Hotel Booking Cancellation

BlueCode

PT. BlueCode

Member of data scientist team


Agustina Sri Wardani


Fatchul Arifin


Ferry Setefanus


Gigih Septian


Kornelius Rio


M. Harun Arrasyid


Raza Aqil Maulana

Role

Kami bertanggung jawab memberikan rekomendasi agar tingkat pembatalan hotel dapat menurun dan memberikan analisa guna meningkatkan kemajuan hotel

Table of Contents

Business Understanding

EDA

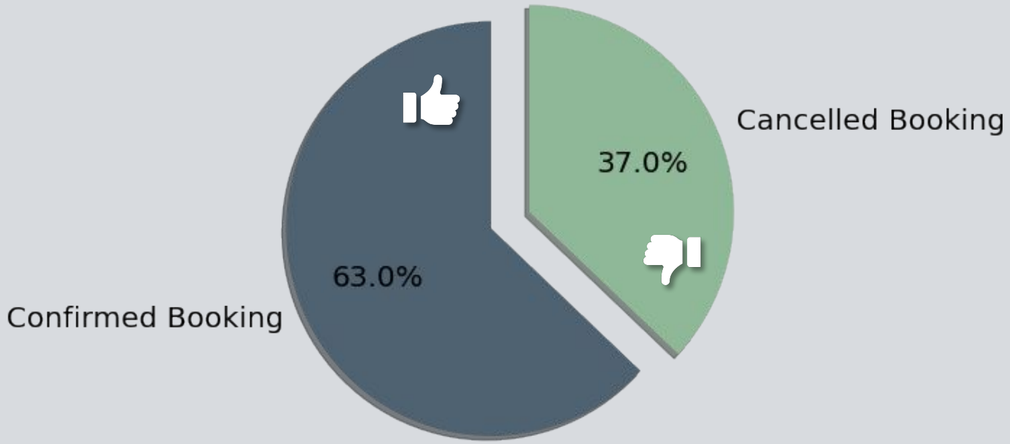
Preprocessing

Modeling & Evaluation

Business Recommendation

Business Understanding

Problem



“booking cancellation merupakan aspek kunci hotel revenue management karena dampaknya pada sistem reservasi kamar” (J. Agustín dkk, 2020)

“rigid cancellation policies yang ditetapkan hotel dapat menimbulkan dampak negatif terhadap revenue dan reputasi hotel” (Antonio dkk, 2021)

Banyaknya pembatalan pesanan hotel menjadi masalah yang menjadi perhatian tim management karena berhubungan langsung dengan revenue. Cancellation policies yang tepat diperlukan agar customer tidak lari ke competitor

Cancellation Rate Prediction solusi tepat untuk mengurangi cancellation dengan strategi bisnis yang tepat

Goal, Objective, Business Metric

GOAL

Menurunkan *Cancellation Rate* Hotel

OBJECTIVE

Model machine learning untuk memprediksi apakah customer akan melakukan cancel atau tidak, sehingga pihak hotel dapat melakukan pendekatan dengan customer yang diprediksi akan cancel

BUSINESS METRIC

Cancellation Rate

Exploratory Data Analysis

Hotel Booking Cancellation Dataset

Dataset terdiri dari 119390 rows & 36 colums

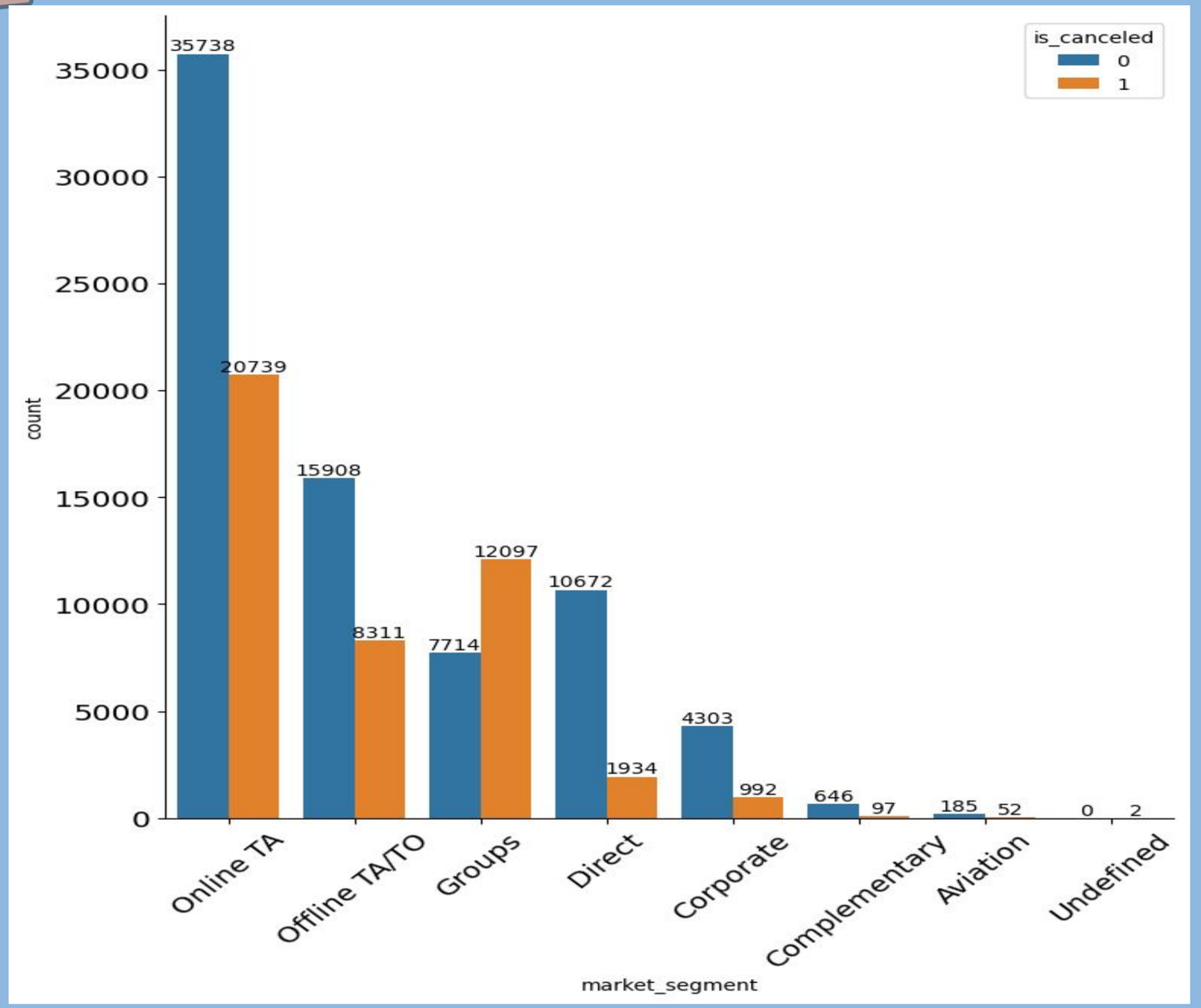
Target: is_canceled

fitur: 35 column selain target

see the Data

Data pada dataset ini didapat dari PMS
(Property Management System) 2 hotel bintang 4 di Portugal

