalah tamu reguler atau bukan



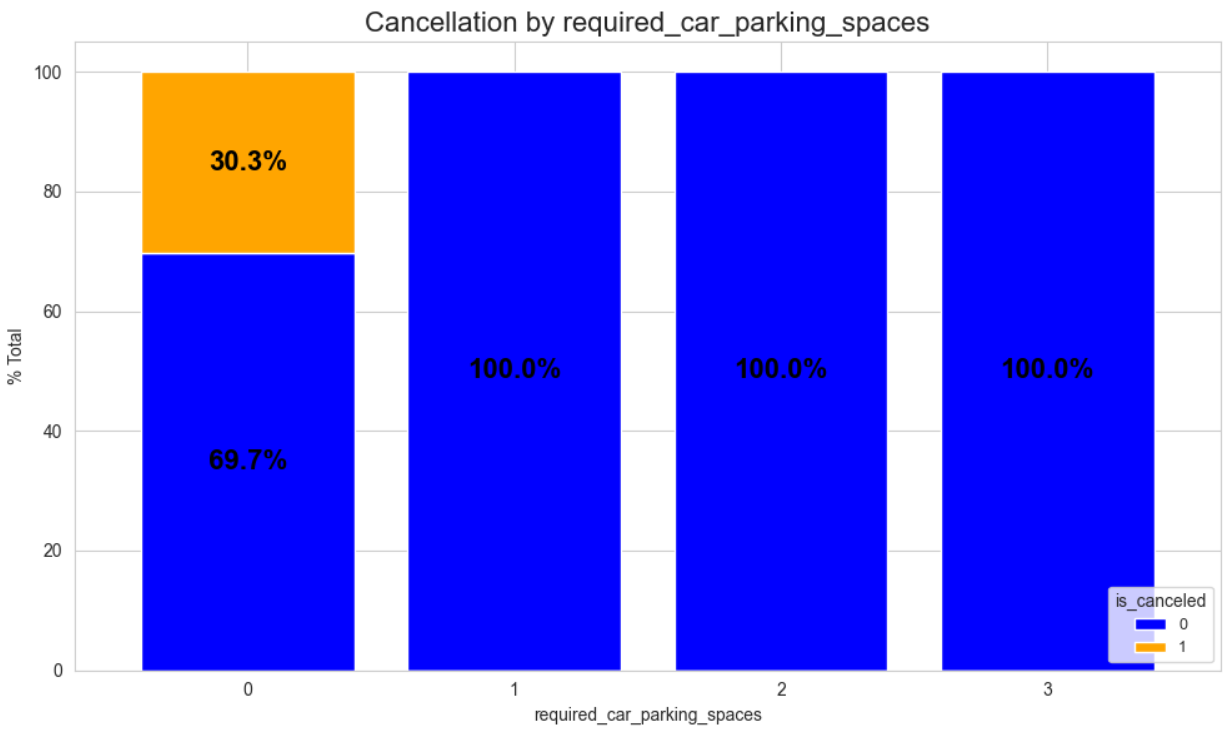
Tamu reguler sangat kecil potensinya melakukan pembatalan (hanya 8%) sedangkan tamu-tamu baru yang bukan tamu reguler memiliki potensi pembatalan sebesar 28%.

- Keputusan pembatalan pelanggan berdasarkan tipe pelanggan



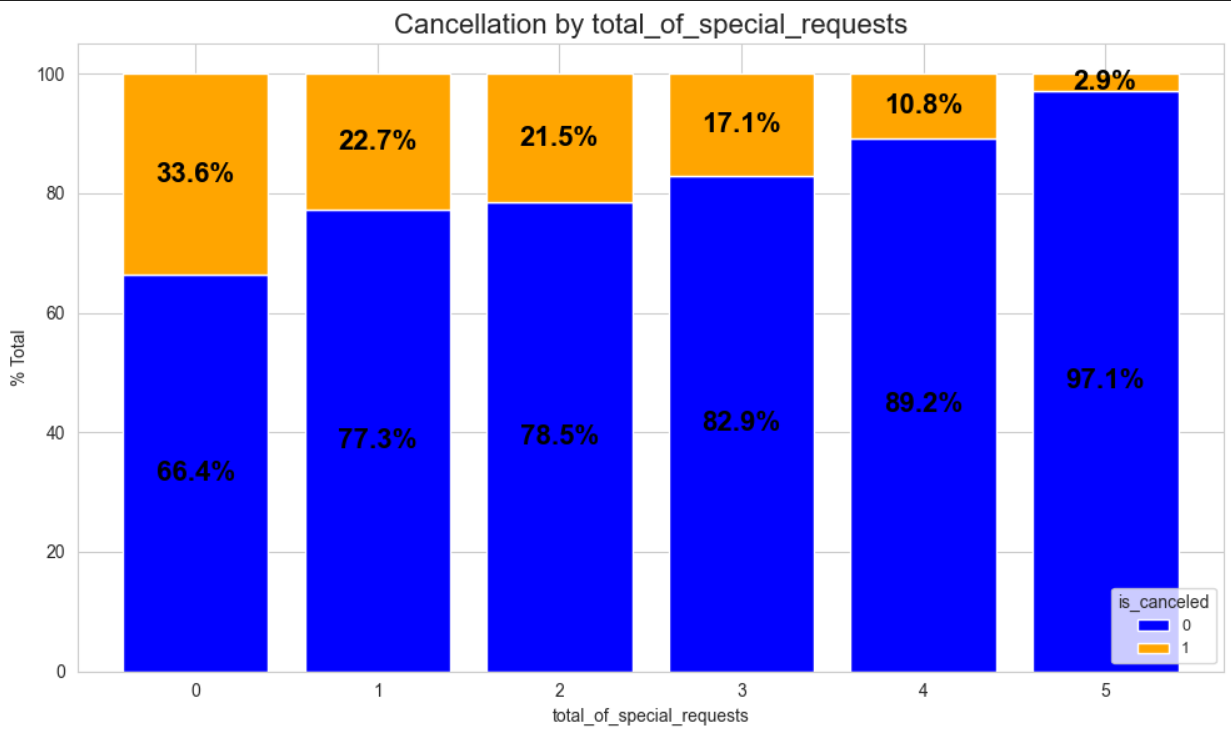
Tipe pelanggan Transient memiliki tingkat pembatalan paling tinggi yaitu sebesar 30,3%, sedangkan untuk tipe customer Group (7%) dan Contract (16%) lebih jarang membatalkan. Pelanggan Transient lebih fleksibel dalam membatalkan dibanding Contract atau Group.

