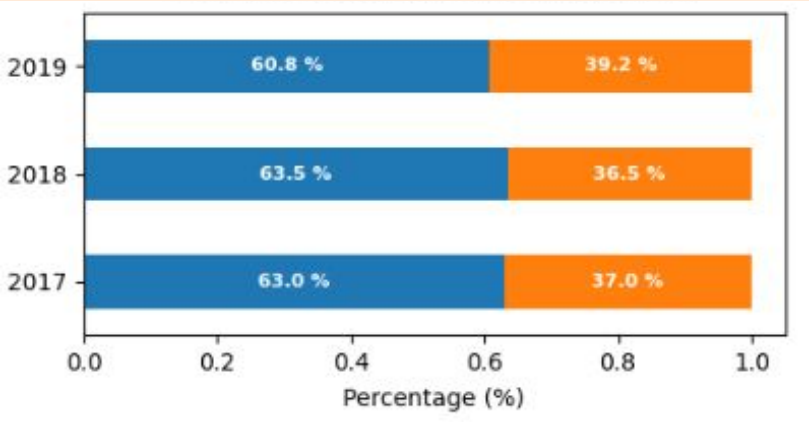
Exploring Customer Characteristics and Predicting Hotel Booking Cancellations Using Machine Learning

By: Ramadhoni Nasri

Latar Belakang Masalah



Tingkat pembatalan pemesanan kamar hotel (cancelation rate) di tahun **2017** mencapai angka **37%**. Kemudian pada tahun **2018**, tingkat cancelation rate mengalami **stagnasi** atau sedikit penurunan menjadi **36,5%** meski tidak signifikan. Namun pada tahun **2019**, angka cancelation rate booking hotel **melonjak** cukup tajam menjadi **39%**, angka tertinggi dalam **3 tahun terakhir.**

Data tersebut menunjukkan, meski sempat stagnan dua tahun belakangan, tingkat pembatalan reservasi kamar hotel pada akhirnya **meningkat signifikan** di tahun **2019**, bahkan lebih tinggi dari 2017. Hal ini mengindikasikan tamu hotel semakin mudah untuk membatalkan booking kamar meskipun sudah melakukan pemesanan sebelumnya. Kecenderungan pembatalan ini terus meningkat dari tahun ke tahun.

Dampak Terhadap Hotel

Pendapatan hotel berkurang
Hotel kehilangan kesempatan mendapat tamu dan pendapatan karena kamar yang sudah dibooking dibatalkan

Kerusakan Reputasi
Banyaknya pembatalan bisa diartikan tamu hotel tidak puas. Reputasi dan citra hotel bisa terpengaruh di mata calon tamu.

Kesulitan Perencanaan
Hotel sudah mempersiapkan sejumlah kamar, makanan, dan fasilitas lain berdasarkan jumlah booking. Pembatalan menyebabkan perencanaan dan persiapan ini sia-sia.

Biaya tambahan
Hotel harus mengeluarkan usaha lebih besar untuk promosi dan mendapatkan tamu pengganti agar kamar yang dibatalkan tetap terisi.

Objektif

- Mengetahui karakteristik pelanggan yang melakukan pembatalan dan menemukan pola pembatalan pemesanan dengan melakukan eksplorasi data yang mendalam.
- Membangun model *machine learning classifier* untuk memprediksi pembatalan pemesanan hotel.

Data Understanding

```
RangeIndex: 83293 entries, 0 to 83292
Data columns (total 33 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   hotel                           83293 non-null  object
 1   is_canceled                     83293 non-null  int64
 2   lead_time                       83293 non-null  int64
 3   arrival_date_year               83293 non-null  int64
 4   arrival_date_month              83293 non-null  object
 5   arrival_date_week_number        83293 non-null  int64
 6   arrival_date_day_of_month       83293 non-null  int64
 7   stays_in_weekend_nights         83293 non-null  int64
 8   stays_in_week_nights            83293 non-null  int64
 9   adults                          83293 non-null  int64
 10  children                        83290 non-null  float64
 11  babies                          83293 non-null  int64
 12  meal                            83293 non-null  object
 13  country                         82947 non-null  object
 14  market_segment                  83293 non-null  object
 15  distribution_channel            83293 non-null  object
 16  is_repeated_guest               83293 non-null  int64
 17  previous_cancellations          83293 non-null  int64
 18  previous_bookings_not_canceled  83293 non-null  int64
 19  reserved_room_type              83293 non-null  object
 20  assigned_room_type              83293 non-null  object
 21  booking_changes                 83293 non-null  int64
 22  deposit_type                    83293 non-null  object
 23  agent                           71889 non-null  float64
 24  company                         4734 non-null   float64
 25  days_in_waiting_list            83293 non-null  int64
 26  customer_type                   83293 non-null  object
 27  adr                             83293 non-null  float64
 28  required_car_parking_spaces     83293 non-null  int64
 29  total_of_special_requests       83293 non-null  int64
 30  reservation_status              83293 non-null  object
 31  reservation_status_date         83293 non-null  object
 32  bookingID                       83293 non-null  int64
dtypes: float64(4), int64(17), object(12)
```