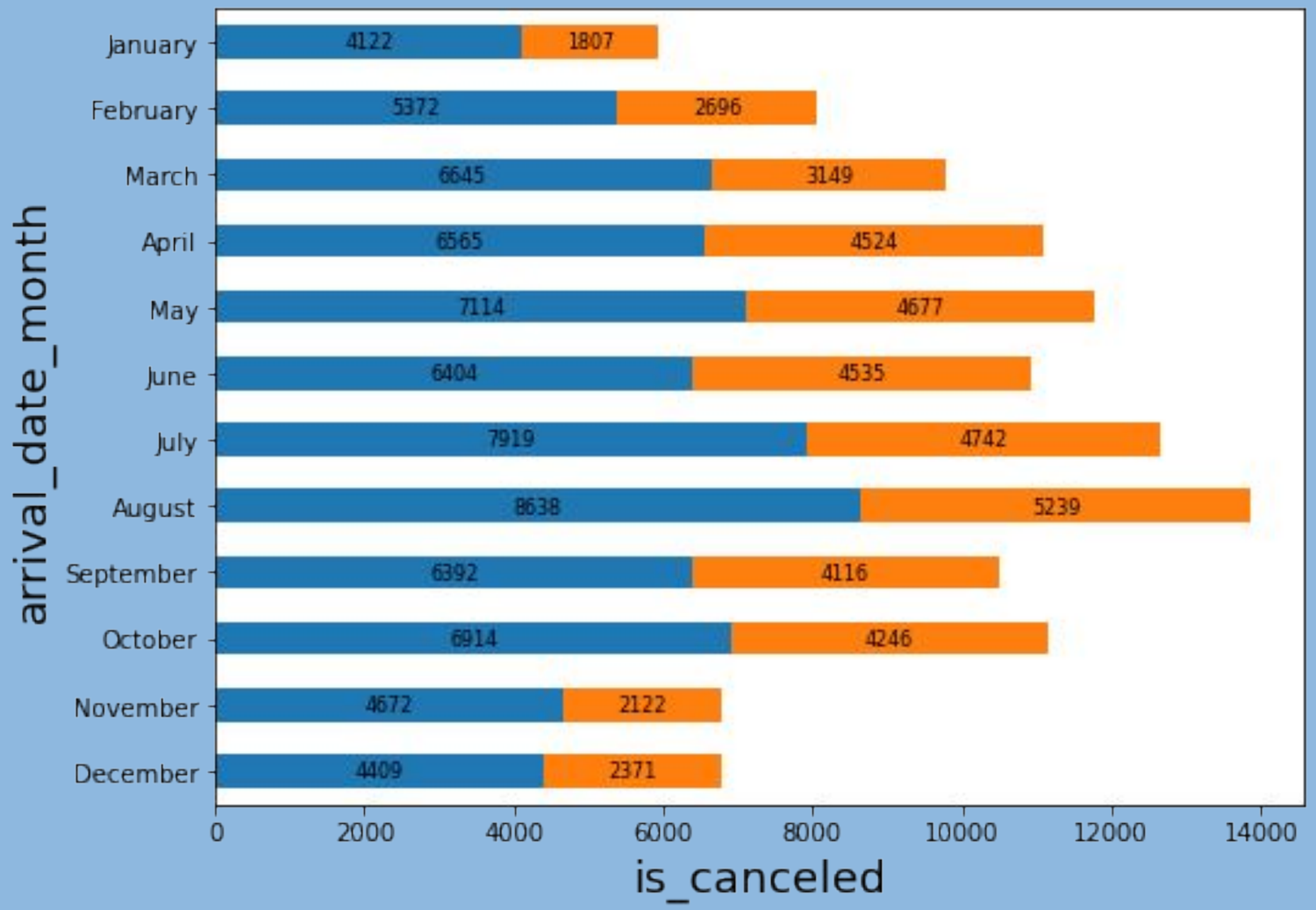
Market Segment & Cancellation



Pelanggan cenderung menyukai melakukan booking melalui online travel agent

Melalui platform online travel agent ini juga, terjadi paling banyak cancel yang dilakukan oleh customer