- Keputusan pembatalan pelanggan berdasarkan kebutuhan jumlah space parkir



Berdasarkan permintaan parking space, sebanyak 30% pembatalan dilakukan oleh tamu yang tidak meminta parking space, sedangkan tamu yang meminta parking space mereka adalah tamu yang tidak pernah melakukan pembatalan pemesanan.

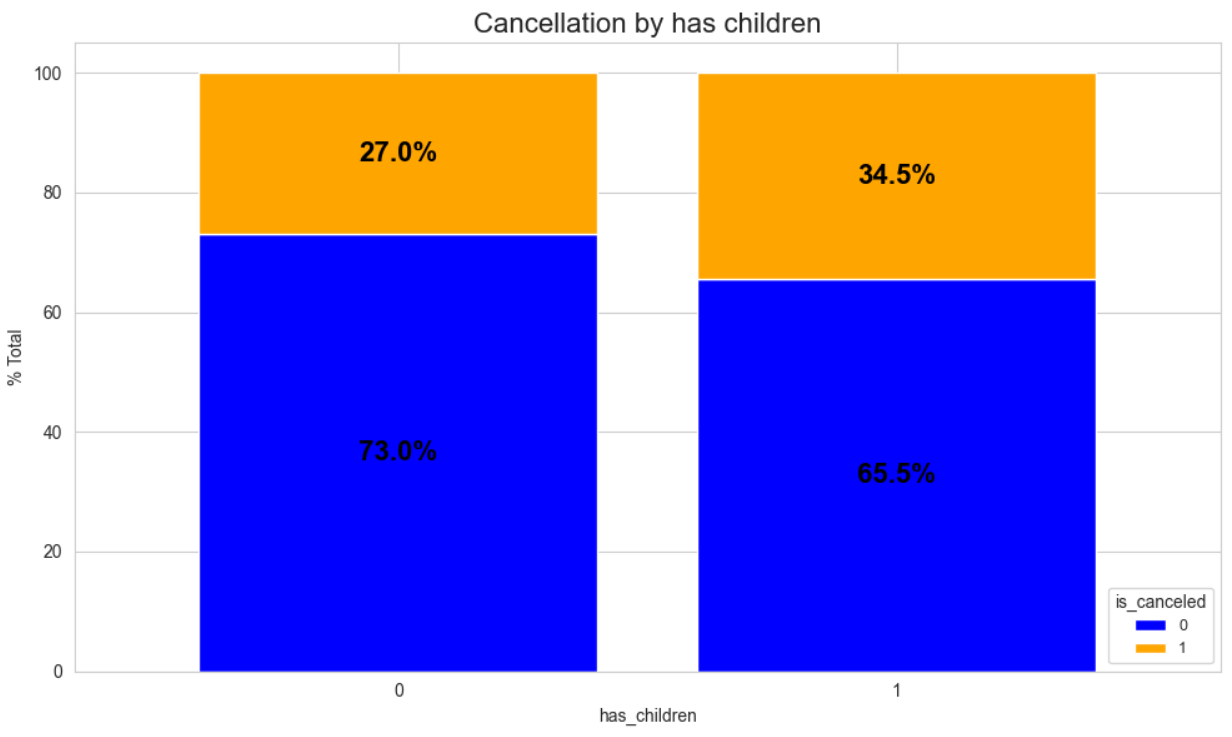
- Keputusan pembatalan pelanggan berdasarkan total permintaan khusus



Tamu-tamu yang memiliki permintaan khusus memiliki komitmen tinggi dengan pesanannya.  Terbukti bahwa semakin banyak permintaan khusus, pembatalan yang terjadi semakin sedikit.

- Keputusan pembatalan pelanggan berdasarkan menginap bersama anak atau tidak

Untuk menganalisis ini, kita membuat satu kolom baru yang menunjukkan apakah mereka datang bersama anak atau tidak, dengan jika ada minimal 1 bayi dan/atau 1 anak-anak, maka mereka membawa anak. Jika jumlah bayi dan anak-anaknya keduanya nol, artinya mereka tidak membawa anak.



Presentase tamu yang membawa anak (has\_children = 1) cenderung lebih tinggi untuk melakukan pembatalan pemesanan yaitu sebesar 34.5%, dibandingkan dengan tamu yang tidak membawa anak (has\_children = 0) dengan presentase pembatalan sebesar 27%.

- Keputusan pembatalan pelanggan berdasarkan perubahan pada pemesanan

Untuk menganalisis hal ini, kita buat kategori sebagai berikut.

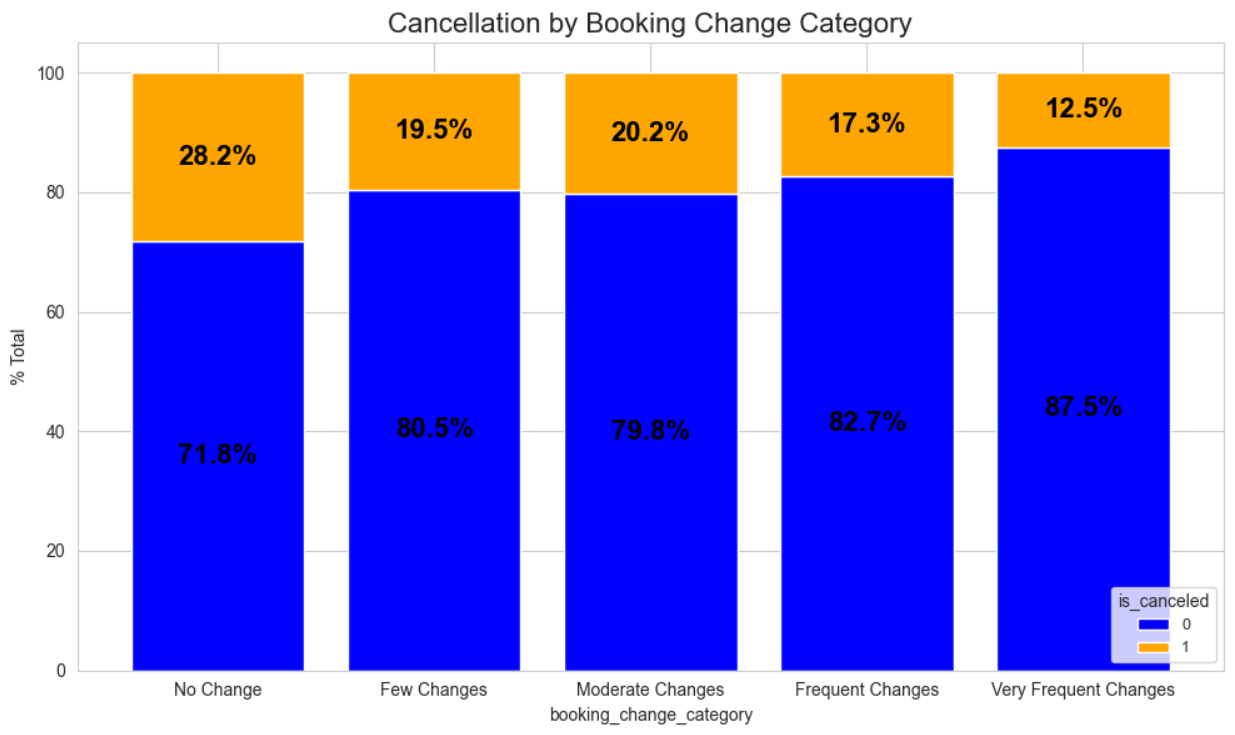
- Tidak ada perubahan artinya nilai booking\_changes = 0