- Data terdiri dari 32 kolom dan 83.293 baris.
- Dari 32 kolom tersebut, 12 kolom berisi data kategorikal, 21 kolom lainnya berisi data numerik
- Variabel Target adalah is_canceled
- Terdapat 4 kolom yang memiliki missing value

Data Cleaning

Missing value

	fitur	missing_count	percentage
0	company	78559	94.316
1	agent	11404	13.691
2	country	346	0.415
3	children	3	0.004

- Kolom **company** memiliki *missing value* sebesar 94,31%, oleh karena itu kolom tersebut akan dihapus.
- Kolom **agent** akan diisi dengan nilai 0 dengan asumsi bahwa reservasi tidak melalui agen.
- Kolom **country dan children** memiliki persentase *missing value* yang sangat rendah (<1%), oleh karena itu baris-baris *missing value* tersebut akan dihapus.

Removing unreasonable values

	adults	adr
count	82944.000000	82944.000000
mean	1.856337	101.888512
std	0.605626	48.018623
min	0.000000	0.000000
25%	2.000000	70.000000

- NIlai **0** pada kolom **adults** adalah tidak masuk akal, karena minimal harus ada satu orang dewasa yang memesan hotel.
- Pada dataset ini diasumsikan bahwa setiap pemesanan memiliki **biaya kamar** yang harus dibayarkan sehingga nilai **0** pada kolom **adr** dihapus.

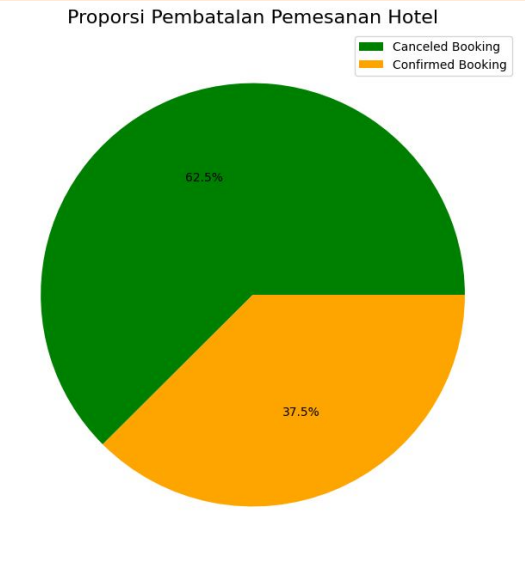
Change the correct Data type

```
children                 float64
reservation_status_date   object
agent                    float64
```

```
children                      int32
reservation_status_date    datetime64[ns]
agent                         int32
```