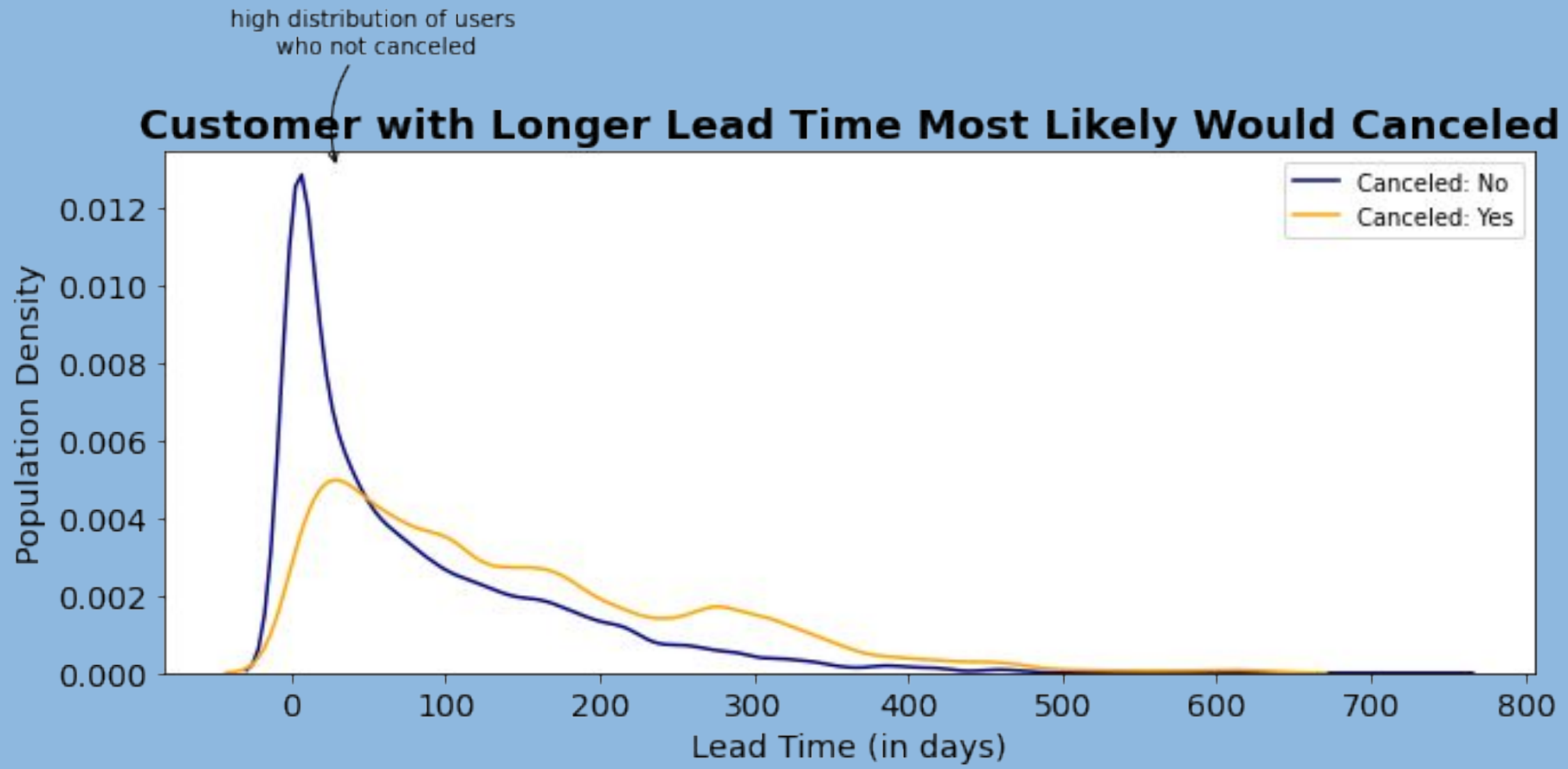
Arrival Month & Cancellation



Bulan Agustus menjadi bulan yang **paling banyak terjadi booking**

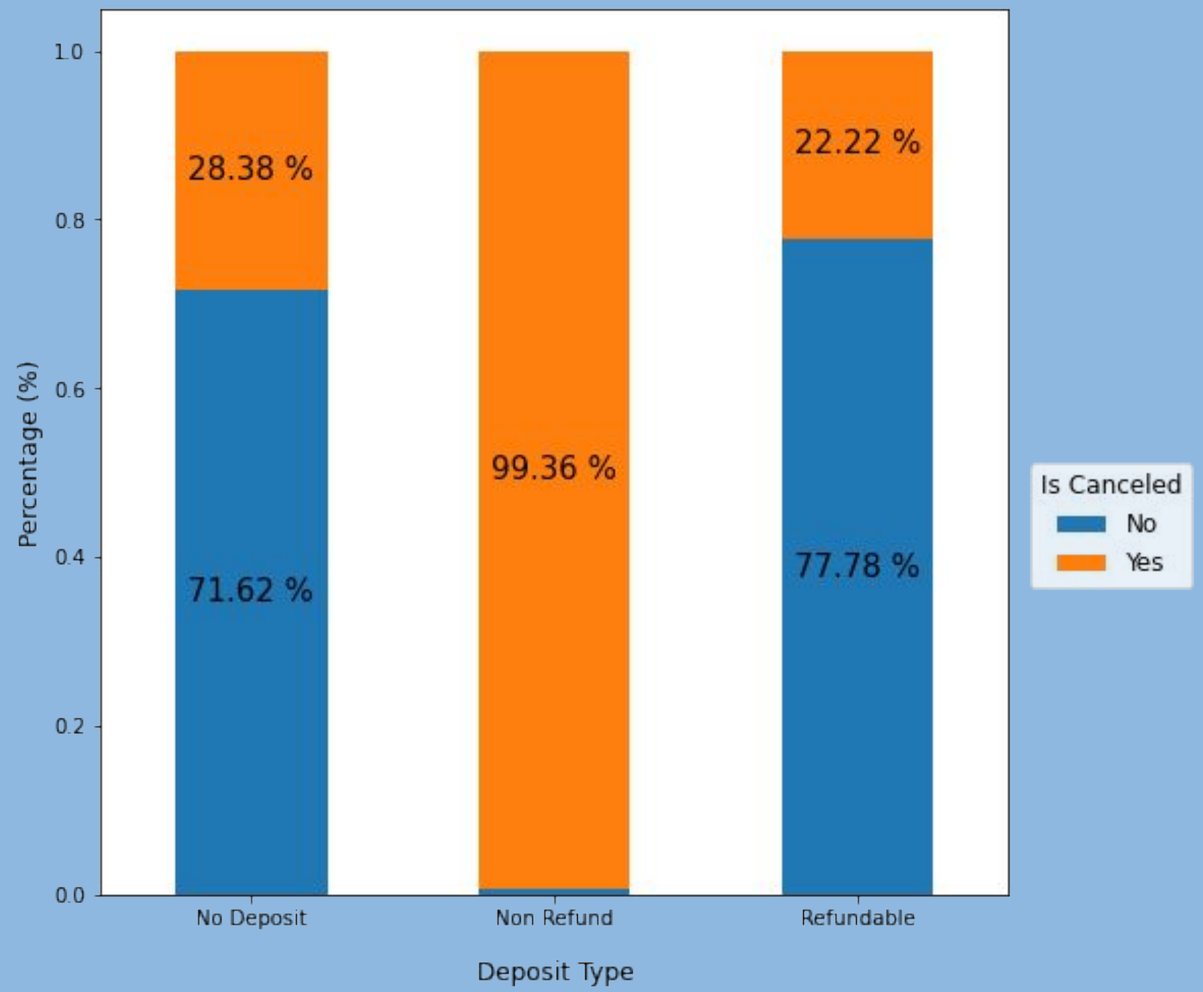
Bulan **Agustus** merupakan puncak summer di Portugal