- Jarang melakukan perubahan untuk nilai booking\_changes = 1 sampai 2

- Frekuensi melakukan perubahan sedang untuk nilai booking\_changes = 3 sampai 5

- Sering melakukan perubahan untuk nilai booking\_changes = 6 sampai 10

- Sangat sering melakukan perubahan untuk nilai booking\_changes lebih dari 10



Semakin sedikit perubahan pada pemesanan maka semakin tinggi presentase pembatalannya

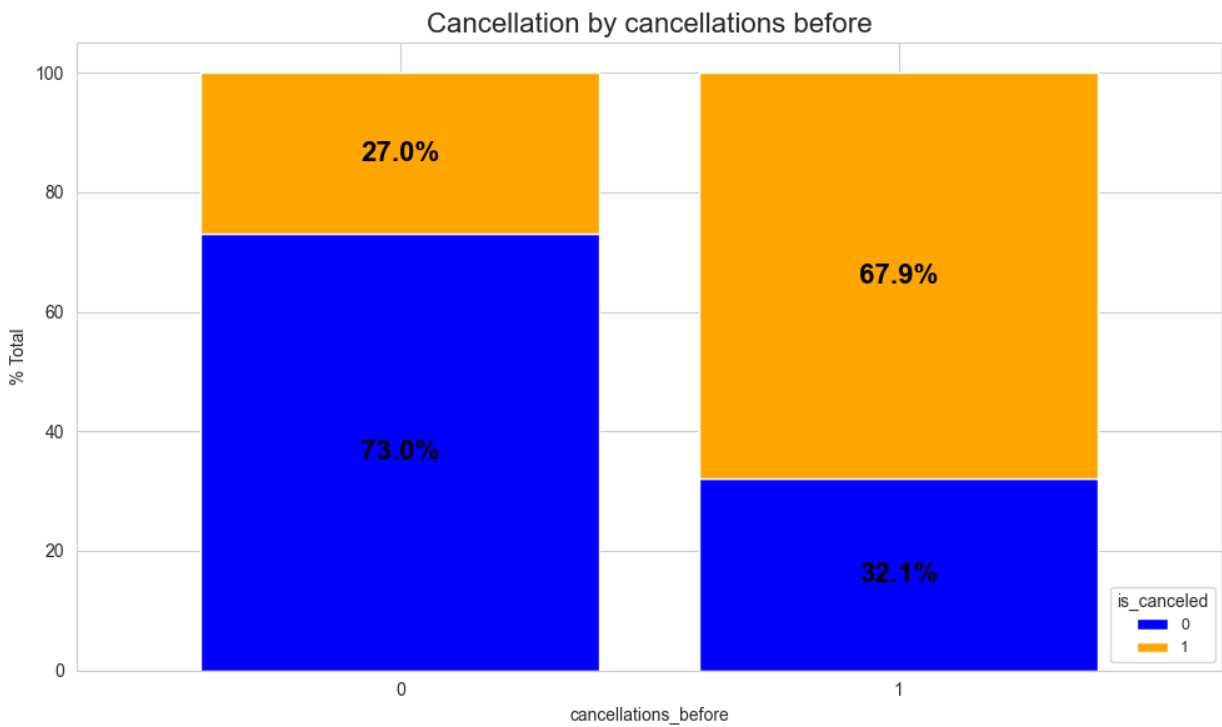
- Keputusan pembatalan pelanggan berdasarkan pernah tidaknya pelanggan membatalkan pesanan sebelumnya

Untuk menganalisis hal ini, kita akan menambah kolom baru yang menunjukkan apakah pelanggan tersebut sebelumnya pernah membatalkan pesanan atau tidak dengan :

- 1 : pelanggan pernah membatalkan pemesanan

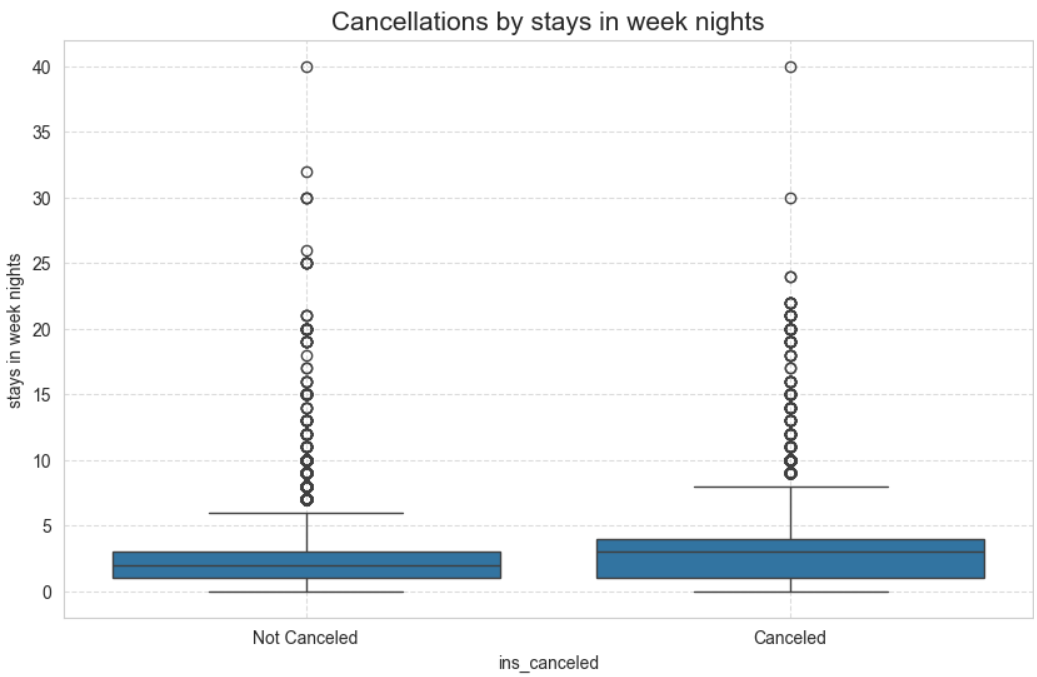
- 0 : pelanggan belum pernah membatalkan pemesanan

Pelanggan yang pernah membatalkan pemesanan, nilai pada kolom previous\_cancellations akan lebih dari 0, sedangkan yang belum pernah membatalkan pemesanan, nilai pada kolom previous\_cancellations akan sama dengan 0.