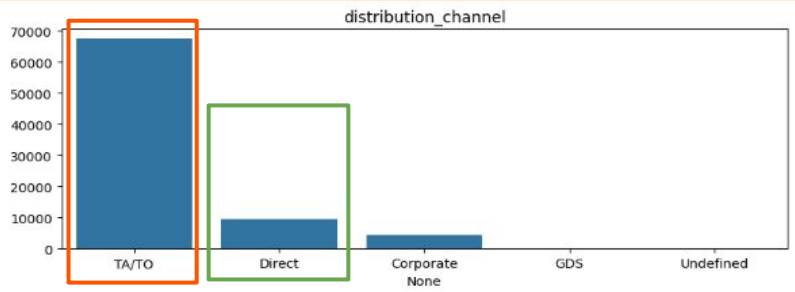
Exploratory Data Analysis

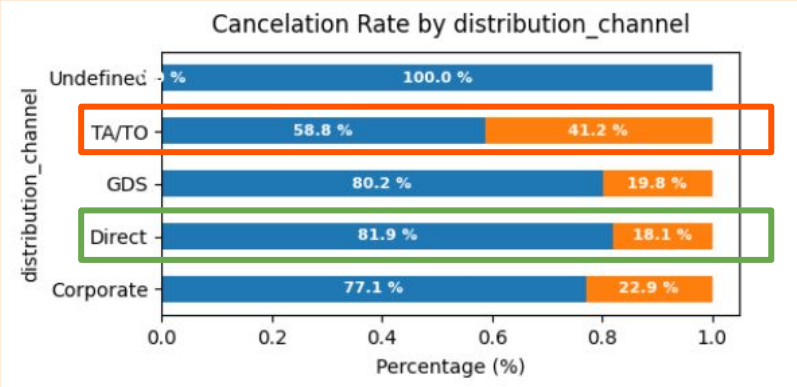
Pelanggan yang membatalkan pemesanan kamar hotel berjumlah 37.5%, sedangkan pelanggan yang tidak membatalkan pemesanan kamar hotel berjumlah 62.5%.



Hal ini dapat menyebabkan **kesulitan** dalam pemodelan, karena model dapat **cenderung** untuk mengklasifikasikan pelanggan sebagai **tidak membatalkan** pemesanan kamar hotel. Oleh karena itu, perlu dilakukan **penyeimbangan kelas** untuk membuat proporsi pelanggan yang membatalkan dan yang tidak membatalkan pemesanan kamar hotel menjadi sama pada saat dilakukan pemodelan.

Distribution Channel & Cancellation Rate



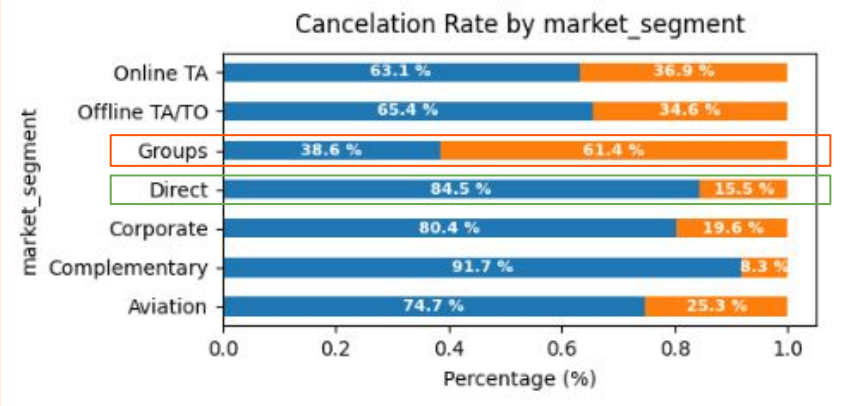


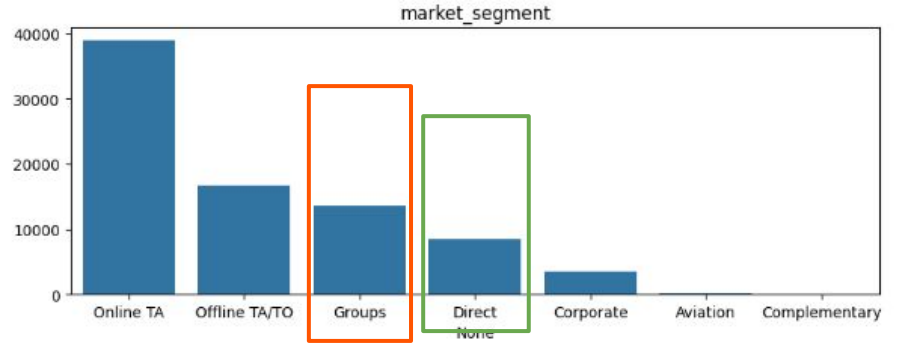
Tingkat pembatalan reservasi hotel melalui distribution **channel TA/TO** mencapai 41.2%, sementara channel ini merupakan pilihan teratas pelanggan. Channel kedua terbanyak adalah **direct** namun tingkat pembatalannya paling rendah yaitu 18.1%

Action:

- Memperkuat **kerja sama** dan **komunikasi** dengan **agen travel/tour** operator untuk meminimalisir pembatalan.
- Mendorong dan **memberi promosi khusus** untuk pemesanan melalui **channel direct** agar tingkat pembatalan rendah dapat dipertahankan atau bahkan ditingkatkan.