Lead Time & Cancellation



Customer yang memiliki **lead time yang lebih lama memiliki tingkat canceled yang lebih tinggi**

Deposit Type & Cancellation



Pemesanan hotel dengan jenis deposit **non refund memiliki cancellation rate terbesar** yaitu mencapai 99%. Salah satu alasannya adalah karena non refund memiliki median lead time tertinggi

Data Preprocessing

Data Cleansing

Handle missing value & duplicate

Drop kolom personal data

Adjust data type

Handle outlier

Data Treatment

Log Transformasi &
Normalisasi

Encoding:
Labelling & OHE

Dataframe

Handling outlier dengan metode Z Score

df1

- Tidak handling data cleansing
- Tidak feature transformasi
- Tidak menghapus outlier dengan metode Z Score

df2

- Handling data cleansing
- Handling feature transformasi
- Tidak menghapus outlier dengan metode Z Score

df3

- Handling data cleansing
- Handling feature transformasi
- Menghapus outlier dengan metode Z Score

Feature Extraction

- reserved_room_types
- assigned_room_types

→

- reserved_vs_assigned

- arrival_date_month

→

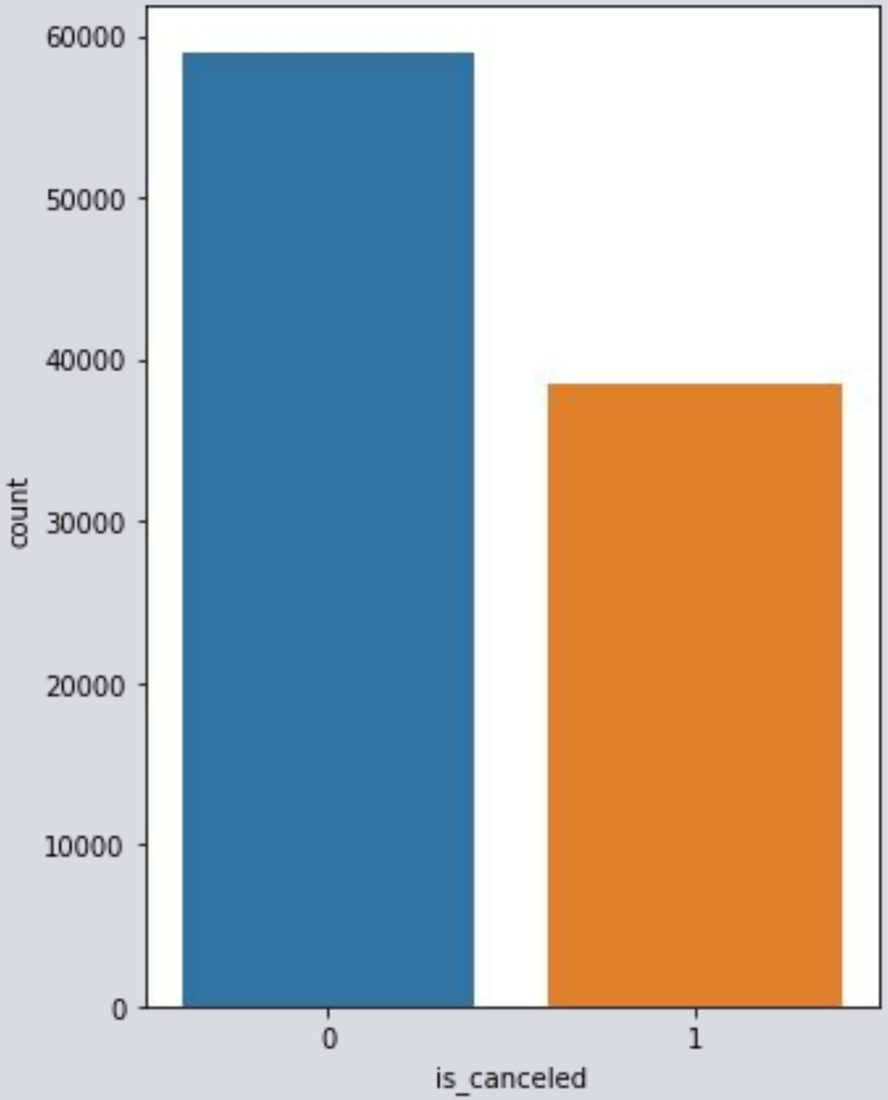
- season

- country

→

- origin_type

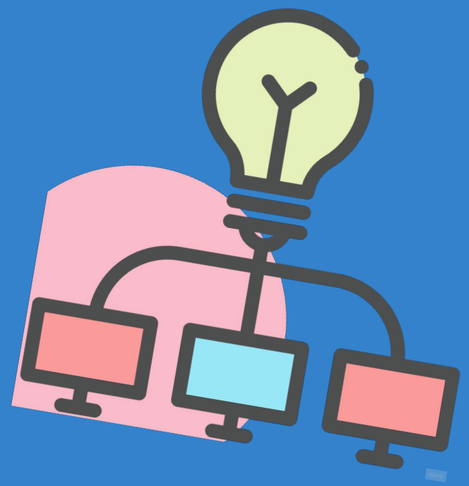
Class Imbalance



Tidak dilakukan **oversampling/ undersampling** karena target cukup proporsional

not canceled = 60,5% : was canceled = 39,5%

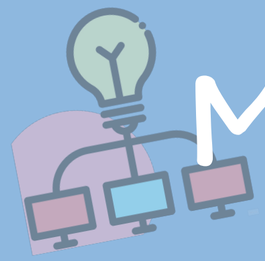
Modeling & Evaluation



Split Data
Train & Test

70 : 30

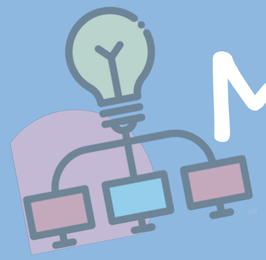
70% menjadi data train dan 30% sisanya data test



Modeling Preparation

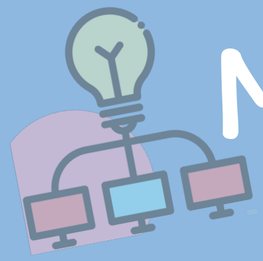
Kandidat Modeling

- Logistic Regression
- kNN
- Decision Tree
- GaussiaNB
- Random Forest
- XGBoost
- CatBoost
- AdaBoost



Model Evaluation

- Metriks yang digunakan adalah **recall**.
- Bertujuan untuk menekan angka false negatif (**diprediksi tidak cancel ternyata iya**).
- Selain dapat menurunkan cancellation rate, **tim bisnis juga dapat fokus kepada customer yang diprediksi false positif** (diprediksi cancel ternyata tidak) sehingga cost untuk mempertahankan customer agar tidak cancel menjadi berkurang.

