Rakamin Academy

Laporan Project

Stage 2

1. Apakah sudah melakukan pengecekkan data bermasalah seperti missing values, invalid values, atau data duplicate dan sudah membersihkannya?

Pengecekkan untuk data bermasalah seperti missing values, duplicate data, outliers, feature transformation, feature encoding, serta class imbalance sudah dilakukan.
Berikut hasil dari treatment tersebut:

A. Missing values

- Ditemukan 4 kolom yang memiliki missing values, yaitu : company, agen, country, children.
- Beberapa treatment terkait temuan tersebut, yaitu :

1. Kolom children karena hanya memiliki 4 baris data yang kosong maka missing value tersebut akan di hapus
2. Kolom agent karena merupakan ID maka nilai akan di set 0 untuk data yang tidak memiliki ID dan 1 untuk data yang memiliki ID.
3. Kolom company karena banyak missing value (94%) maka feature company dapat di drop
4. Kolom country karena sebagian besar nilainya adalah PRT dan missing value rationya kecil, maka missing value tersebut akan diisi dengan nilai PRT.

B. Duplicate data

Setelah dilakukan pengecekkan menggunakan fungsi duplicated, tidak ditemukan adanya indikasi data ganda atau duplicate.

C. Outliers

- **Mengubah outlier menjadi batas atas (high limit)**
  Tujuannya adalah membatasi value feature agar persebarannya tidak terlalu luas. Hal tersebut digunakan untuk feature : lead_time dan adr
- **Ganti Nilai Outlier Menjadi Nilai Terdekat**
  Tujuannya adalah untuk memberikan variasi data dikarenakan data hanya bertumpuk pada 1 value. Hal tersebut digunakan untuk feature : stays_in_weekend_nights, adults, children, babies, dan total_special_request.
- **Ubah Nilai Outlier Menjadi 2 Variasi**
  Karena persebaran hanya tertumpuk pada 1 nilai, dan secara logika feature tersebut dapat disederhanakan menjadi 2 nilai. Handling tersebut dilakukan pada feature : previous_cancellations, previous_bookings_not_cancelled, booking_changes, days_in_waiting_list, dan required_car_parking_spaces.

D. Feature transformation

Melakukan normalized / re-scale kolom numerical yang memiliki angka variance tinggi guna mempermudah interpretasi beberapa model Machine Learning.

E. Feature encoding

Dilakukan treatment dengan Label Encoding untuk fitur hotel (karena ada 2 unique), reserved_room_type dan assigned_room_type (karena unique data nya ordinal). Selain itu juga menggunakan One Hot Encoding untuk fitur distribution_channel, deposit_type, customer_type, reservation_status dimana fitur tersebut memiliki nilai unique lebih dari 2 dan tidak bersifat ordinal.

F. Handle Class Imbalance

Melakukan beberapa step untuk tahapan ini, yaitu :

1. Memisahkan kolom y untuk target, x untuk kolom selain target
2. Melakukan split data menjadi train dan test
3. RUS dan ROS dengan package Imblearn

2. Apakah sudah menentukan feature apa saja yang akan digunakan, atau perlu ditambahkan, dan reformatting feature sesuai dengan kebutuhan?

Penentuan feature apa saja yang akan digunakan, atau perlu ditambahkan, dan reformatting feature sesuai dengan kebutuhan sudah ditentukan. Berikut rinciannya :

A. Feature selection

- **Remove irrelevant features**
  Kolom name, email, phone-number, credit_card adalah kolom yang tidak memiliki pola, sehingga kolom tersebut dapat dihapus.
- **Remove redundant column**
  Kolom stays_in_weekend_nights dan stays_in_week_nights dihapus karena sudah ada kolom total_stays, kemudian Kolom children dan babies dihapus karena sudah ada kolom kids.