Market Segment & Cancellation Rate



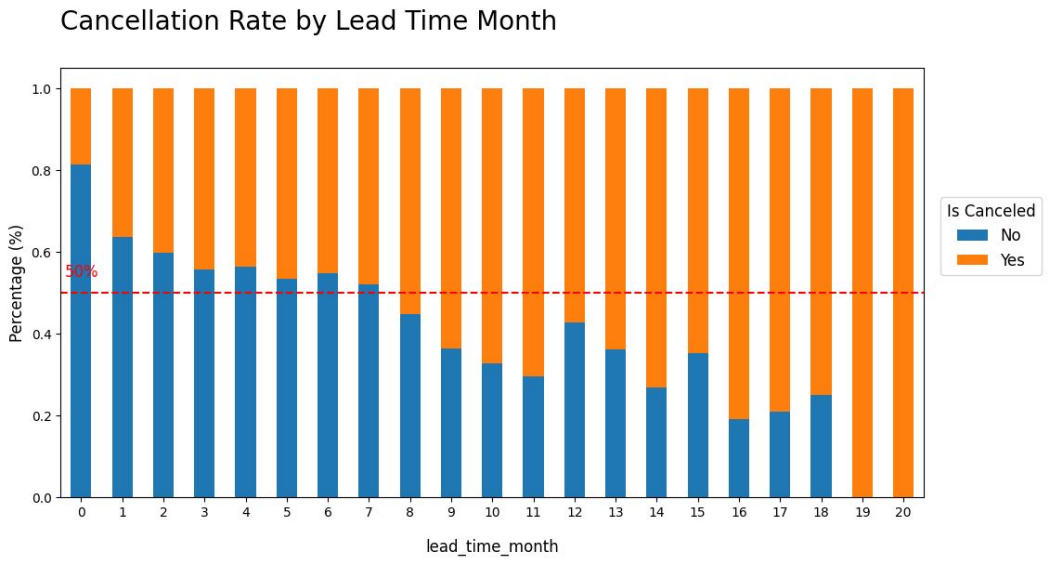


Pada market **segment Group**, terlihat persentase pembatalan pemesanan (cancelation rate) yang tinggi, yaitu sebesar **61%**. Angka ini menunjukkan lebih dari setengah tamu yang awalnya memesan kamar di segmen ini kemudian membatalkan pemesanan mereka.
Selain, itu pada **segment Direct,** persentase pembatalannya paling rendah (selain complementary) yaitu sebesar **15.6%.**

Action:

- Memberlakukan kebijakan yang **lebih ketat** pada pemesanan **segmen Group**.
- Sumber daya pemasaran dan distribusi bisa lebih **difokuskan** pada **segmen Direct** yang berpotensi memberikan pendapatan lebih stabil dengan risiko pembatalan rendah.

Monthly Lead Time & Cancellation Rate



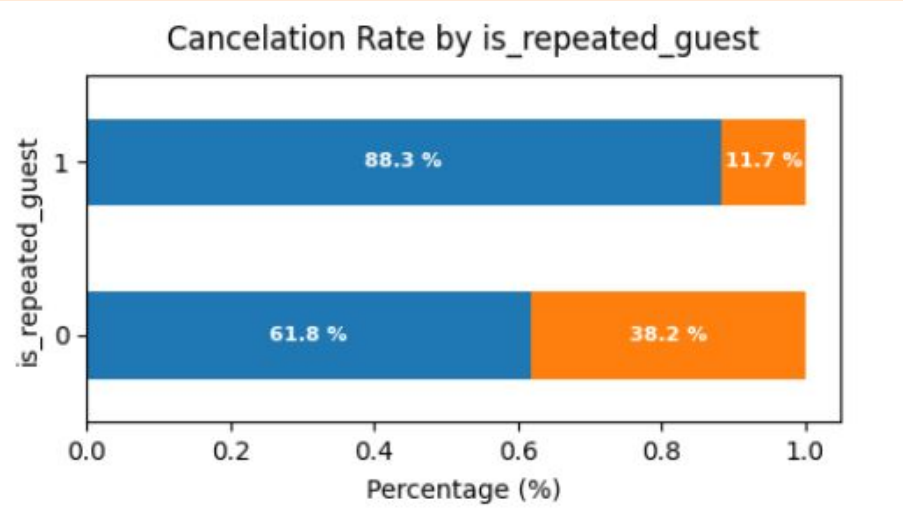
- Pemesanan yang memiliki **lead time ⩽ 7 bulan** memiliki tingkat **konfirmasi** pemesanan yang lebih tinggi (>50%) terhadap tingkat pembatalan.
- Pemesanan yang memiliki **lead time ⩾ 7 bulan** memiliki tingkat **pembatalan** yang lebih tinggi (>50%) dibandingkan dengan tingkat konfirmasinya.