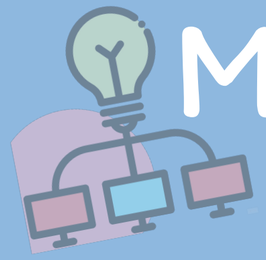


Modeling

Modeling 1:
menggunakan semua fitur

Modeling 2:
feature selection dengan featurewiz

model masih overfitting / recall masih rendah

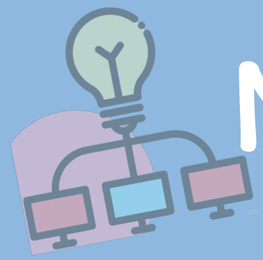


Modeling Result

memilih model yang memiliki **recall** terbaik dengan treshold **roc_auc** validation **tidak lebih dari 10%**

	df1 XGBoost	df2 CatBoost	df3 XGBoost
Recall	0.76	0.76	0.75
Precision	0.83	0.83	0.83
Accuracy	0.85	0.85	0.84
ROU_AUC (test)	0.87	0.87	0.87
ROU_AUC (train)	0.94	0.94	0.92

- Semua model sudah bagus.
- Sehingga dipilih model yang tidak ada indikasi overfitting / underfitting
- Juga merujuk ke insight yang kita inginkan “tidak ada lead time yang tinggi” sehingga seharunya tidak ada outlier

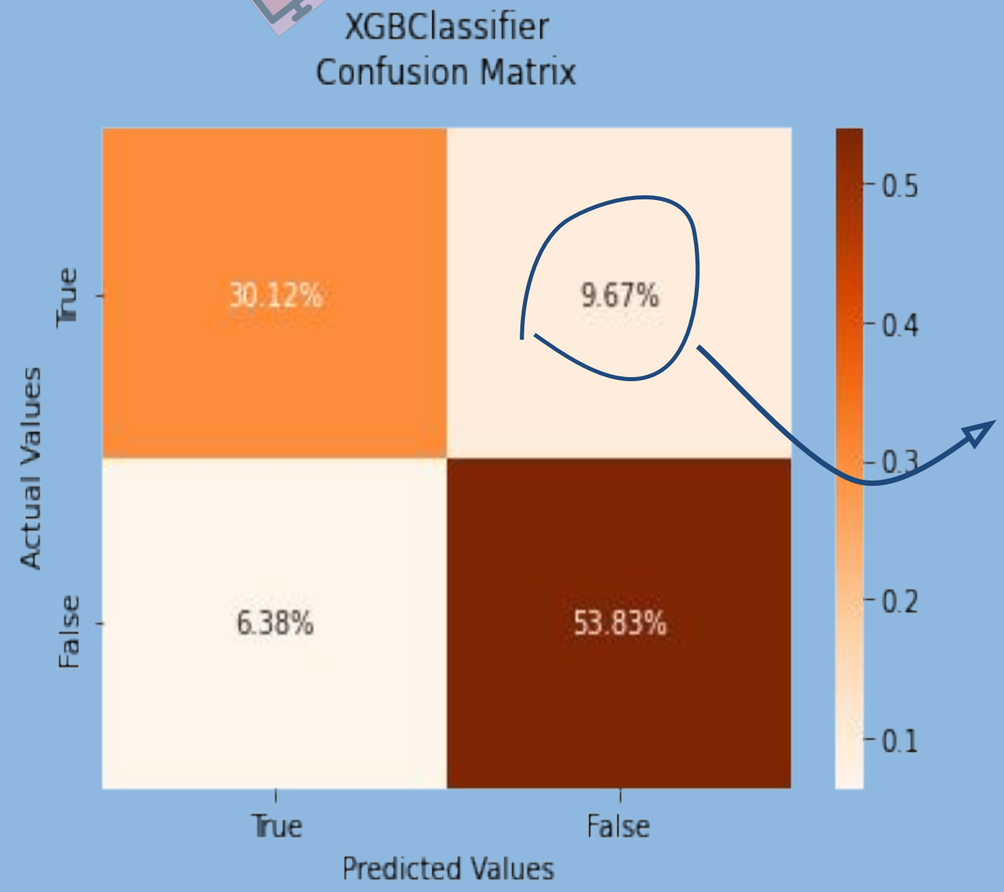


Model Result Hyperparameter Tuning

	XGBoost (df3)
Recall	0.76
Precision	0.83
Accuracy	0.84
ROU_AUC (test)	0.87
ROU_AUC (train)	0.94

Setelah dilakukan hyperparameter tuning, score recall mengalami peningkatan dari 75% menjadi 76%