B. Feature extraction

1. **Total stays,** dari kolom stays_in_weekend_nights dan stays_in_week_nights, kita bisa mendapatkan total jumlah hari menginap. Terlihat data total stays bernilai 0, hal ini memungkinkan sebab konsumen dapat check in dan checkout dihari yang sama.
2. **Total guest**, dari kolom `adults, children and babies` kita bisa mendapatkan total jumlah orang yang menginap.
3. **Kids,** Dari kolom children and babies kita bisa mendapatkan total jumlah anak yang menginap. Karena biasanya hanya ada kategori tamu dewasa dan anak (dibawah 17 tahun) dalam pemesanan kamar hotel.

4. **Guest locations,** karena kolom country memiliki data yang sangat besar, maka akan dikategorikan menjadi local & international. Pengkategorian mengacu pada lokasi hotel pada dataset yang berada di negara Portugal untuk penduduk local dan selain itu akan di kategorikan sebagai international.
5. **Arrival columns,** dari kolom ini akan digenerate kolom baru berdasarkan bulan dan hari.
6. **Meal columns,** dari kolom ini didapat data bernilai 'undentified'. Dari source dataset ini menyebutkan bahwa undentified sama dengan SC (no meal package), untuk itu kami akan mengganti nilai undentified menjadi SC.
7. **Distribution channel**, dari kolom ini didapat data bernilai Undefined dan akan di ganti menjadi TA/TO yang merupakan nilai modus.
8. **Market segment**, dari kolom ini didapat data bernilai Undefined dan akan di ganti menjadi online TA yang merupakan modus.

C. Feature tambahan

1. **Customer Satisfaction Rate (Rating)**: Fitur ini perlu ditambahkan karena dapat digunakan untuk menilai kinerja pelayanan hotel dan mungkin saja berpengaruh terhadap cancellation rate. Hipotesis dari kami adalah semakin tinggi rating yang diberikan kepada hotel, maka akan semakin rendah cancellation rate.
2. **Total Revenue:** Pada Dataset hanya disajikan data mengenai Average Daily Rate (ADR), yang didapatkan dari pendapatan kamar/kamar terisi. Fitur ini belum belum menggambarkan pengeluaran tiap customer per order. Total Revenue dapat digunakan untuk menganalisis efektifivitas pricing yang telah ditetapkan dan mungkin mungkin saja berpengaruh terhadap cancellation rate.
3. **Reason For Staying:** Alasan seseorang melakukan pemesanan hotel diantaranya adalah Business Trip atau Holiday. Dengan mengetahui alasan tersebut, kedepannya dapat dibuat Marketing Campaign dengan bekerja sama dengan sebuah perusahaan atau tempat wisata di daerah setempat. Selain itu, kita juga akan mengetahui hubungan antara Reason For Staying dan Cancellation rate di hotel tersebut.

4. **Reason For Cancelled:** Alasan seseorang melakukan pembatalan reservation hotel diantaranya adalah Change Mind, Find Better Hotel, Weather Issue, dan lain sebagainya. Faktor alasan pembatalan yang dikarenakan alasan pribadi customer dapat dimaklumi, namun jika alasan pembatalan karena find better hotel, diperlukan evaluasi kinerja ataupun fasilitas dari hotel tersebut dibandingkan dengan hotel competitor sekitar.
5. **Promotion Apply:** Dengan adanya fitur ini, dapat diketahui efektifitas promosi yang sedang dijalankan, sekaligus mengetahui apakah customer yang order dengan menggunakan promo, kemungkinan cancel nya akan lebih tinggi atau rendah.

Feature yang akan kami gunakan yaitu :

1.Berdasarkan correlation heatmap pada multivariate analysis:

- total_of_special_request
- required_car_parking_spaces
- booking_change
- previous_bookings_not_canceled
- is_repeated guest
- agent
- previous_cancellation
- lead_time_norm
- adr
- total_stay
- total_guest

2.Berdasarkan hasil analisis kami pada stage sebelumnya:

- hotel
- lead_time
- arrival_date_year
- arrival_date_month
- is_repeated_guest
- previous_cancellations
- previous_bookings_not_canceled
- booking_change
- adr
- assigned_room_type
- deposit_type
- days_in_waiting_list
- customer_type
- total_stays

Dari kedua ini akan dipilih salah satu yang menghasilkan akurasi tertinggi.