Artinya tingkat pembatalan **berkorelasi positif** terhadap lead time.

Action:

- Menawarkan harga/tarif yang lebih **rendah** untuk pemesanan dengan **lead time pendek** untuk mendorong pemesanan dengan masa tunggu yang singkat.
- Memberikan **batas waktu** pembatalan yang lebih ketat untuk **pemesanan jangka panjang**, misalnya minimal 30 hari sebelum tanggal check-in.

Repeated Guest & Cancellation Rate

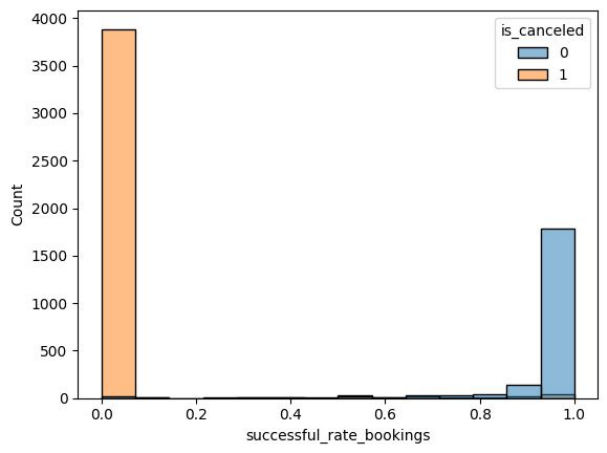


Tingkat pembatalan **berbeda secara signifikan** antara pelanggan yang berulang dan yang baru. Pelanggan yang berulang, meskipun jumlahnya **sedikit**, memiliki tingkat pembatalan yang **sangat rendah**, yaitu **11%**. Sementara itu, pelanggan yang baru, yang merupakan mayoritas, memiliki tingkat pembatalan yang **tinggi**, yaitu **38.2%**. Hal ini menunjukkan bahwa pelanggan yang berulang **lebih loyal** dan **tidak mudah berubah pikiran** dalam pemesanan kamar hotel.

Action:

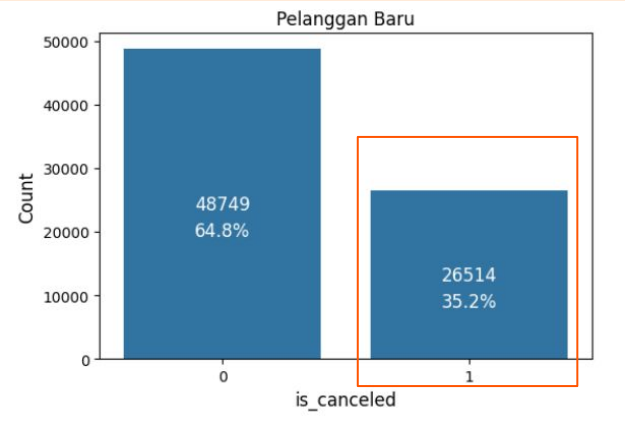
- Mempromosikan penawaran konversi dari tamu baru menjadi member loyalitas untuk menurunkan tingkat pembatalan ke depannya.
- Fokus pada program loyalitas dan retensi pelanggan untuk meningkatkan jumlah pelanggan berulang yang memiliki tingkat pembatalan rendah (11%).