Confusion Matric

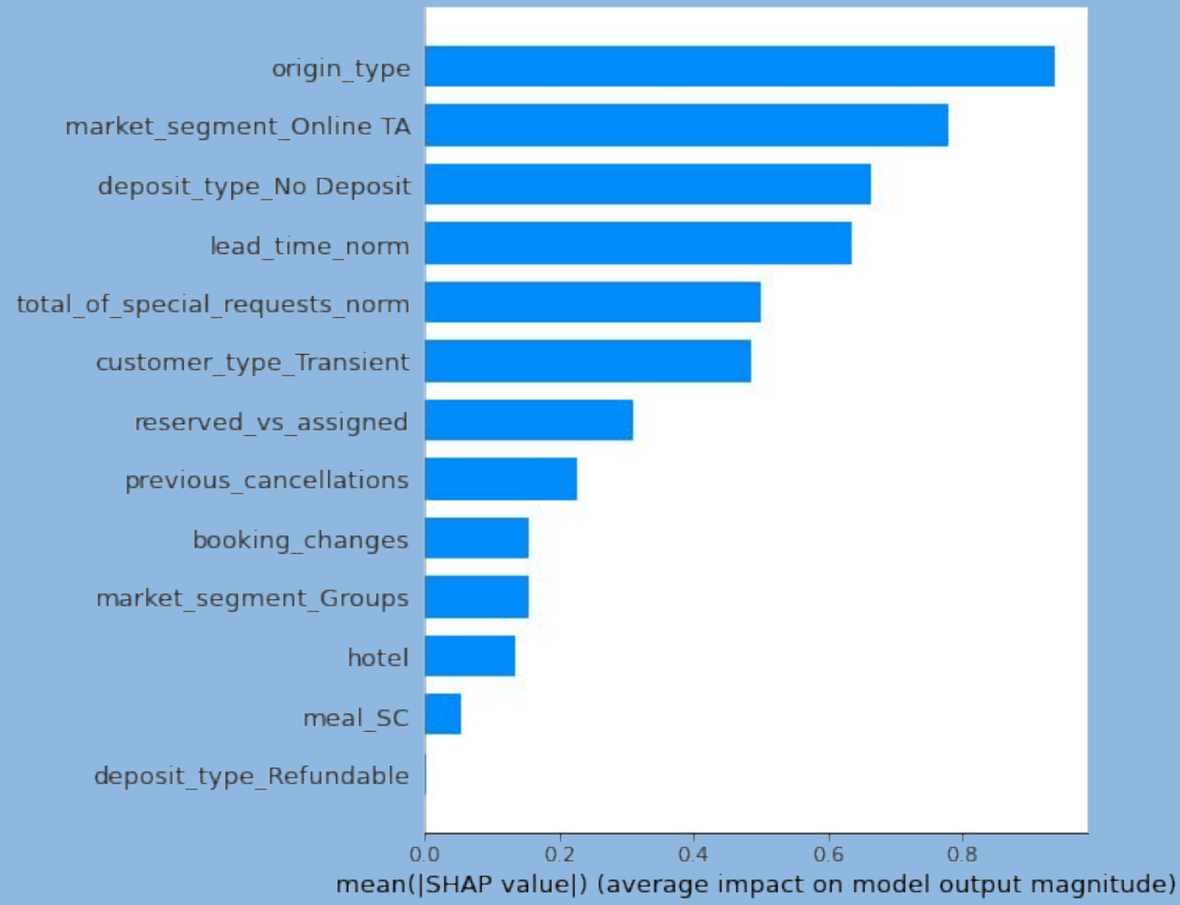


Recall = 76%

9.67% terprediksi tidak cancel tapi kenyataannya cancel

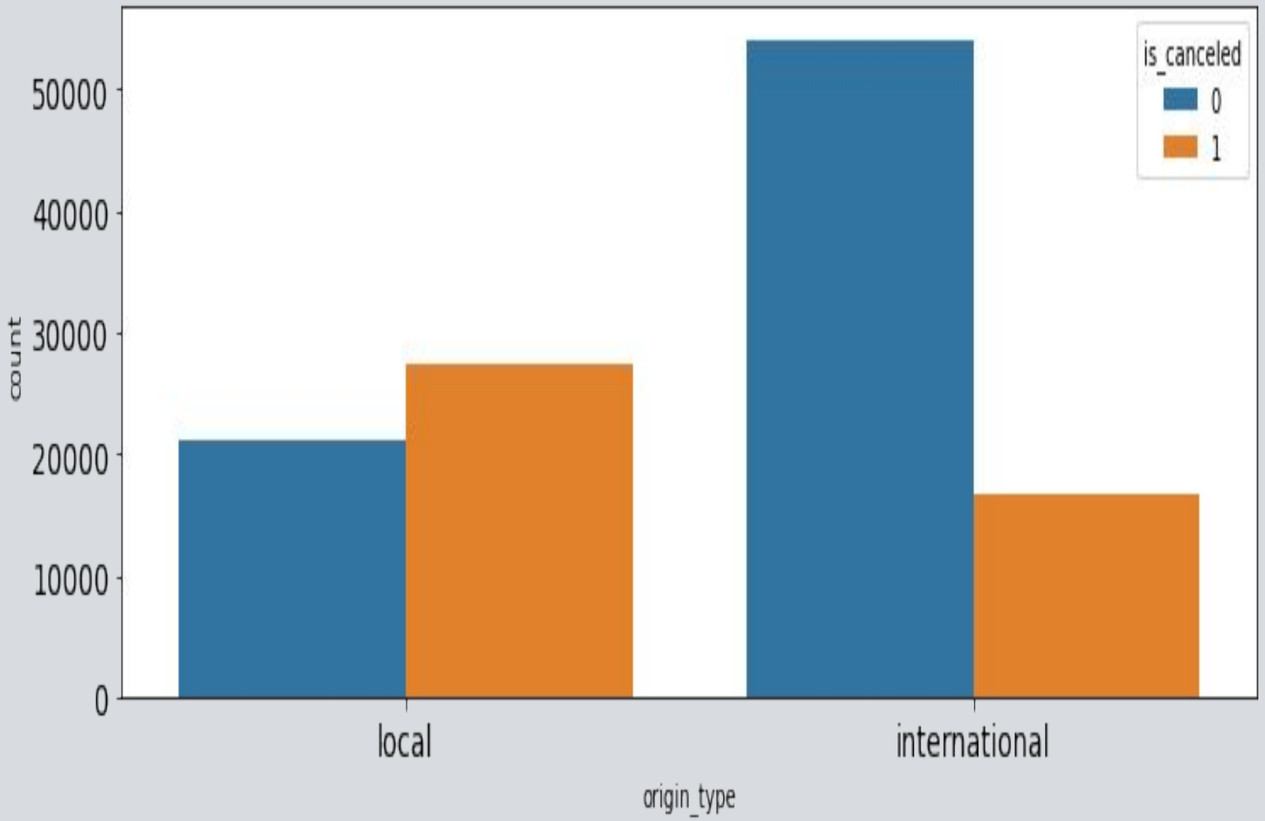
Angka ini diharapkan dapat ditekan dengan penambahan fitur untuk meningkatkan performa ML, seperti price, age, travel purpose, bed type

