Data Science Project

ANALYSIS AND PREDICTION OF HOTEL RESERVATION CANCELLATION

Presented by : Aziz Nur Rohman
Data Science Batch 22

ibimbing

BUSINESS UNDERSTANDING

Background

In recent years there has been a rapid increase in hotel cancellations. More and more frequently, guests tend to cancel their room reservation. Cancellations are an important aspect of hotel operations, impacting revenue streams, guest satisfaction and operational efficiency.

Problem

Cancellations can have a bad effect on the hotels. A loss of income occurs as a result of unsold rooms. By identifying which bookings are likely to be cancelled, revenue managers and other members of the hotel's staff can take measures to avoid potential cancellations.

BUSINESS OBJECTIVE

1. Analayze Trend and Seasonality related to booking cancellation
2. Discover about the highest reasons of cancellation
3. Build machine learning model to predict cancellation

DATA UNDERSTANDING

Source:

From the link given by dibimbing.id:
Link:
https://bit.ly/DatasetUserInterviewDS

Data Description:

The columns

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

The columns:

- 33 columns
- 83,293 rows
- 4 missing value columns:
    a. children (0.003%)
    b. country (0.42%)
    c. agent (13.69%)
    d. company (94%)

DATA PREPARATION

Data Cleaning

Missing value handling:

Null value in:

- children replaced by 0
- country replaced by unknown
- agent replaced by 0
- company replaced by 0

Removing some rows:

- adults = 0
- adults + children + babies = 0
- stays_in_week_nights + stays_in_week_nights = 0
- adr = 0

Duplicated rows:

No duplicated values

Result:

There is no drop column. The data is missed 1.88% from original dataset.

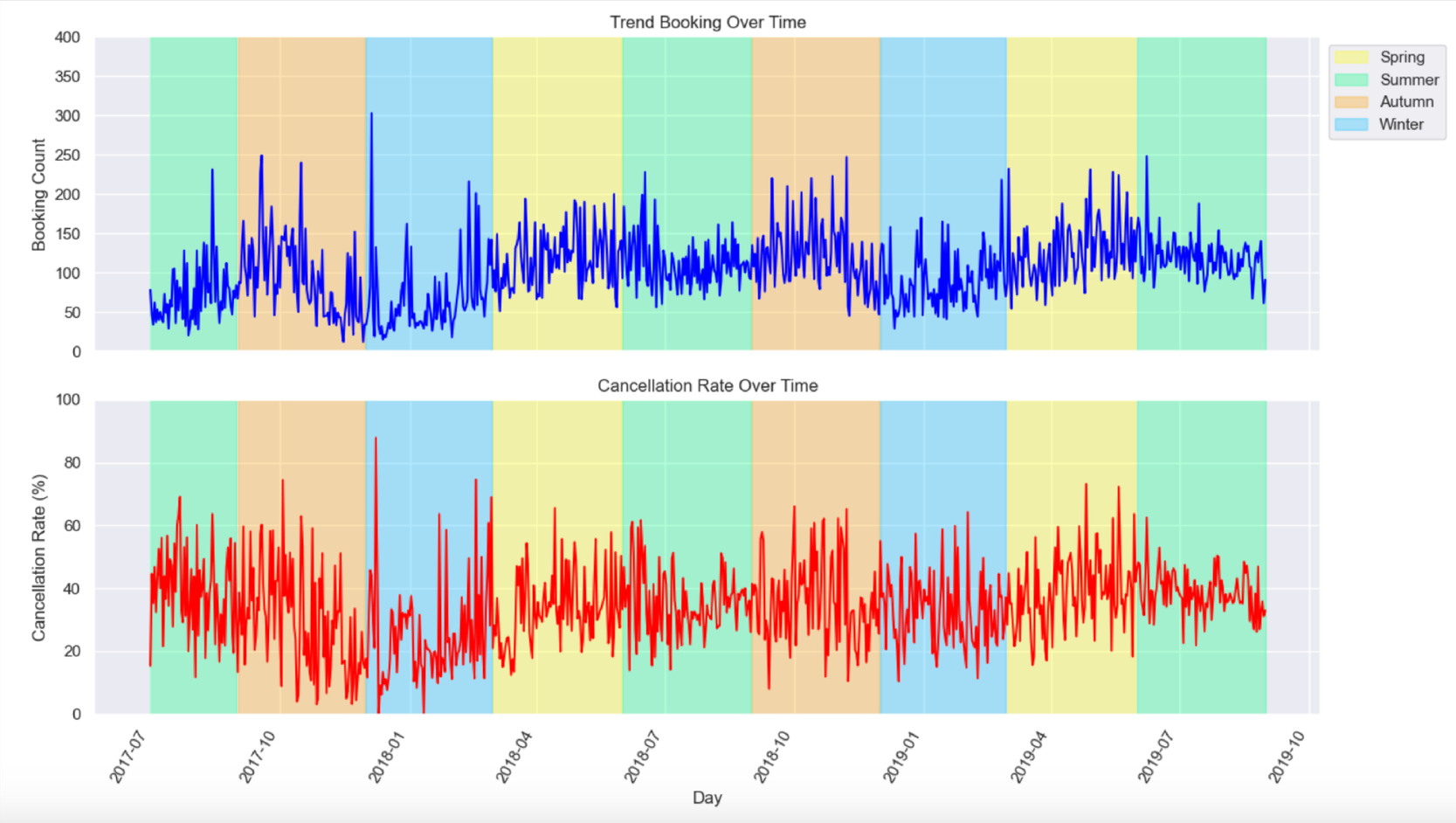
Now the data is:

- 33 columns
- 81724 rows

Exploratory Data Analysis (EDA)

TREND

Total Booking and Cancellation Rate

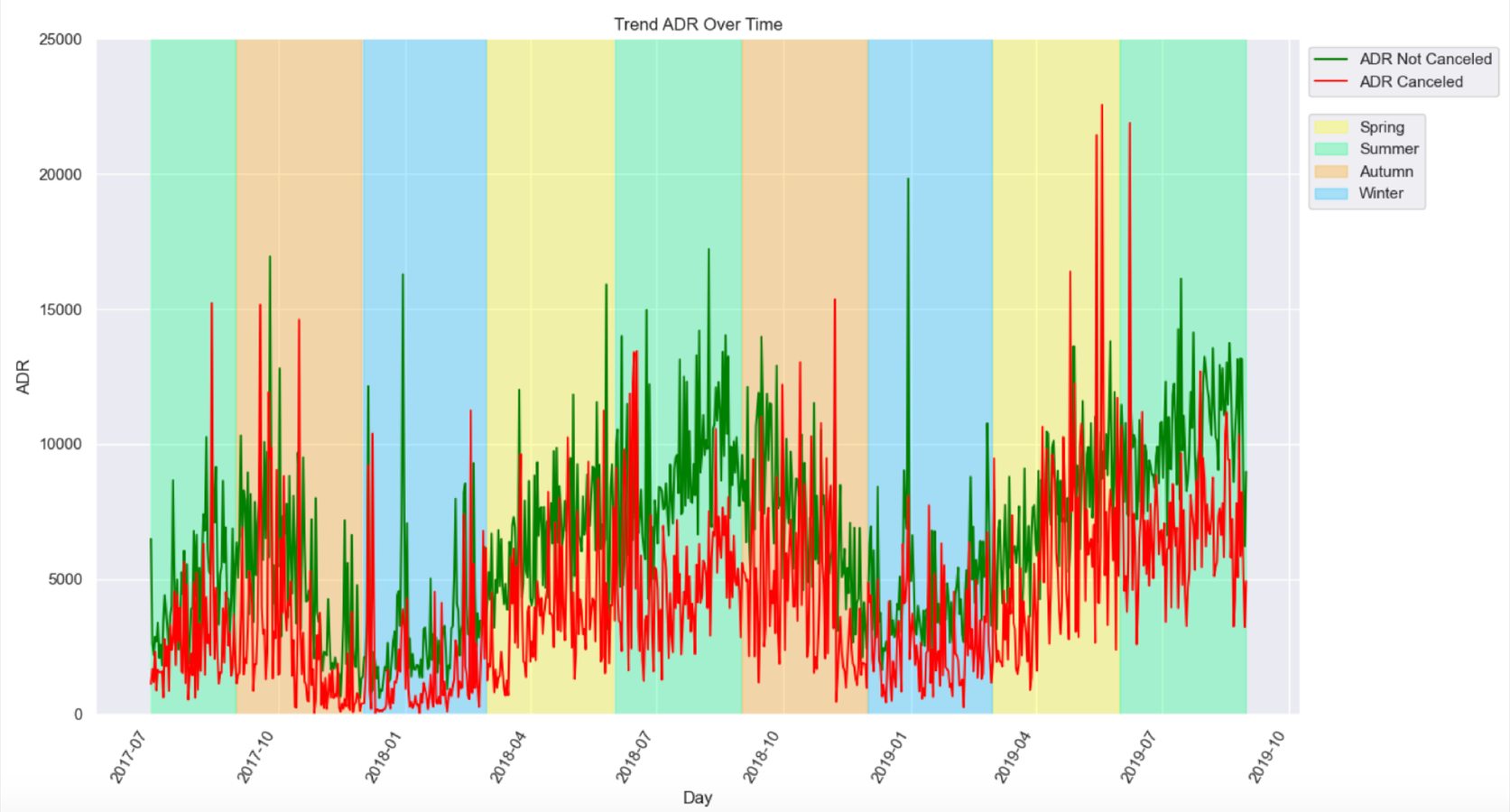


Insight

1. Both total booking and cancellation rate is fairly fluctuative over time
2. Total booking has always been in the lowest count in winter (Dec-Feb) and the highest count in spring (Mar-Apr) and autumn (Sep-Nov)
3. Cancellation rate is more stable than total booking. It is only in Dec 17 - Feb 18 (winter) cancellation rate is fairly sagging.
4. Total booking and cancellation rate is fairly influenced by month and year