Successful Rate Bookings Historis & Cancellation Rate

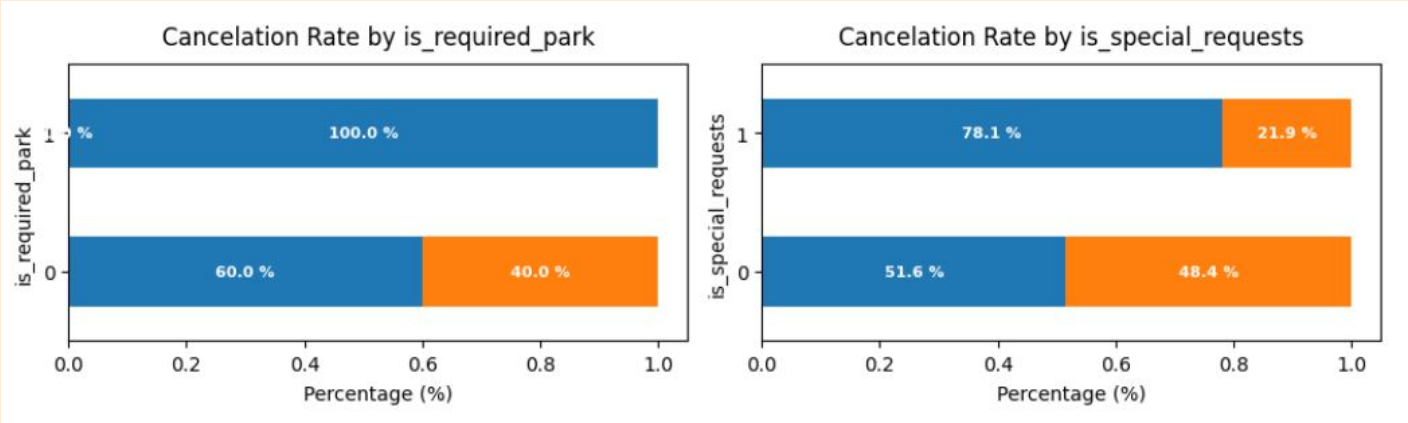


Pelanggan yang memiliki **Successful Rate Bookings Historis** yang **tinggi** cenderung tidak membatalkan pesanan mereka, sedangkan pelanggan yang memiliki **Successful Rate Bookings** cenderung membatalkan pesanan mereka kembali.

Pelanggan yang memiliki yang **tidak memiliki riwayat pemesanan hotel sebelumnya** (pelanggan baru) memiliki memiliki tingkat pembatalan yang cukup tinggi, yaitu **35.2%.**



Parking Space and Special Request & Cancellation Rate



- Pelanggan yang meminta **fasilitas parkir** memiliki tingkat pembatalan sangat rendah (0%), sedangkan yang tidak meminta parkir tingkat pembatalannya tinggi (40%).
- Pelanggan dengan **permintaan khusus** memiliki tingkat pembatalan rendah (21,9%), sementara tanpa permintaan khusus tingkat pembatalannya tinggi (48,4%).

Insight ini menunjukkan bahwa pemenuhan fasilitas seperti parkir dan akomodasi permintaan khusus tamu berkorelasi positif terhadap konfirmasi pemesanan dengan rendahnya tingkat pembatalan pemesanan hotel mereka. Artinya, preferensi dan kepuasan tamu terhadap penyediaan fasilitas/layanan sesuai kebutuhannya dapat meningkatkan komitmen untuk tidak membatalkan pemesanan.

Action:

- Mengoptimalkan **penyediaan dan promosi** fasilitas parkir kendaraan kepada calon tamu, terutama untuk area dengan banyak pengunjung menggunakan kendaraan pribadi.
- **Meningkatkan komunikasi** saat pemesanan untuk menggali dan mengakomodasi permintaan khusus dari calon tamu, seperti fasilitas disabilitas, pilihan makanan, dll.

Rekomendasi EDA

Peningkatan Monitoring

Hotel disarankan untuk meningkatkan pemantauan terhadap channel pemesanan dan segmen pasar yang memiliki tingkat pembatalan tinggi. Dengan pemantauan yang lebih cermat, hotel dapat mengidentifikasi pola-pola pembatalan dan mengambil langkah-langkah yang sesuai untuk mengurangi tingkat pembatalan.

Penyesuaian Strategi Harga

Berdasarkan analisis terhadap faktor-faktor yang mempengaruhi pembatalan, hotel dapat menyesuaikan strategi harga kamar secara dinamis. Misalnya, dengan menawarkan harga kamar yang lebih rendah untuk pemesanan dengan lead time yang lebih lama atau untuk segmen pasar yang cenderung memiliki tingkat pembatalan tinggi.

Penawaran Khusus untuk Pelanggan Berulang

Hotel dapat memberikan penawaran khusus atau insentif kepada pelanggan yang berulang untuk mendorong mereka tetap setia dan mengurangi kemungkinan pembatalan pemesanan.

Rekomendasi EDA

Peningkatan Komunikasi dengan Tamu

Hotel dapat meningkatkan komunikasi dengan tamu, terutama sebelum tanggal check-in, untuk mengkonfirmasi pemesanan dan mengurangi kemungkinan pembatalan dekat waktu check-in.

Evaluasi Permintaan Khusus

Hotel dapat mengevaluasi permintaan khusus dari tamu dan mempertimbangkan untuk menyediakan fasilitas atau layanan tambahan yang dapat meningkatkan kepuasan tamu dan mengurangi kemungkinan pembatalan.