SHAP Value & Summary Plot



Business Recommendation

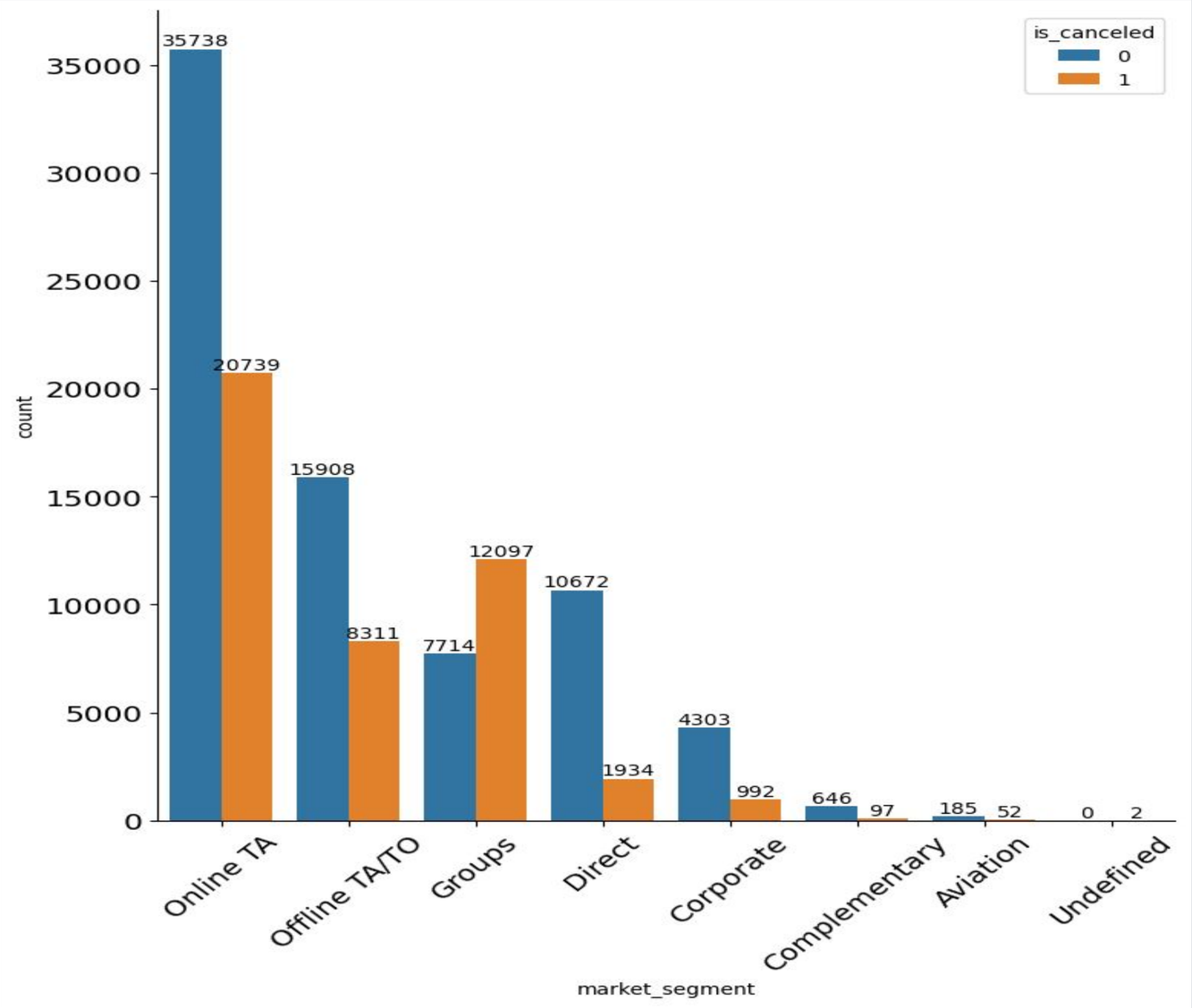
Business Recommendation - origin_type



Business Action:

- Membuat **promo khusus untuk turis lokal** seperti time-limited offer yang hanya dapat diklaim dalam waktu tertentu.
- Promo khusus turis local juga dilakukan dengan campaign hotel menggunakan google ads dengan penargetan lokasi geografis Portugal

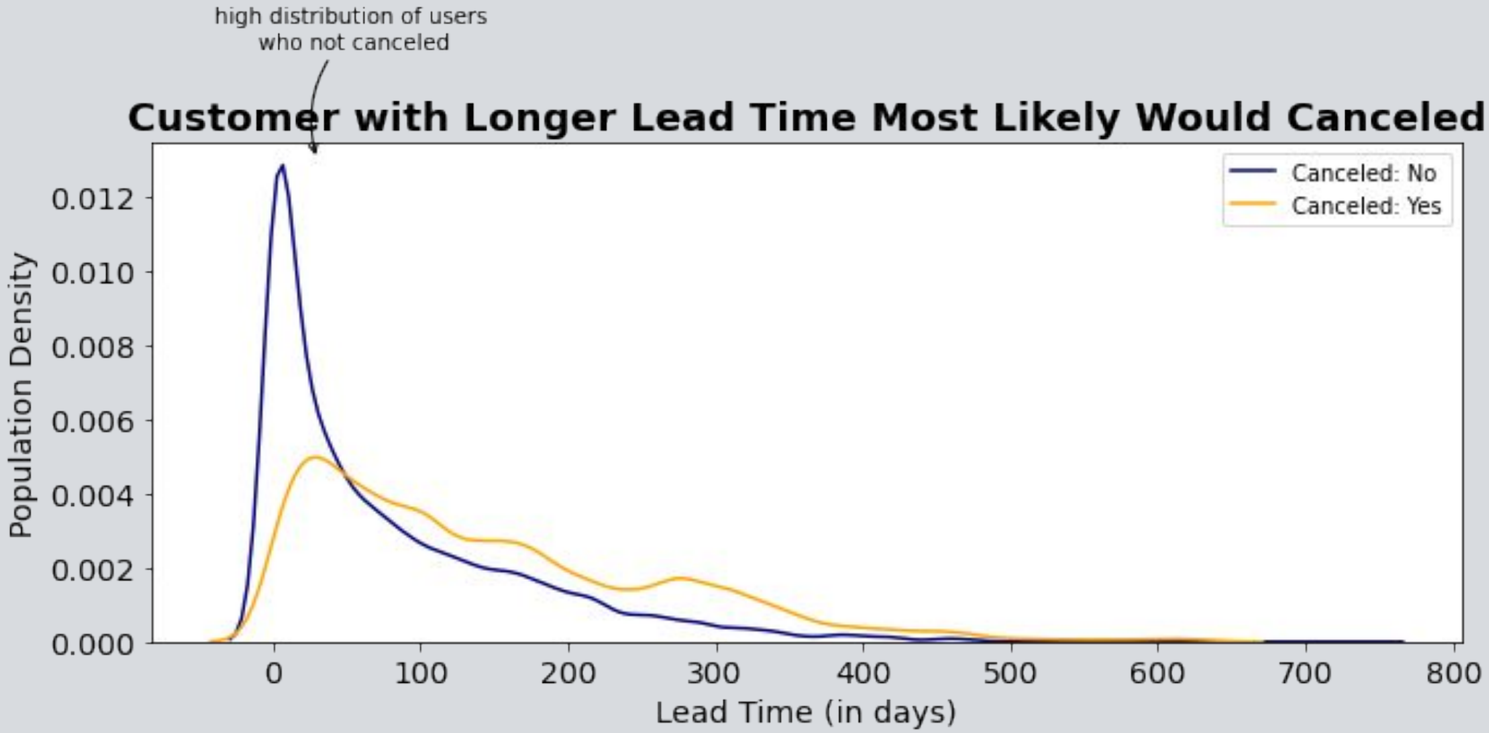
Business Recommendation - market_segment