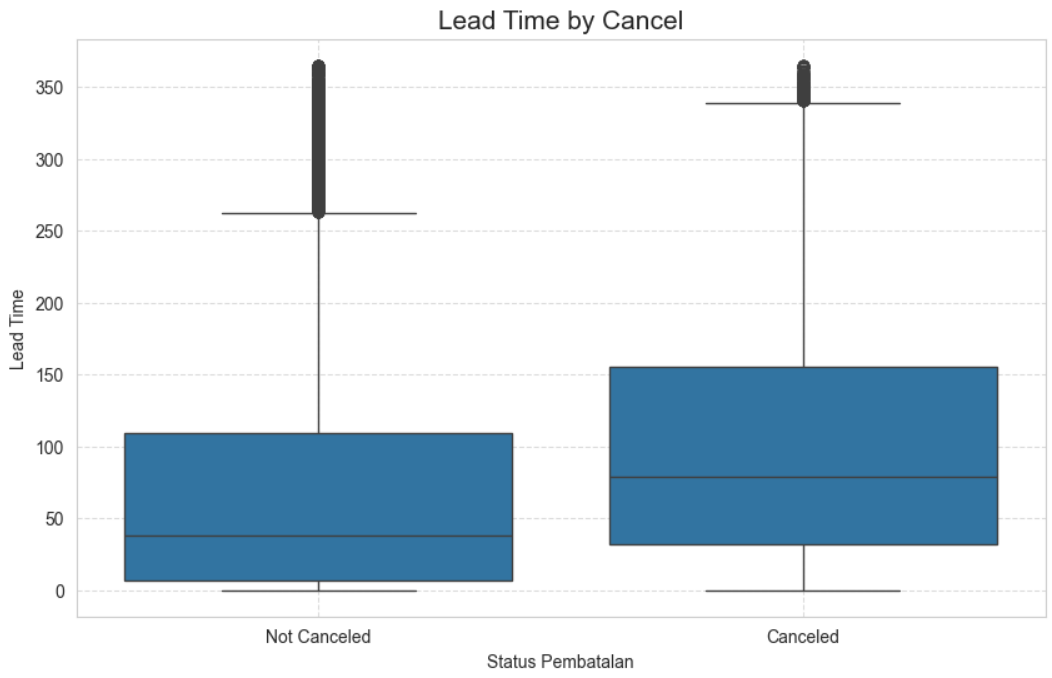
Tamu yang sebelumnya pernah melakukan pembatalan pemesanan cenderung lebih tinggi persentase tingkat pembatalannya yaitu sebesar 67.9% dibandingkan dengan tamu yang belum pernah melakukan pembatalan pemesanan pada pemesanan sebelumnya yang mana tingkat pembatalannya hanya sebesar 27%.

- Keputusan pembatalan pelanggan berdasarkan lama waktu menginap di hari Senin-Jumat



Median lama waktu menginap di antara hari Senin-Jumat untuk tamu yang membatalkan pemesanannya lebih tinggi daripada median lama waktu menginap untuk tamu yang tidak membatalkan pemesanannya. Mayoritas tamu yang membatalkan pemesanannya adalah yang lama waktu menginapnya 3 malam, sedangkan yang tidak membatalkan pemesanannya adalah yang lama waktu menginapnya hanya 2 malam.

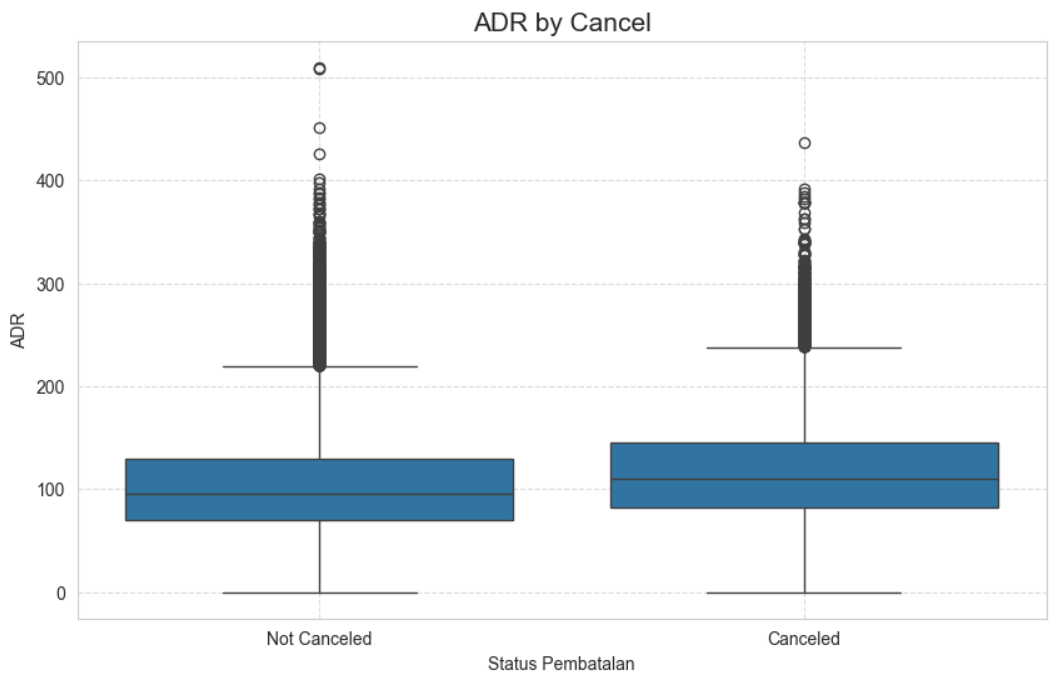
- Keputusan pembatalan pelanggan berdasarkan jarak waktu pemesanan hingga tanggal kedatangan



- Tamu yang melakukan pembatalan pemesanan adalah mayotitas mereka adalah yang jarak waktu pemesanan sampai waktu kedatangan adalah 79 hari.

- Tamu yang tidak melakukan pembatalan pemesanan mayoritas mereka adalah yang jarak waktu pemesanan sampai waktu kedatangan adalah 38 hari.