TREND

ADR Not Canceled and Canceled

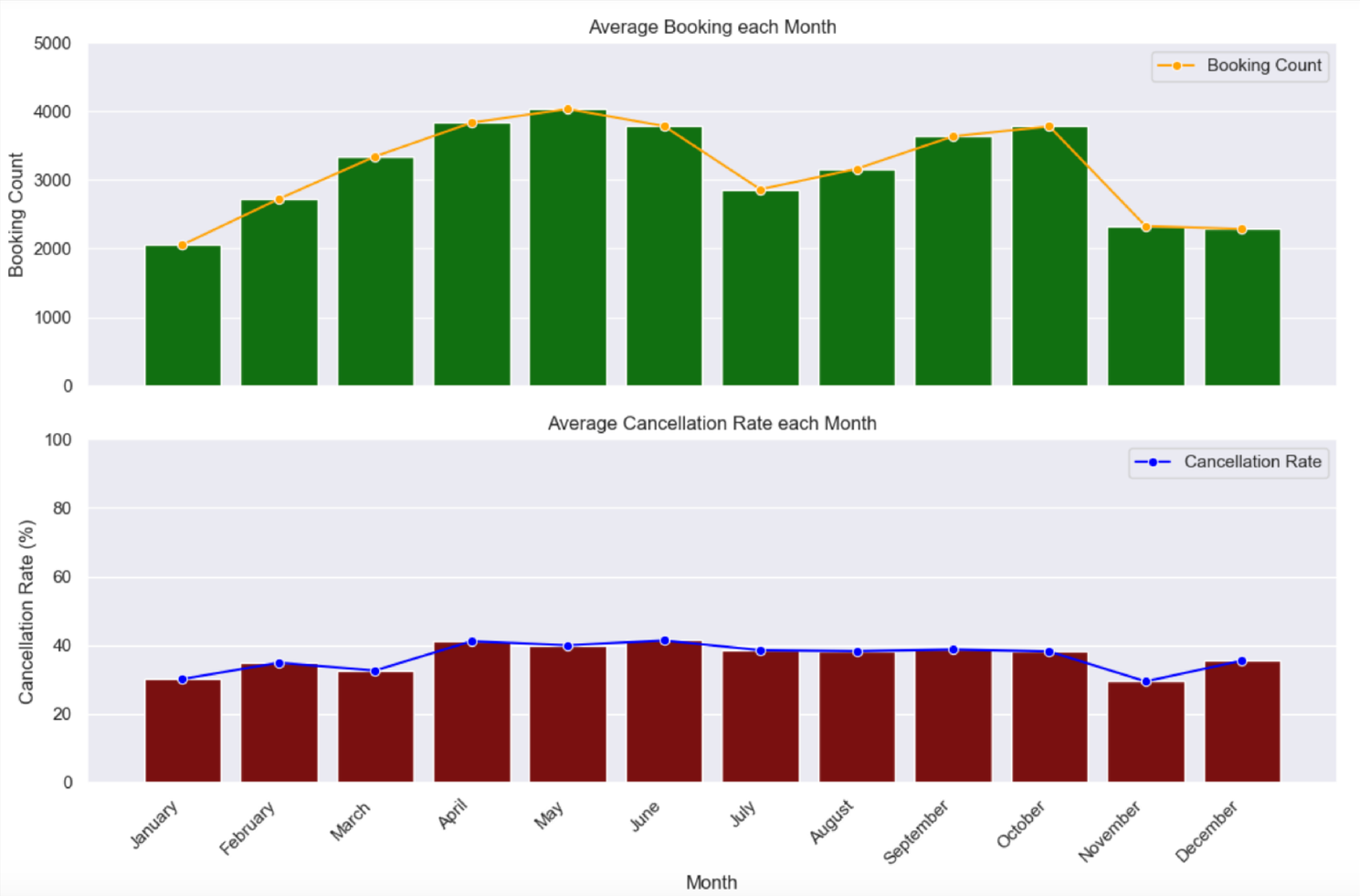


Insight

1. In general, Those ADR are quite erratic over time
2. Both ADR are always decreasing in the winter (Dec-Feb) and are always uplifitng in summer-autumn (Jun - Nov)
3. Both ADR are influenced by month and year.

SEASONALITY

Total Booking and Cancellation Rate by month

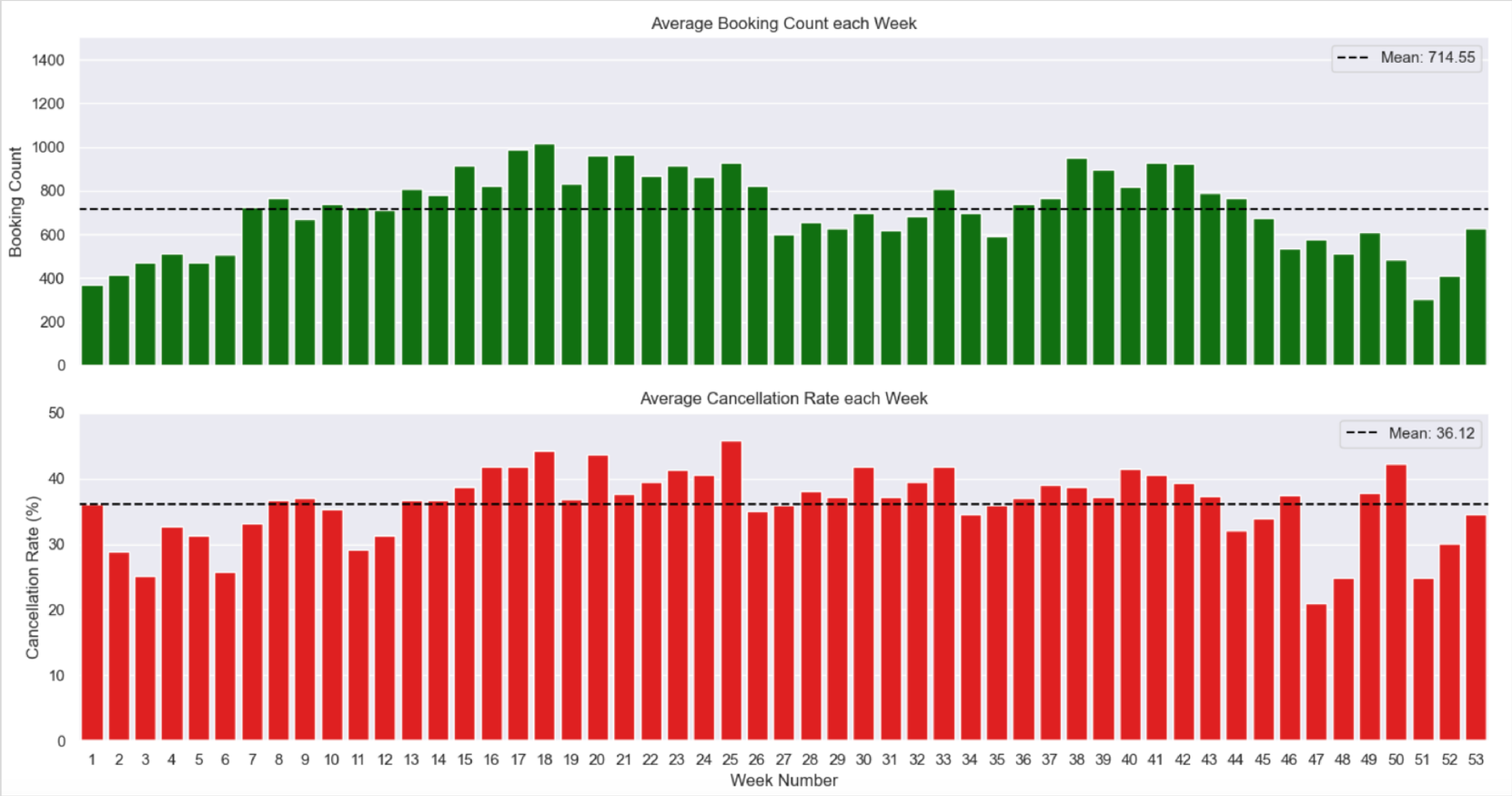


Insight

1. Generally, Both graph has their own unique pattern
2. Total book has two peak points. It is in Apr-Jul and Oct-Nov. Those month is the busiest.
3. The highest cancelation rate happens in April-October Which has more less 40%.
4. From those pattern, the leverage of month is fairly high for total booking and for cancellation rate