Modeling
Machine Learning

Pre-Processing

- **Encoding data categorical**
- **Feature Selection dengan chi-square contingency (data kategorik) dan ANOVA satu arah (data numerik)**
  Setelah dilakukan seleksi fitur, dari total 76 fitur setelah dilakukan proses encoding, hanya 62 fitur yang dipertahankan untuk digunakan.
- **Membagi dataset menjadi data train dan data test**
- **Melakukan feature scaling untuk fitur numerical**
- **Melakukan undersampling pada data train untuk menyeimbangkan kelas yang tidak seimbang**

Model Used

1. Logistic Regression
2. K-Neighbors Classifier
3. Decision Tree Classifier
4. Random Forest Classifier
5. Gradient Boosting Classifier
6. LightGBM Classifier
7. XGBoost Classifier

Model Result

Berdasarkan performa **Recall** model yang terbaik adalah model **XGBoost** yang telah dilakukan **hyperparameter tuning** dengan score **92.43%**. Dengan kata lain, jika ada **100 orang** yang berpotensi membatalkan pemesanan hotel, model ini mampu memprediksi sekitar **92 orang**.

Selain nilai recall yang tinggi, nilai **f1-score** dari model tersebut juga tinggi, yaitu **80.70%** yang menandakan model tidak hanya dapat memprediksi sebagian besar pelanggan yang berpotensi **cancel booking** kamar hotel dengan akurat, tetapi juga dapat **menghindari** kesalahan dalam memprediksi pelanggan yang **confirmed booking** kamar hotel sebagai **cancel booking**.

	Model	AUC Score	F1-Score	Recall	Time
0	XGBoost_Tuned	0.9456	0.8070	0.9243	44.3125
1	LightGBM_Tuned	0.9483	0.8288	0.8674	14.5938
2	XGBoost	0.9416	0.8168	0.8653	12.5781
3	XGBoost_Scalling	0.9416	0.8168	0.8653	13.3438
4	LightGBM_Scalling	0.9384	0.8120	0.8585	2.4844
5	LightGBM	0.9387	0.8117	0.8583	2.2500
6	Random Forest_Tuned	0.9340	0.8001	0.8580	24.2656
7	Random Forest_Scalling	0.9445	0.8240	0.8527	4.2344
8	Random Forest	0.9445	0.8242	0.8522	4.3281
9	Gradient Boosting_Scalling	0.9248	0.7914	0.8363	4.8438
10	Gradient Boosting	0.9248	0.7914	0.8363	4.8281
11	Decision Tree_Scalling	0.8162	0.7655	0.8257	0.2500
12	Decision Tree	0.8159	0.7651	0.8254	0.2656
13	Logistic Regression_Tuned	0.9004	0.7608	0.7917	1.9688
14	KNN Neighbors_Scalling	0.8677	0.7342	0.7911	14.9375
15	Logistic Regression_Scalling	0.8645	0.7278	0.7514	0.7969
16	KNN Neighbors	0.8168	0.6730	0.7410	15.6875
17	Logistic Regression	0.8485	0.7126	0.7397	0.6562

Strategi Penggunaan Model

Strategi 1: Pengelolaan Overbooking Berbasis Probabilitas Prediksi

Hasil prediksi model, terutama **probabilitas pembatalan**, dapat menjadi panduan bagi hotel untuk merancang strategi mitigasi seperti overbooking, dengan memperhatikan kemungkinan kesalahan prediksi.

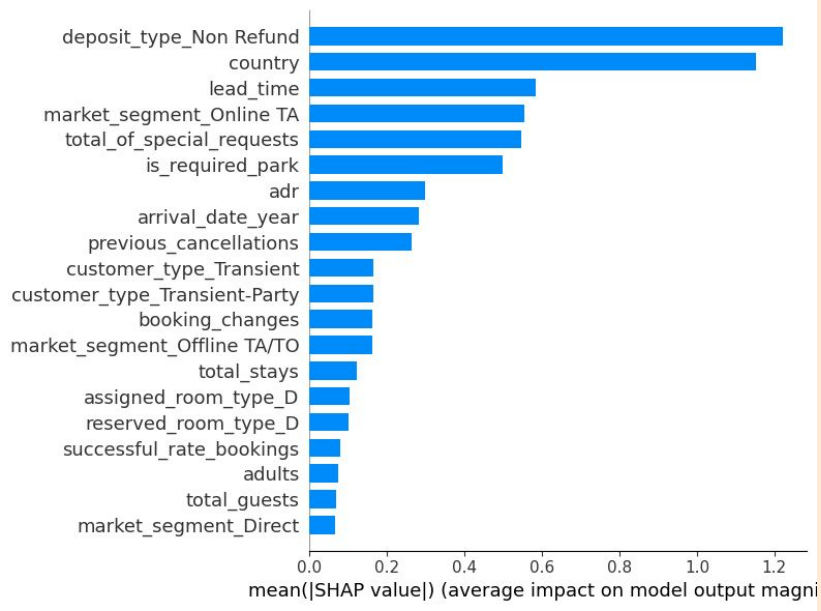
Misalnya:

- Probabilitas < 50%: Monitor status pemesanan dengan cermat.
- 50% ≤ Probabilitas ≤ 75%: Lakukan overbooking secara ringan.
- Probabilitas > 75%: Lakukan overbooking dengan tegas dan berikan insentif. Insentif dapat berupa diskon atau voucher menarik kepada tamu lain yang bersedia melakukan pemesanan (overbooking), dengan risiko harus pindah kamar jika tamu dengan probabilitas pembatalan tinggi ternyata tidak membatalkan.

Strategi 2: Kebijakan Pembatalan Dinamis

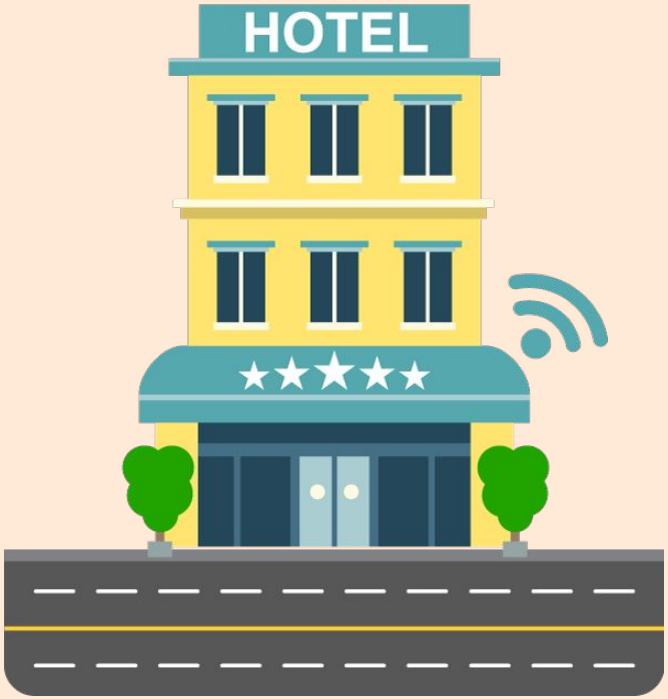
Penerapan **kebijakan pembatalan yang dinamis**, berdasarkan prediksi probabilitas pembatalan, memungkinkan hotel untuk menyesuaikan batas waktu pembatalan sesuai dengan risiko pembatalan yang diprediksi. Misalnya, untuk tamu dengan probabilitas pembatalan tinggi, hotel dapat menerapkan batas waktu pembatalan yang lebih ketat. Sebagai contoh, hotel dapat meminta pembatalan dilakukan 72 jam sebelum tanggal check-in, daripada kebijakan standar yang meminta pembatalan dilakukan 24 jam sebelumnya. Pendekatan ini membantu hotel memiliki lebih banyak waktu untuk menyesuaikan inventaris kamar dan meminimalkan dampak pembatalan terhadap pendapatan.

Feature Important



Berdasarkan hasil feature importance yang menunjukkan **deposit_type_Non Refund** dan **country** menjadi dua faktor teratas yang paling berpengaruh terhadap pembatalan pemesanan hotel, berikut beberapa strategi khusus yang dapat dilakukan:

- Mengkaji ulang kebijakan deposit non-refundable yang ternyata berkorelasi tinggi dengan pembatalan. Pertimbangkan untuk membatasi atau memodifikasi kebijakan ini.
- Menyediakan layanan pelanggan dalam bahasa asli pelanggan juga bisa membantu. Ini dapat meningkatkan kepuasan pelanggan dan mengurangi kemungkinan pembatalan.



Thanks!

Github:
https://github.com/hubble99

Email:
ramadhoninasri09@gmail.com

Linkedin:
https://www.linkedin.com/in/ramadhoni-nasri-4b514b220/