- Keputusan pembatalan pelanggan berdasarkan ADR (harga rata-rata per kamar per hari)



- Tamu yang melakukan pembatalan pemesanan mayoritas dari mereka adalah yang harga kamarnya 110 Euro

- Tamu yang tidak melakukan pembatalan pemesanan mayoritas dari mereka adalah yang harga kamarnya adalah 95 Euro

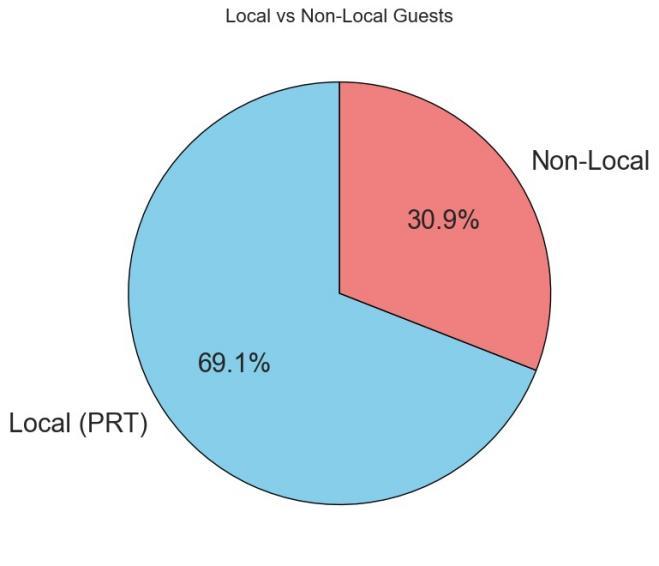
Tamu yang melakukan pembatalan pemesanan memiliki median ADR yang lebih tinggi dari pada tamu yang melakukan pembatalan

- Keputusan pembatalan pelanggan berdasarkan asal negara

Untuk menganalisis hal ini, kita akan membuat satu kolom baru yang menunjukkan apakah tamu ini berasal dari dalam negeri, artinya berasal dari Portugal, ataukah berasal dari luar negeri atau berasal dari luar Portugal. Kolom tersebut kita namakan is\_local. Isi dari kolom is\_local adalah :

- 1 : tamu berasal Portugal (dari dalam negeri)

- 0 : tamu berasal dari luar Portugal (dari luar negeri)

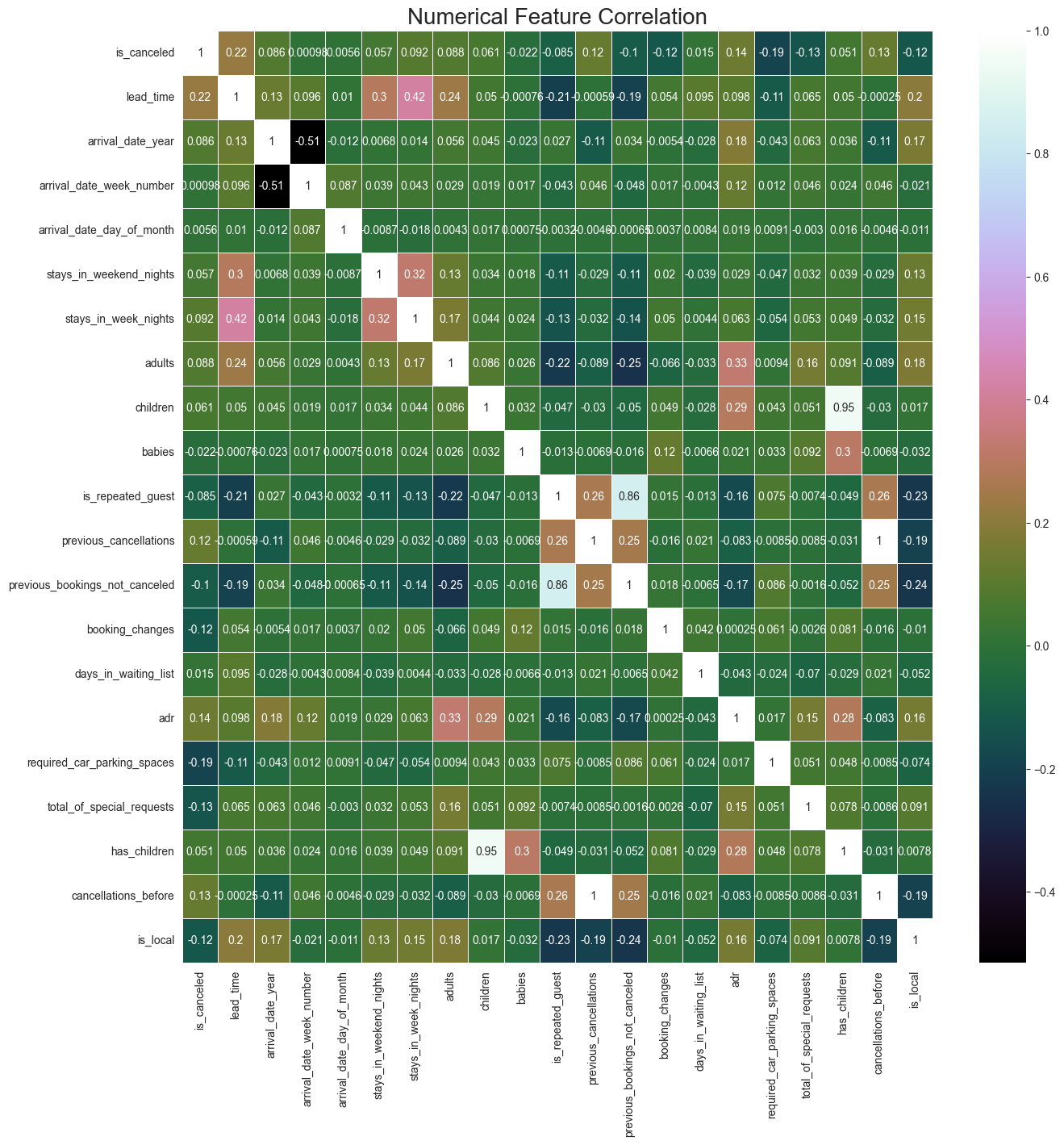


Dari pie chart di atas, diperoleh tamu yang berasal dari Portugal lebih banyak daripada tamu yang berasal dari luar Portugal

1. Membangun Model
2. Data Correlation and Features Selection

Setiap hubungan baik data kategorik maupun numerik terhadap kolom target (is\_canceled). Pada kolom numerik menggunakan heatmap terhadap kolom is\_canceled. Sedangkan untuk kolom kategorik menggunakan metode chi-square dimana bila p-value < 0.05 maka kolom tersebut memiliki pengaruh terhadap kolom is\_canceled.

* Korelasi pada kolom numerik terhadap kolom target.