SEASONALITY

Total Booking and Cancellation Rate by Week



Insight

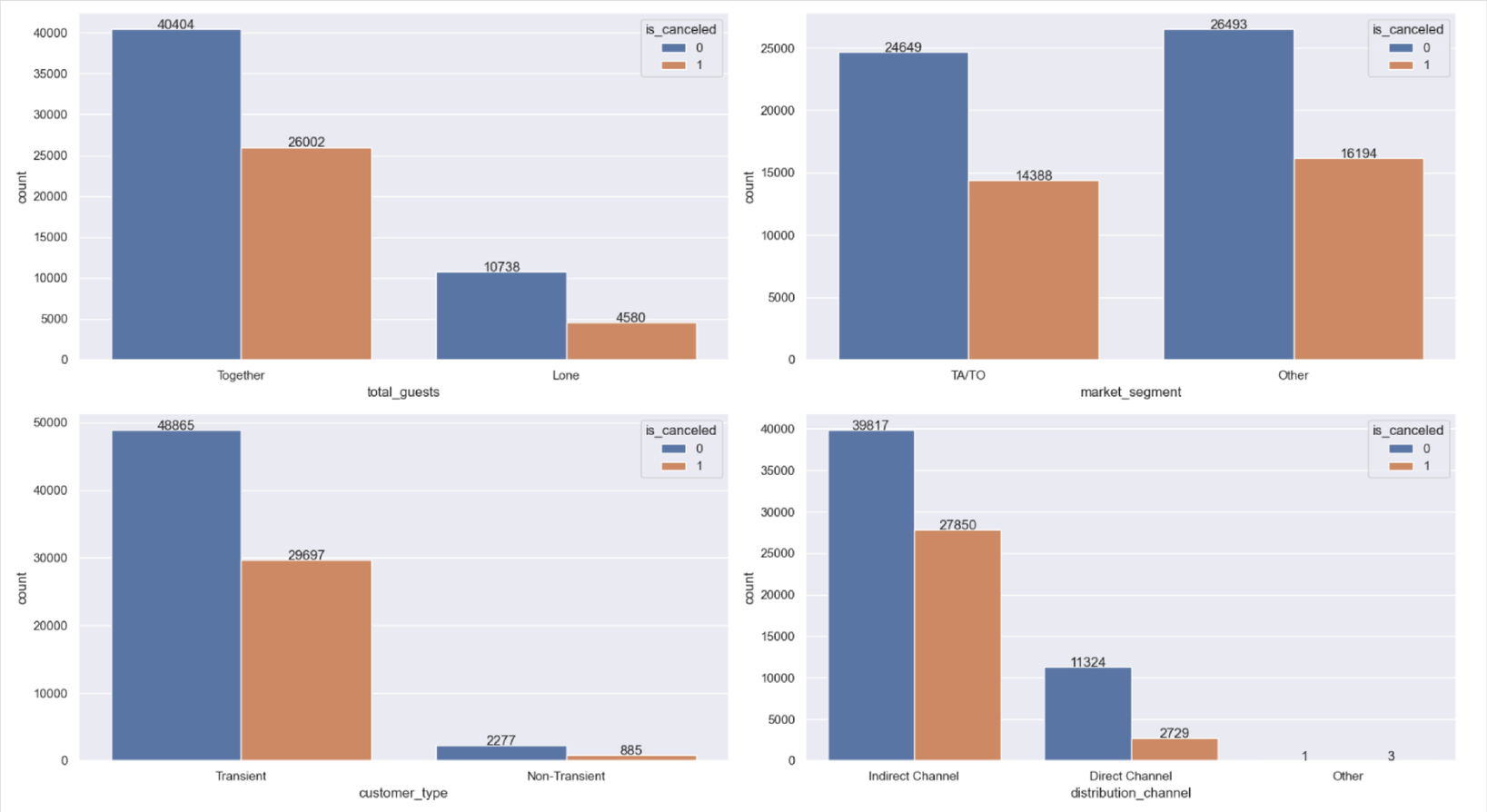
1. Total booking's pattern is like wave. it has peak season in 13th -26th week and 36th-44th week. It means those week sare the busiest.
2. Most cancellation rate in 13th-43th is above average. These weeks are the highest cancellation rate
3. From those pattern, the change of week could affect the number of booking. While, cancellation rate is slightly affected.

DEEP-DIVE ANALYSIS

To make it easier, In this case, I grouped the features based on the same character:

1. **Demographics** (adults, children, babies, country, market_segment, customer_type, distribution_channel)
2. **Historical Booking** (is_repeated_guest, previous_cancellation, previous_bookings_not_canceled)
3. **Meal and Room Type** (meal, reserved_room_type, assigned_room_type)
4. **Booking Process** (booking_changes, deposit_type, agent, company, days_in_waiting_list)
5. **Special Requests** (required_car_parking_spaces, total_of_special_requests)
6. **Duration of Stay** (stays_in_week_nights, stays_in_weekend_nights)
7. **Lead Time** (lead_time)
8. **ADR** (adr)

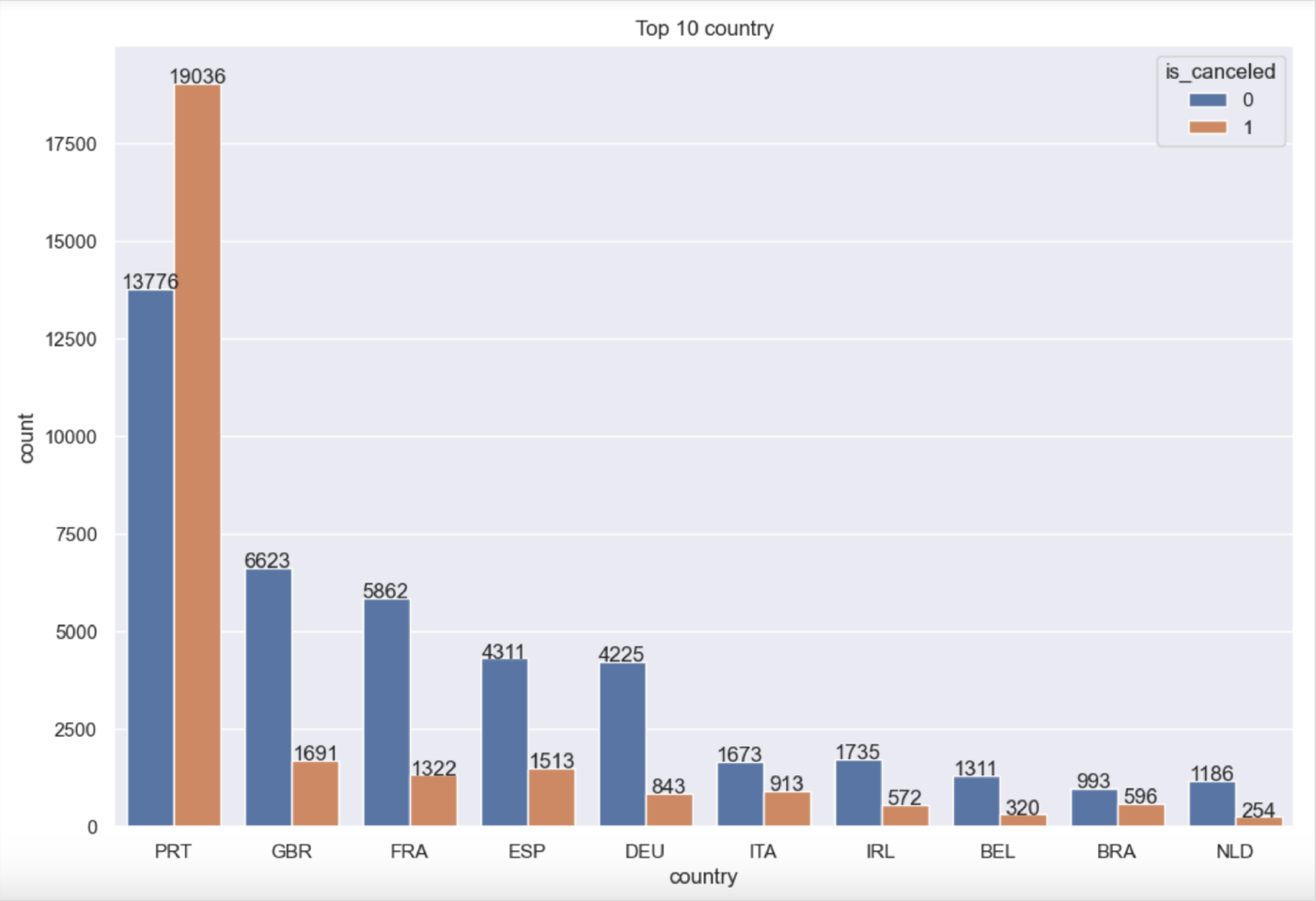
DEMOGRAPHICS ANALYSIS



Insight

1. Total guests more than 1 (Together) is slightly higher to cancel reservation.
2. each category of market_segement almost has the same potential to cancel reservation
3. Transient customer type is more likely to cancel booking.
4. Indirect channel is the highest probable to make cancellation.

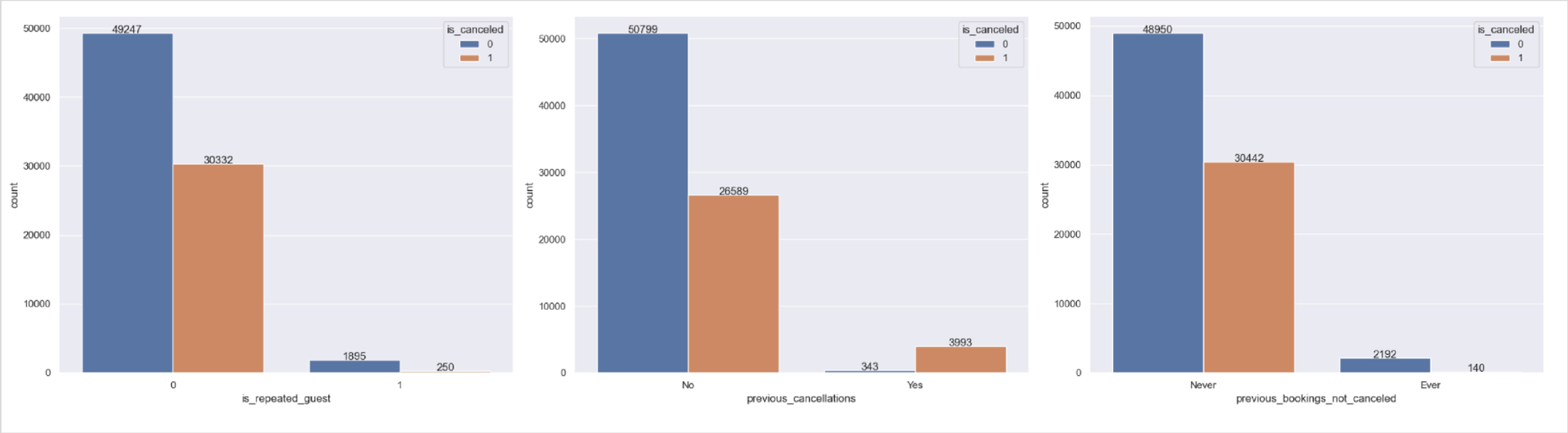
DEMOGRAPHICS ANALYSIS



Insight

1. Portugal is the highest total booking and cancellation

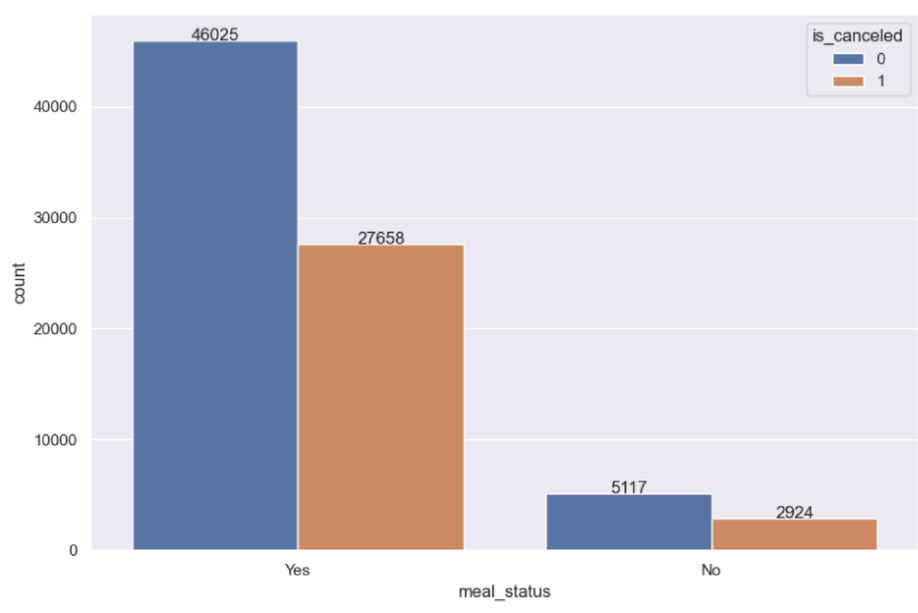
HISTORICAL BOOKING ANALYSIS

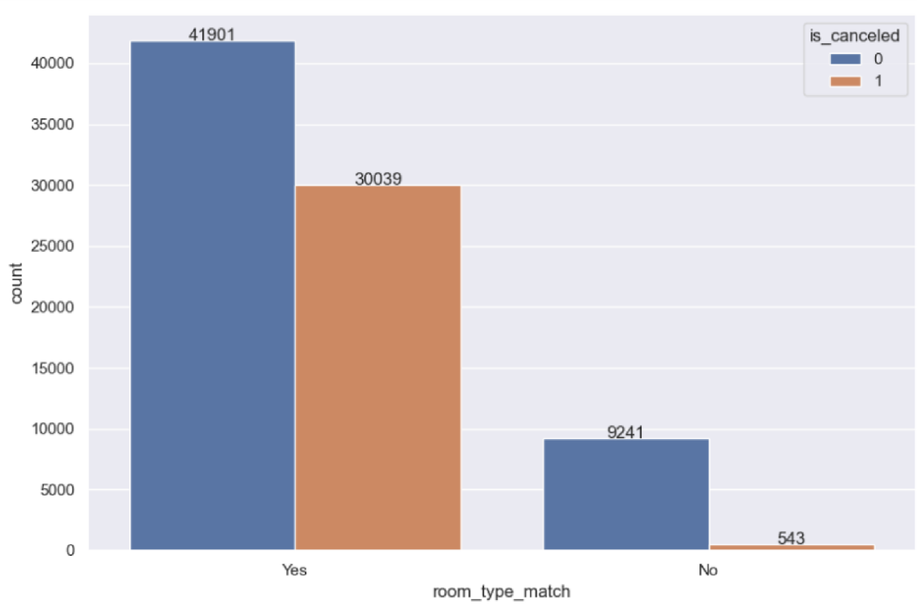


Insight

1. Most cancellation is conducted by newcomers
2. People with previous cancellation background is highly to cancel again in the future
3. People with never conducting booking is tend to make cancellation