Diperoleh kolom numerik yang akan dijadikan feature adalah :

1. required\_car\_parking\_spaces

2. is\_local

3. total\_of\_special\_requests

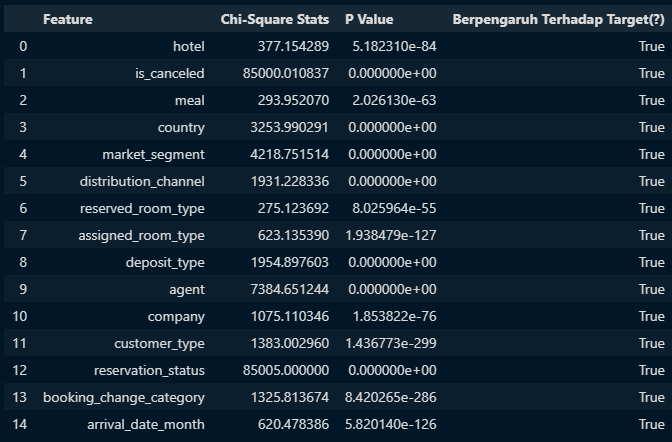
4. cancellations\_before

5. adr

6. lead\_time

Hal ini dikarenakan nilai korelasi untuk fitur-fitur tersebut dengan kolom target (`is\_canceled`) lebih dari 10%

* Korelasi kolom kategorik terhadap kolom target



* Kolom reserved\_room\_type akan didrop dari fitur karena hanya menunjukkan jenis kamar yang dipesan, sedangkan assigned\_room\_type lebih relevan karena mencerminkan jenis kamar yang benar-benar diterima oleh tamu. Oleh karena itu, analisis akan difokuskan pada kamar yang sebenarnya didapatkan
* Kolom company akan didrop dari feature karena sekitar 94% pelanggan adalah yang tidak berkaitan dengan perusahaan tertentu.
* Kolom reservation\_status akan didrop dari feature karena kolom ini menggambarkan kolom is\_canceled.
* Kolom country akan dihapus karena telah digantikan oleh kolom is\_local, yang secara langsung merepresentasikan apakah tamu berasal dari dalam atau luar negeri. Hal ini dilakukan untuk menyederhanakan analisis tanpa kehilangan informasi penting.
* Kolom is\_canceled merupakan kolom target, sehingga tidak dijadikan fitur

Kolom-kolom yang bersifat kategorik yang akan dijadikan feature adalah :

1. hotel
2. arrival\_date\_month
3. meal
4. market\_segment
5. distribution\_channel
6. assigned\_room\_type
7. deposit\_type
8. agent
9. customer\_type
10. booking\_change\_category
11. Modeling

* Preprocessing

Pada proses ini, hal yang perlu kita lakukan adalah,

1. Menentukan kolom fitur dan target

Dimana kolom x adalah fitur, dan kolom y adalah target (is\_canceled)

1. Split data

Data dibagi menjadi data train dan terest dengan perbandingan data train dan data test yaitu 80% : 20%.

1. Skema Preprocessing

Pada tahap ini, perlu dilakukannya feature engineering dimana kolom fitur akan dilakukan proses encoding dan scalling bergantung pada tipe kolom fitur tersebut.

1. One Hot Encoding

Kolom yang dilakukan proses dengan OHE adalah kolom 'hotel', 'meal', 'distribution\_channel', 'deposit\_type', 'customer\_type'.

1. Binary Encoding

kolom yang dilakukan proeses Binary Encoding adalah 'market\_segment', 'assigned\_room\_type', 'agent', 'arrival\_date\_month', 'booking\_change\_category'

1. Robust Scaler

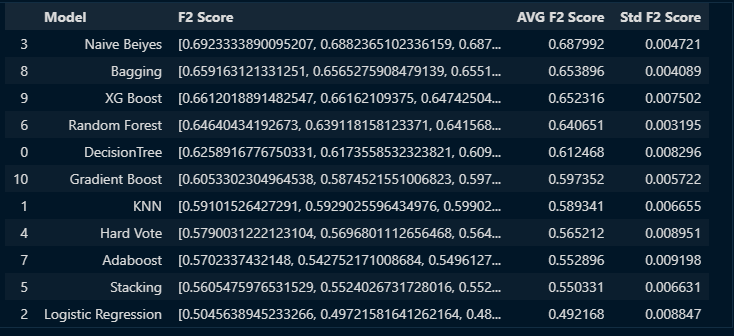
Kolom yang di

* Model

Pada proyek ini, kita akan memprediksi seorang tmau akan membatalkan pesanan atau tidak. Maka dari itu, model yang digunakan adalah model klasifikasi.

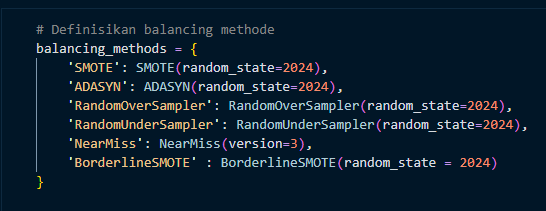
* Pipeline

Proses ini dilakukan untuk melihat f2 score jika data tidak diseimbangkan lebih dulu, mengingat data target yang jumlahnya tidak seimbang.



* Metode Balancing pada Target

Balancing dilakukan untuk menyeimbangkan data target. Metode yang dilakukan seperti gambar dibawah ini.



Kemudian model ini juga diproses dengan Pipeline dan memperoleh.

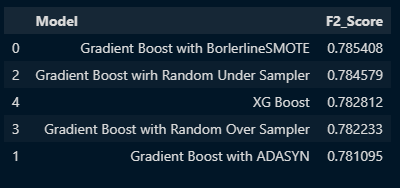


Dari perbandingan model dengan balancing, diperoleh lima model dengan rata-rata F2 score tertinggi, yaitu:

1. XG Boost dengan metode balancing Random Under Sampler
2. Gradient Boost dengan metode balancing BorderlineSMOTE
3. Gradient Boost dengan metode balancing Random Under Sampler
4. XG Boost dengan metode balancing Random over Sampler
5. Gradient Boost dengan metode balancing ADASYN