MEAL AND ROOM TYPE ANALYSIS

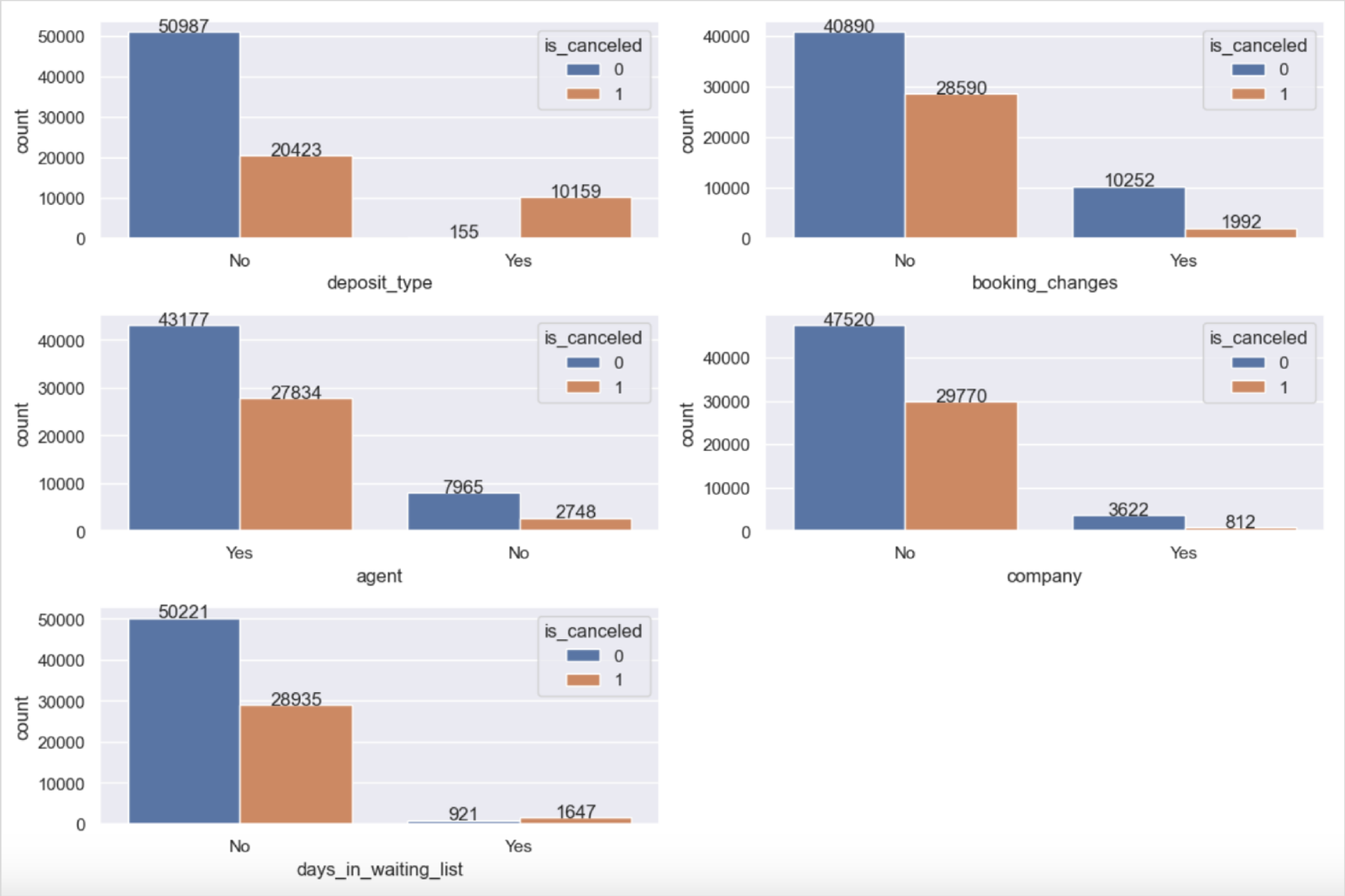




Insight

1. Both reservation with meal and no meal are the same potential to make cancellation
2. Surprisingly, although the reserved room type is the same as the assigned romm type, it does not make people to not cancel the booking. While people with different room type from the reservation is extremely low to cancel.

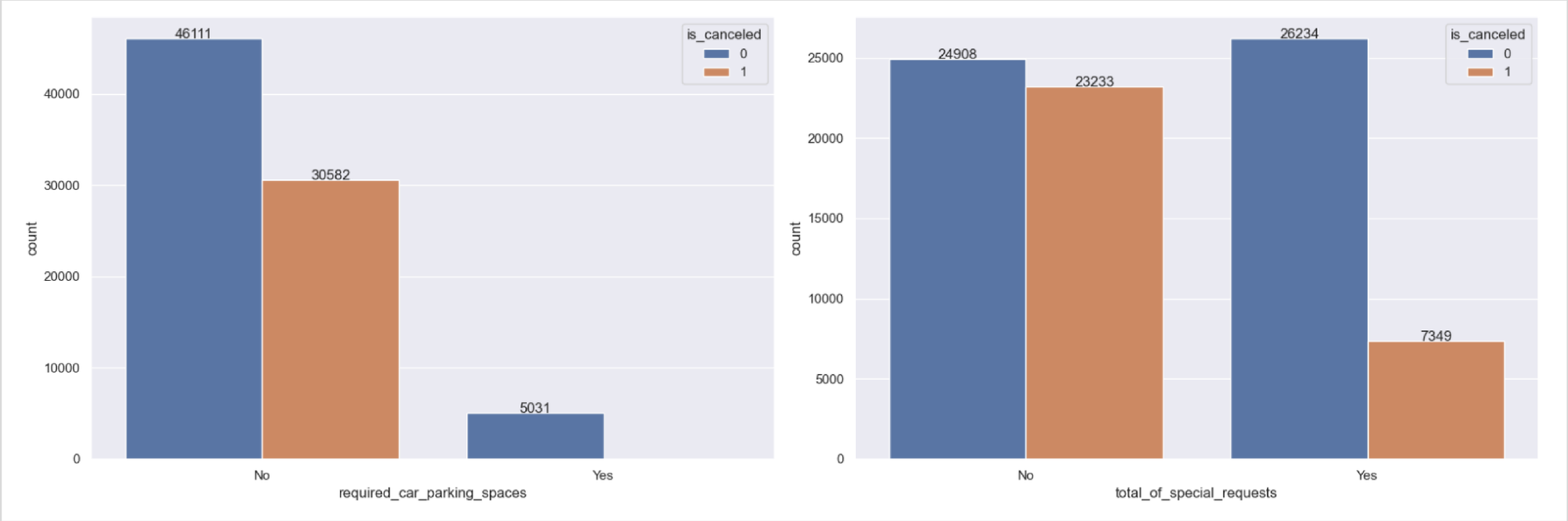
BOOKING PROCESS ANALYSIS



Insight

1. Payment with deposit is the highest reason for cancellation. Almost all the people cancel reservation.
2. No booking change is tend to cancel.
3. Booking via agent is more likely to make cancellation
4. Reservation which is conducted independently is higher to cancel
5. Booking process with waiting day is highly potential to make cancellation.

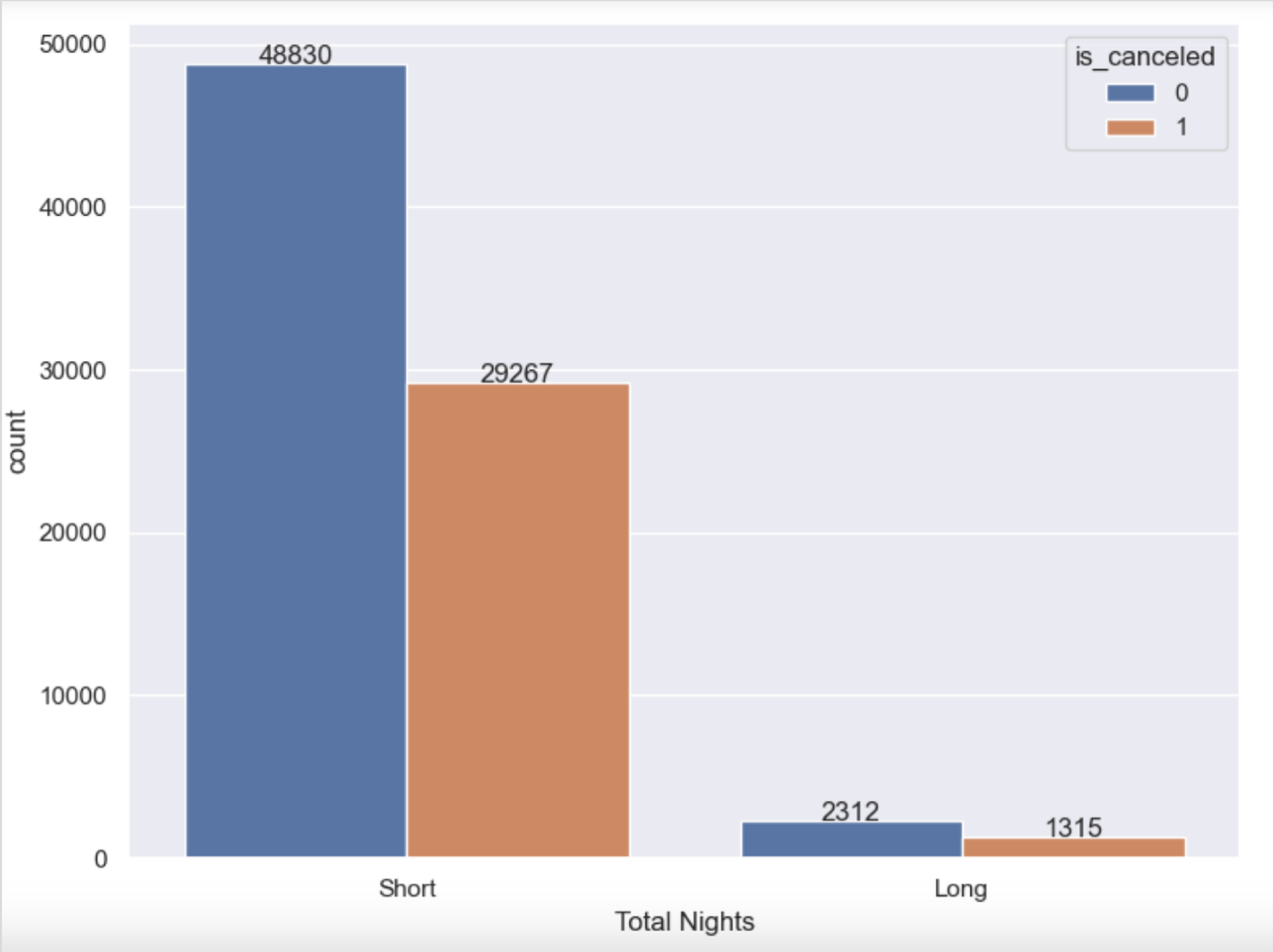
SPECIAL REQUEST INFORMATION ANALYSIS



Insight

1. All people with car parking space request do not make cancellation. While people without the request mostly cancel reservation.
2. No special request is highly to make cancellation.

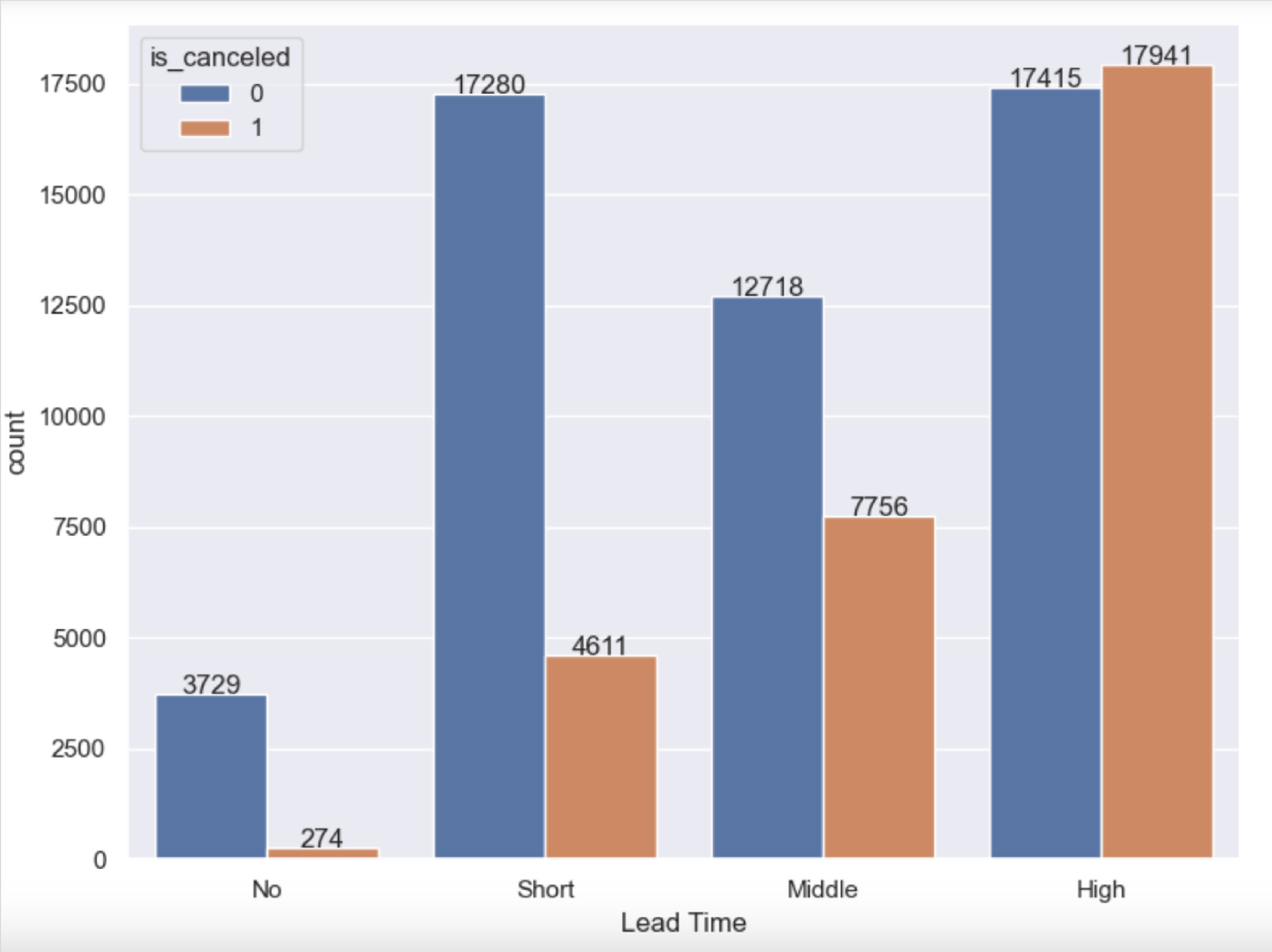
DURATION OF STAY ANALYSIS



Insight

1. Both Short and Long stay in hotel have the same potential to cancel reservation

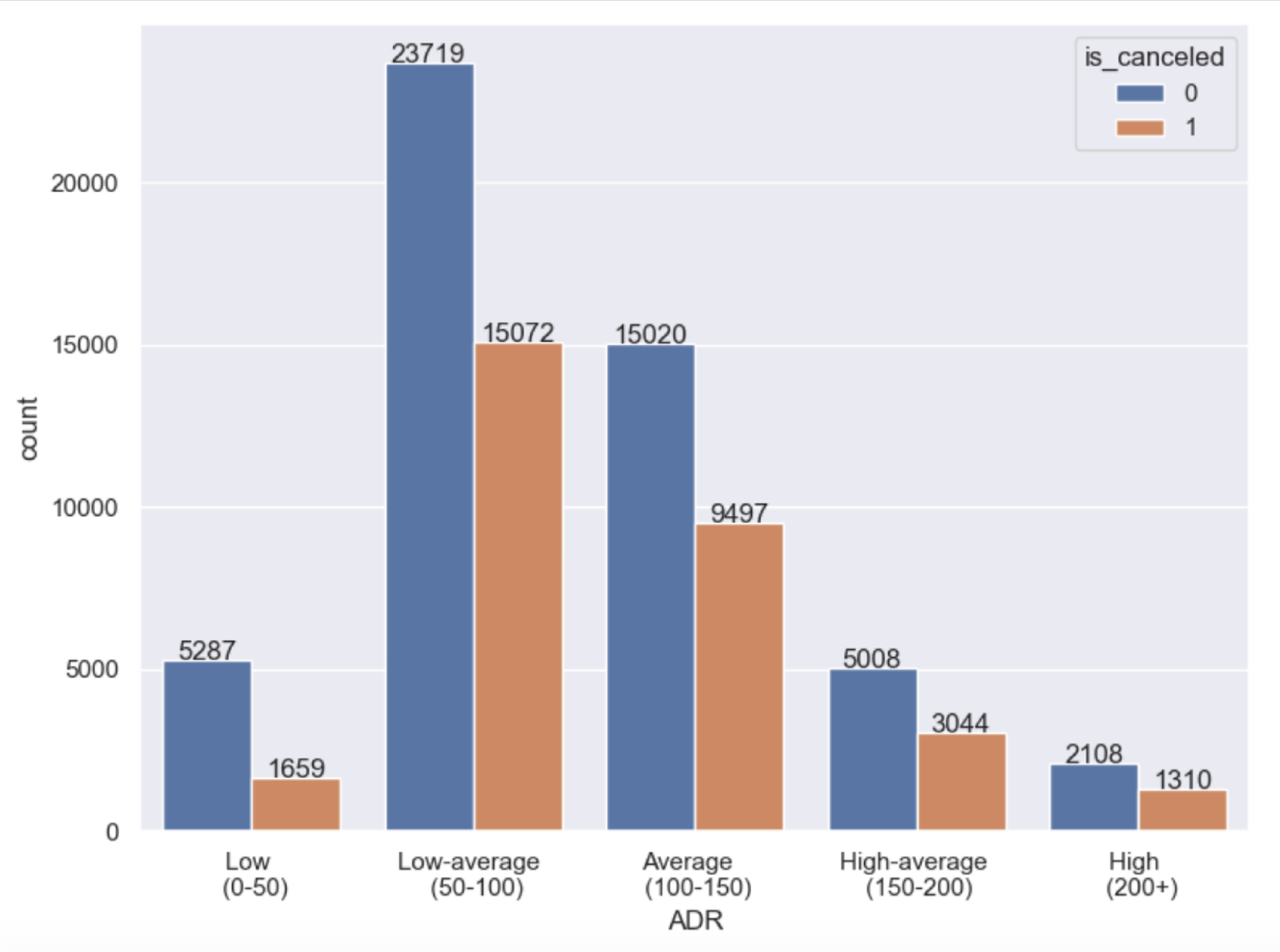
LEAD TIME ANALYSIS



Insight

1. The higher lead time, the higher cancellation

ADR ANALYSIS



Insight

1. Low ADR is the lowest cancellation. While Low-average to High has the same cancellation potential

Data Preprocessing

DATA PREPROCESSING

Flow of this step

1. t-test and chi-square test

To Select feature which has high relationship with target column.
t-test -> for numerical features
chi-square test -> for categorical