* Hyperparameter Tuning

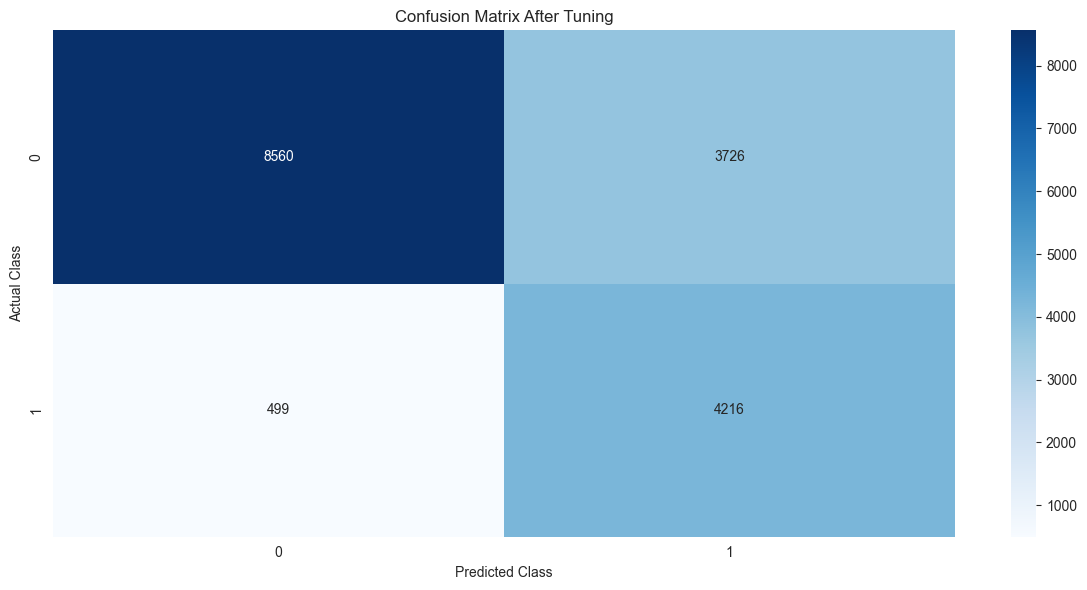
Model dengan nilai rata-rata F2 score tertinggi yang disebutkan sebelumnya, dilakukan tuning menggunakan RandomizedSearch dan menggunakan Cross Validation. Kemudian diperoleh hasil sebagai berikut.



1. Prediksi Data Uji

Dari hasil hyperparameter tuning, dipilih model Gradient Boost dengan metode balancing Border Line Smote karena model tersebut memiliki f2 score tertinggi. Diperoleh f2 score dengan model Gradient Boosting dengan metode Borderline SMOTE tanpa hyperparameter tuning adalah 0,78.

1. Confusion Matrix



Model **sangat baik dalam mendeteksi pembatalan** (F-2 Score tinggi). **Cocok untuk kasus di mana gagal mendeteksi pembatalan (FN)** lebih rendah nilainya daripada kesalahan memprediksi batal yang tidak terjadi (FP).

1. Limitasi Model
2. Model ini terbatas hanya jika informasi yang diketahui adalah :

* 'hotel'
* 'lead\_time'
* 'arrival\_date\_month'
* 'meal'
* 'market\_segment'
* 'distribution\_channel'
* 'cancellations\_before'
* 'booking\_change\_category'
* 'assigned\_room\_type'
* 'deposit\_type'
* 'agent'
* 'customer\_type'
* 'adr'
* 'required\_car\_parking\_spaces'
* 'total\_of\_special\_requests'
* 'is\_local'
* 'is\_canceled'

Dengan demikian, untuk memprediksi keputusan pembatalan di luar informasi tersebut, peluang akurasi terhadap prediksi tidak akurat akan lebih besar.

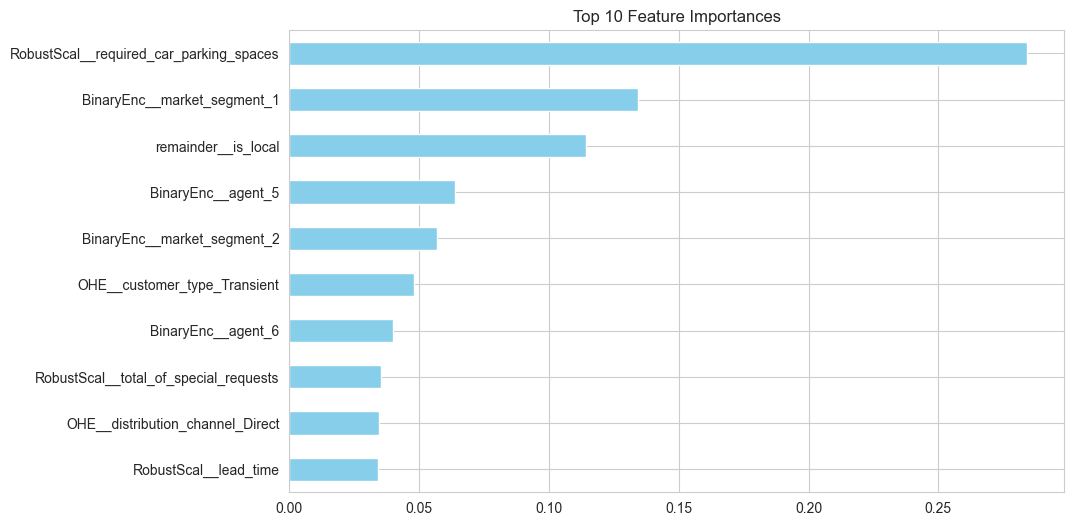
1. Data targetnya tidak seimbang. Data tamu yang membatalkan lebih sedikit daripada tamu yang tidak membatalkan
2. Karena keterbatasan waktu, model yang diuji coba terbatas, yaitu

* DecisionTree,
* KNN,
* Logistic Regression
* Naive Beiyes
* Hard Vote
* Stacking
* Random Forest
* Adaboost
* Bagging
* XG Boost
* Gradient Boost

1. Karena keterbatasan waktu, metode hyperparameter tuning yang dapat digunakan adalah randomized search
2. Features Importance

Diperoleh fitur-fitur yang paling berpengaruh terhadap model adalah :

1. required\_car\_parking\_spaces
2. market\_segment
3. is\_local
4. agent
5. customer\_type
6. total\_of\_special\_requests
7. distribution\_channel
8. lead\_time



1. Model Impact

Dengan model ini, pihak manajemen mampu mengetahui prediksi tamu-tamu yang berpotensi untuk melakukan pembatalan. Selain itu, pihak manajemen juga akan mengetahui pelanggan mana yang berpotensi untuk tetap melanjutkan pesanannya sehingga pihak manajemen dapat melakukan strategi yang lebih efektif untuk mengatasi demand di hotel mereka.