2. Feature Engineering

a. adults + children + babies -> total_guests

b. stay_in_week_nights + stay_in_weekend_nights -> total_nights

c. reserved_room_type and assigned_room_type -> room_type_match

d. {total_guests, market_segment, customer_type, previous_cancellations, previous_bookings_not_canceled, deposit_type, booking_changes, agent, company, days_in_waiting_list, meal, required_car_parking_spcace, total_of_special_requests, total_nights }. These column is transformed into binary column.

e. distribution_channel -> transformed into three categories

f. lead_time -> transormed into four categories

g. adr -> transformed into five categories

DATA PREPROCESSING

Flow of this step

3. Encoding

- Manually encoding (arrival_date_month, adr, lead_time)
- Label encoding (all binary and tertiary columns)
- Frequency encoding (country)

4. t-test and chi-square test

To recheck which column have high correlation with column target

5. Outlier Checking

DATA PREPROCESSING

Flow of this step

Before

```
Data columns (total 33 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   hotel                           81724 non-null  object
 1   is_canceled                     81724 non-null  int64
 2   lead_time                       81724 non-null  int64
 3   arrival_date_year               81724 non-null  int64
 4   arrival_date_month              81724 non-null  object
 5   arrival_date_week_number        81724 non-null  int64
 6   arrival_date_day_of_month       81724 non-null  int64
 7   stays_in_weekend_nights         81724 non-null  int64
 8   stays_in_week_nights            81724 non-null  int64
 9   adults                          81724 non-null  int64
 10  children                        81724 non-null  int64
 11  babies                          81724 non-null  int64
 12  meal                            81724 non-null  object
 13  country                         81724 non-null  object
 14  market_segment                  81724 non-null  object
 15  distribution_channel            81724 non-null  object
 16  is_repeated_guest               81724 non-null  int64
 17  previous_cancellations          81724 non-null  int64
 18  previous_bookings_not_canceled  81724 non-null  int64
 19  reserved_room_type              81724 non-null  object
 20  assigned_room_type              81724 non-null  object
 21  booking_changes                 81724 non-null  int64
 22  deposit_type                    81724 non-null  object
 23  agent                           81724 non-null  int64
 24  company                         81724 non-null  int64
 25  days_in_waiting_list            81724 non-null  int64
 26  customer_type                   81724 non-null  object
 27  adr                             81724 non-null  float64
 28  required_car_parking_spaces     81724 non-null  int64
 29  total_of_special_requests       81724 non-null  int64
 30  reservation_status              81724 non-null  object
 31  reservation_status_date         81724 non-null  object
 32  bookingID                       81724 non-null  int64
```

After

```
Data columns (total 21 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   is_canceled                     81724 non-null  int64
 1   lead_time                       81724 non-null  int64
 2   arrival_date_month              81724 non-null  int64
 3   meal                            81724 non-null  int64
 4   country                         81724 non-null  float64
 5   market_segment                  81724 non-null  int64
 6   distribution_channel            81724 non-null  int64
 7   is_repeated_guest               81724 non-null  int64
 8   previous_cancellations          81724 non-null  int64
 9   previous_bookings_not_canceled  81724 non-null  int64
 10  booking_changes                 81724 non-null  int64
 11  deposit_type                    81724 non-null  int64
 12  agent                           81724 non-null  int64
 13  company                         81724 non-null  int64
 14  days_in_waiting_list            81724 non-null  int64
 15  customer_type                   81724 non-null  int64
 16  adr                             81724 non-null  int64
 17  required_car_parking_spaces     81724 non-null  int64
 18  total_of_special_requests       81724 non-null  int64
 19  total_guests                    81724 non-null  int64
 20  room_type_match                 81724 non-null  int64
```

Modelling and Evaluation

MODELLING

Metric

- F1 Score
- False Negative Rate

Models to Compare:

- K-NN
- Logistic Regression
- Decision Tree
- Random Forest
- Extra Tree
- Naive Bayes
- Adaboost
- LightGBM
- XGBoost
- CatBoost

Methodology:

a. Target Column : is_canceled
b. Target Column is imbalanced
c. Splitting: 80:20 (training: testing)
d. Feature Scaling: MinMaxScaler
e. Conducting model based on:
   - Baseline (BS)
   - Cross Validation (CV)
   - Baseline + CV
   - Undersampling
   - Oversampling

EVALUATION

The criterion of the best model in this case:

- The model is good fit (not overvitting or Underfitting).
- The highest value of f1 score
- The lowest value of FNR

How to evaluate ?

a. Compare the criterion based on scenario:
   - Baseline + CV
   - Undersampling + CV
   - Oversampling + CV

b. Next, select the first winner in each scenario

c. Then, compare those models and choose one model with best of the best criterion.

NB: Acceptable difference is 10%. More than 10%, Model is overfitting

EVALUATION

a. Baseline + CV

	model	f1_score_BS+CV	f1_score_BS+CV_training	difference_f1_score (%)	fnr_score_BS+CV
0	CatBoost	78.096856	79.974541	2.347853	24.852845
1	XGBoost	77.669243	79.460907	2.254774	25.245258
2	Random Forest	77.664551	87.245542	10.981641	23.986266
3	LightGBM	77.212579	78.100275	1.136610	26.536952
4	Extra Tree	77.068116	86.976931	11.392463	25.817528
5	Decision Tree	76.408835	86.976931	12.150458	25.621321
6	K-NN	75.591216	81.351016	7.080182	26.831262
7	AdaBoost	73.394658	74.064601	0.904540	32.161543
8	Logistic Regression	72.461428	73.114725	0.893523	33.960105
9	Naive Bayes	62.835330	62.717715	0.187532	3.106606

- Random Forest, Decision Tree, and Extra Tree are overfitting. So, they are eliminated.
- The highest f1 score is Catboost followed XGBoost, and LightGBM.
- The lowest FNR is Naive Bayes followed CatBoost, XGBoost.

Although Naive Bayes is the lowest score for FNR, it has the lowest f1 score. Hence, the best model of this scenario is:

CatBoost.

- **F1 Score 78.10%**
- **FNR 24.85%**

EVALUATION

b. Undersampling + CV

	model	f1_score_US	f1_score_US_training	difference_f1_score (%)	fnr_score_US
0	CatBoost	78.991337	84.970720	7.036992	16.514061
1	XGBoost	78.807176	84.889935	7.165467	17.037279
2	LightGBM	78.646934	83.611238	5.937363	17.887508
3	Random Forest	77.509549	91.000141	14.824803	17.053630
4	Extra Tree	77.066791	90.820321	15.143670	19.064748
5	Decision Tree	76.260553	90.820321	16.031399	19.506213
6	AdaBoost	75.580655	80.363858	5.951932	22.318509
7	Logistic Regression	74.994112	80.085760	6.357745	21.909745
8	K-NN	71.508575	80.429453	11.091556	33.191629
9	Naive Bayes	62.741364	73.318195	14.425930	3.041203

- Random Forest, Decision Tree, Extra Tree , and Naive Bayes are overfitting. Therefore, they are eliminated
- The highest f1 score is Catboost followed XGBoost, and LightGBM.
- The lowest FNR is CatBoost followed XGBoost, and LightGBM.

Hence, the best model of this scenario is:

CatBoost.

- **F1 Score 78.99%**
- **FNR 16.51%**

EVALUATION

c. Oversampling + CV

	model	f1_score_OS	f1_score_OS_training	difference_f1_score (%)	fnr_score_OS
0	CatBoost	78.960195	87.184055	9.432757	20.536298
1	XGBoost	78.698795	86.704335	9.233148	19.898627
2	LightGBM	78.362195	85.994488	8.875329	20.830608
3	Random Forest	77.504998	92.218994	15.955494	20.765206
4	Extra Tree	77.159091	92.093550	16.216618	22.285808
5	Decision Tree	76.480541	92.093550	16.953422	22.400262
6	K-NN	75.979669	86.487418	12.149454	24.231524
7	AdaBoost	75.779568	81.069258	6.524903	22.318509
8	Logistic Regression	75.041193	79.996889	6.194861	21.811642
9	Naive Bayes	62.837444	73.633303	14.661653	3.139307

- Random Forest, Decision Tree, Extra Tree , KNN, and Naive Bayes are overfitting. Therefore, they are eliminated
- The highest f1 score is Catboost followed XGBoost, and LightGBM.
- The lowest FNR is CatBoost followed XGBoost, and LightGBM.
- When using oversampling, All the models tend to move to overfitting.

Hence, the best model of this scenario is:

CatBoost.

- **F1 Score 78.96%**
- **FNR 20.54%**

EVALUATION

Baseline + CV:

CatBoost.

- F1 Score 78.10%
- FNR 24.85%

Undersampling + CV:

CatBoost.

- F1 Score 78.99%
- FNR 16.51%

Oversampling + CV:

CatBoost.

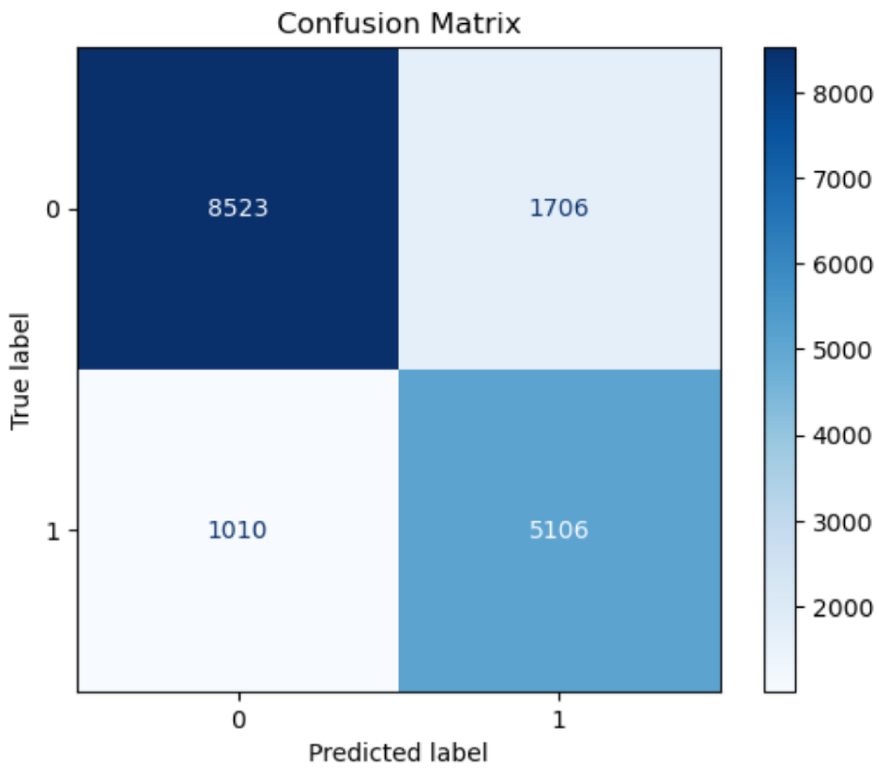
- F1 Score 78.96%
- FNR 20.54%

The best model is CatBoost when it is conducted undersampling. It has the highest f1 score and the lowest FNR.

EVALUATION

CatBoost

```
==========================================================
CatBoost
==========================================================
```



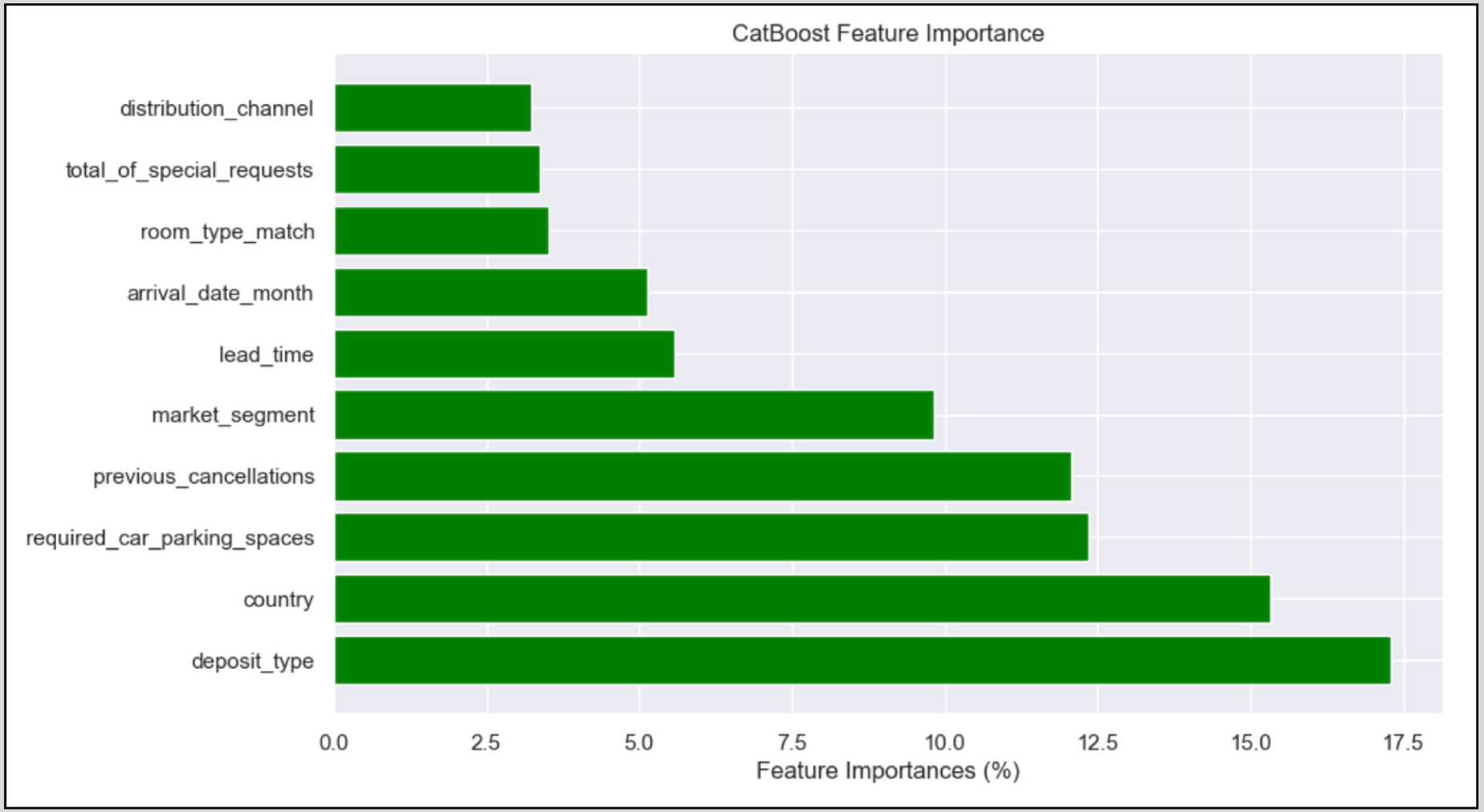
Metric	Value (%)
Precision	74.96
Recall	83.49
F1 Score	78.99
Accuracy	83.38
FNR	16.51

Configuration:

Cross Validation:
Type: Stratified Kfold
n_splits: 10

EVALUATION

CatBoost Feature Importance



IFrom the graph, it can be seen that:
deposit_type is the highest feature that impact to target.

Almost all the features that influence to cancellation are in the top 10 feature importance. They are:

- deposit type
- previous_cancellation
- lead_time
- total_of_special_requests

EVALUATION

Business Impact

Calculation

Total canceled ADR throughout 01-07-2017 to 31-08-2019 (25 months) = $3,216,413.12
ADR/year = (12/25)*$3,216,413.12
ADR/year = $1,543,878.44

canceled ADR mean = $105.17

Catboost:
False Positive = 1706
False Negative = 1010

Assume: ADR loss by model is $105.17/mistake

ADR Loss by model = (1706+1010)*$105.17 = $285,641.72

saving = $1,543,878.44 - $285,641.72 = $1,258,236.72

With Machine Learning Model, The loss can be prevented as $1,258,236.72

Conclusion and Recommendation

CONCLUSION

1. Total booking and cancellation rate is fairly influenced by month and year.
2. ADR are influenced by month and year.
3. The change of week could affect the number of booking. While, cancellation rate is slightly affected.
4. The highest reason for cancellation:
   - Booking with previous cancellation background
   - Payment with deposit
   - Booking process with waiting day
   - Booking without special requests
   - The higher lead time, the higher cancellation
5. The best machine learning model is CatBoost with f1 score 78.99% and FNR 16.51%

RECOMMENDATION

1. Create a new method for payment such as installment payment options. Clearly communicate payment deadlines to guests
2. Adjust pricing strategies based on lead time and offer discounts or special offers for guests who book well in advance.
3. Remind guests by sending notification via email about reservation at certain time. People with high lead time tend to forget about reservation.
4. Create special promotional packages that include additional services or amenities, and offer special benefits to guests who choose those packages.
5. Utilize advanced reservation management systems to monitor and manage bookings more effectively.
6. Provide incentives to guests with a history of cancellations to minimize cancellation risks.

THANK YOU

CONNECT WITH ME

www.linkedin.com/in/aziznurrohmananr

Mail
aziznurrohman.anr@gmail.com