1. Kesimpulan
2. Pelanggan akan cenderung membatalkan pemesanannya jika:

- jarak antara waktu pemesanan dengan waktu kedatangan relatif lama

- harga kamarnya relatif mahal

- pelanggan tersebut berasal dari dalam negeri (dari Portugal)

- pelanggan akan datang pada musim panas (Juni-Agustus)

- pelanggan memesan melalui agen perjalanan online

- pelanggan merupakan pelanggan baru di hotel tersebut

- pelanggan bepergian sendiri, bukan merupakan bagian dari grup ataupun perusahaan

- pelanggan tidak membutuhkan space parkir

- pelanggan tidak memiliki permintaan khusus

- pelanggan tidak memesan paket makan

- pelanggan membawa anak untuk menginap di hotel tersebut

- pelanggan sedikit atau bahkan tidak melakukan perubahan pada pemesanannya

- pelanggan memiliki riwayat membatalkan pemesanan di hotel tersebut sebelumnya

- pelanggan memesan kamar untuk tinggal pada rentang hari Senin-Jumat dalam waktu yang panjang

1. Faktor-faktor yang berpengaruh terhadap keputusan pembatalan pemesanan kamar hotel adalah:

- Kebutuhan space parkir

- jarak antara waktu pemesanan dengan waktu kedatangan (lead time)

- ADR (harga rata-rata per kamar per hari)

- negara asal pelanggan

- bulan kedatangan

- apakah menginap bersama anak atau tidak

- merupakan tamu reguler di hotel tersebut atau tidak

- tipe pelanggan yang bepergian (grup, sendiri, dll)

- saluran pemesanan yang digunakan oleh pelanggan

1. Dari hasil membangun model, diperoleh model yang paling baik adalah XG Boost dengan dilakukan balancing menggunakan metode Random Under Sampler, juga dilakukan hyperparameter tuning. Hal ini terlihat dari nilai F2 Score yang paling tinggi daripada model lainnya, yaitu sebesar 0,786.
2. Model ini dapat diaplikasikan oleh pihak manajemen hotel karena setelah dilakukan tes ke data uji, nilai F2 Score tidak berbeda jauh dengan saat dilakukan tes ke data latih, yaitu sebesar 0,787. Hanya saja, model ini baik diaplikasikan jika terdapat informasi berikut.

- 'hotel'

- 'lead\_time'

- 'arrival\_date\_month'

- 'meal'

- 'market\_segment'

- 'distribution\_channel'

- 'cancellations\_before'

- 'booking\_change\_category'

- 'assigned\_room\_type'

- 'deposit\_type'

- 'agent'

- 'customer\_type'

- 'adr'

- 'required\_car\_parking\_spaces'

- 'total\_of\_special\_requests'

- 'is\_local'

- 'is\_canceled'

1. Saran Dan Rekomendasi
2. Saran
3. Berdasarkan analisis, diperoleh bahwa pelanggan cenderung membatalkan pemesanannya jika jarak antara waktu pemesanan dengan waktu kedatangan relatif lama. Mayoritas pelanggan yang membatalkan pesanannya memiliki jarak waktu pemesanan dengan waktu kedatangan sebesar 79 hari, sedangkan mayoritas pelanggan yang tidak membatalkan pemesanannya memiliki jarak waktu pemesanan dengan waktu kedatangan sebesar 38 hari. Dari data ini, pihak manajemen hotel dapat membatasi waktu pemesanannya, yaitu maksimal 40 hari sebelum waktu kedatangan.
4. Berdasarkan analisis, diperoleh bahwa pelanggan cenderung membatalkan pemesanannya jika harga kamarnya relatif tinggi. Mayoritas pelanggan yang membatalkan pemesanannya adalah yang mendapatkan harga kamar 110 Euro, sedangkan mayoritas pelanggan yang tidak membatalkan pemesanannya adalah yang harga kamarnya 95 Euro. Dari hasil analisis ini, pihak manajemen hotel dapat mempertimbangkan rate harga kamar atau memberikan diskon.
5. Berdasarkan analisis, pelanggan yang cenderung membatalkan pemesanan adalah pelanggan baru atau bukan pelanggan reguler, sehingga pihak manajemen hotel dapat mempertimbangkan untuk membuat member bagi para pelanggan reguler, dimana mereka akan mendapatkan harga kamar yang lebih menarik dan promosi khusus, sehingga pelanggan baru dapat tertarik untuk menjadi member hotel tersebut.
6. Berdasarkan analisis, pelanggan yang cenderung membatalkan pemesanan kamarnya adalah pelanggan yang membawa anak. Pihak manajemen hotel dapat mempertimbangkan untuk membuat fasilitas yang ramah anak, seperti adanya tempat bermain anak-anak, kolam renang yang aman untuk anak-anak, makanan yang sesuai dengan mayoritas lidah anak-anak, sehingga para orangtua yang ingin membawa anaknya menginap dapat merasa aman dan nyaman untuk menginap di hotel tersebut bersama anak mereka, juga anak-anak akan lebih tertarik untuk menginap di hotel tersebut. Karena bisa jadi, tempat menginap adalah pilihan dari anak-anak mereka.
7. Berdasarkan analisis, musim yang paling ramai pelanggan adalah pada musim panas, yaitu pada bulan Juni hingga Agustus, sehingga pihak manajemen hotel dapat memberikan promosi khusus untuk menginap di hotel tersebut saat musim panas, seperti harga yang lebih murah atau mendapatkan fasilitas tambahan secara gratis, seperti fasilitas breakfast.
8. Berdasarkan analisis, pelanggan akan cenderung membatalkan pesanan jika menginap di rentang hari Senin-Jumat dalam waktu yang relatif panjang. Pelanggan yang membatalkan pesanannya untuk menginap di rentang hari Senin-Jumat mayoritas yang memesan untuk menginap selama 3 malam, sedangkan pelanggan yang tidak membatalkan pesanannya mayoritas yang memesan untuk menginap selama 2 hari. Pihak manajemen hotel dapat mempertimbangkan untuk memberikan harga khusus jika menginap 3 hari atau lebih di rentang hari Senin-Jumat, namun dengan adanya deposit di awal dan terdapat cancellation fee.
9. Berdasarkan analisis, pelanggan akan cenderung membatalkan pesanan jika memesan kamar melalui agen perjalanan online. Pihak manajemen hotel dapat menerapkan kebijakan cancellation fee yang lebih tinggi jika pelanggan memesan kamar melalui agen perjalanan online.
10. Rekomendasi

Rekomendasi ini bertujuan untuk meningkatkan keakurasian dan ketepatan model dalam memprediksi pembatalan pemesanan, serta mengurangi kesalahan prediksi yang dapat berdampak pada pengambilan keputusan bisnis.

1. Mencoba algoritma boosting lain yang belum digunakan untuk melihat apakah dapat meningkatkan performa model.
2. Menambah jumlah data, terutama untuk kategori canceled, agar model memiliki lebih banyak informasi dan dapat membuat prediksi yang lebih akurat.
3. Menganalisis data yang masih salah diprediksi, terutama False Positive dan False Negative, untuk memahami pola kesalahan dan mencari cara memperbaikinya.
4. Menggunakan data yang lebih baru agar dapat lebih menggambarkan situasi dan kondisi saat ini.