Business Action:

Melakukan **kerja sama dengan berbagai platform online travel agent** seperti melakukan promo khusus bagi customer yang booking hotel melalui online travel agent tertentu.

Business Recommendation - lead_time

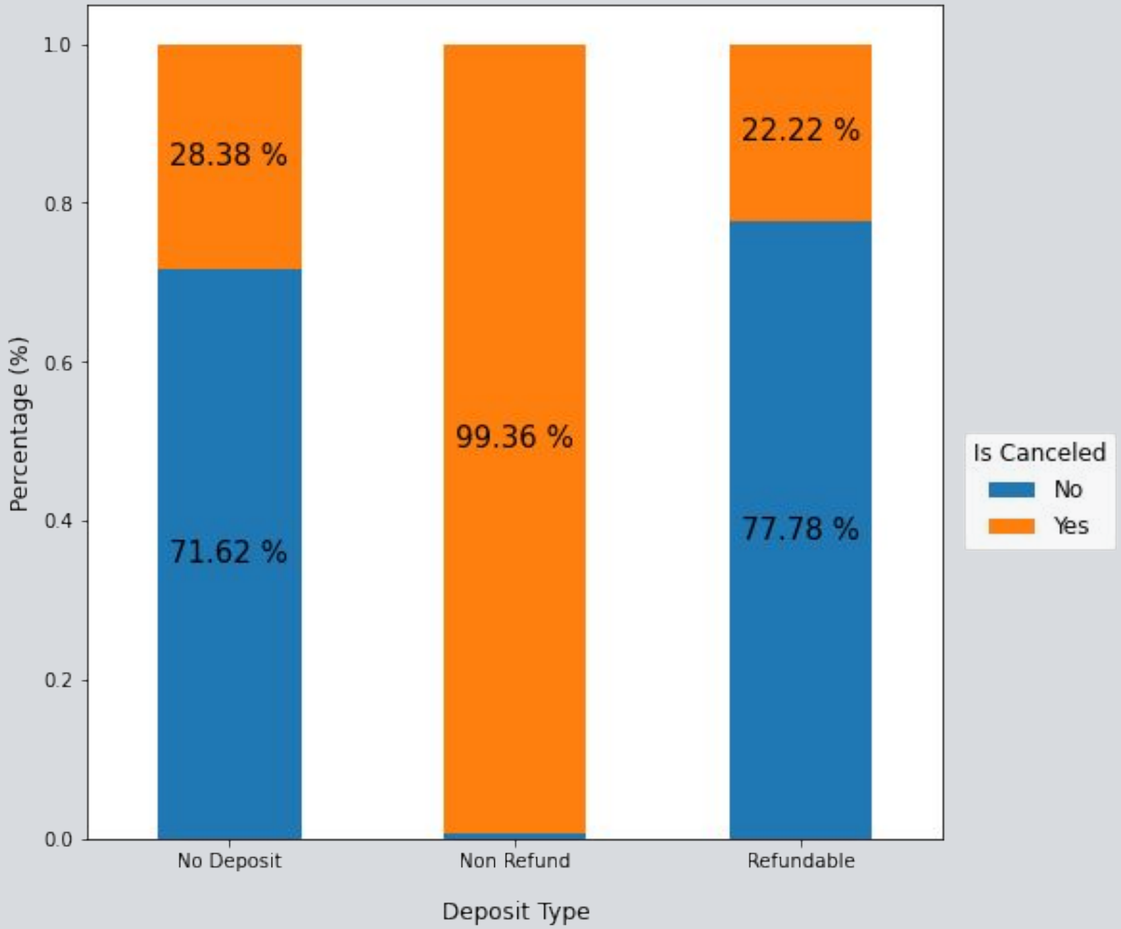


Business Action:

Membuat regulasi **open reserved maksimal 60 hari sebelum hari kedatangan***. Hal ini juga akan memudahkan pihak hotel untuk menerapkan pricing room secara dinamis tergantung event yang akan terjadi pada tanggal yang di booking oleh konsumen

(*)merujuk pada sumber eksternal (D-EDGE, 2019)

Business Recommendation - deposit_type



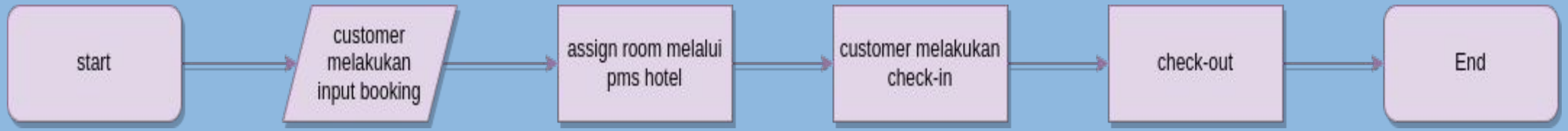
Business Action:

Membuat regulasi maksimal 60 hari sebelum hari kedatangan untuk tipe no deposit. Jenis deposit non refund akan diterapkan pada pemesanan dengan lead time lebih dari 60 hari

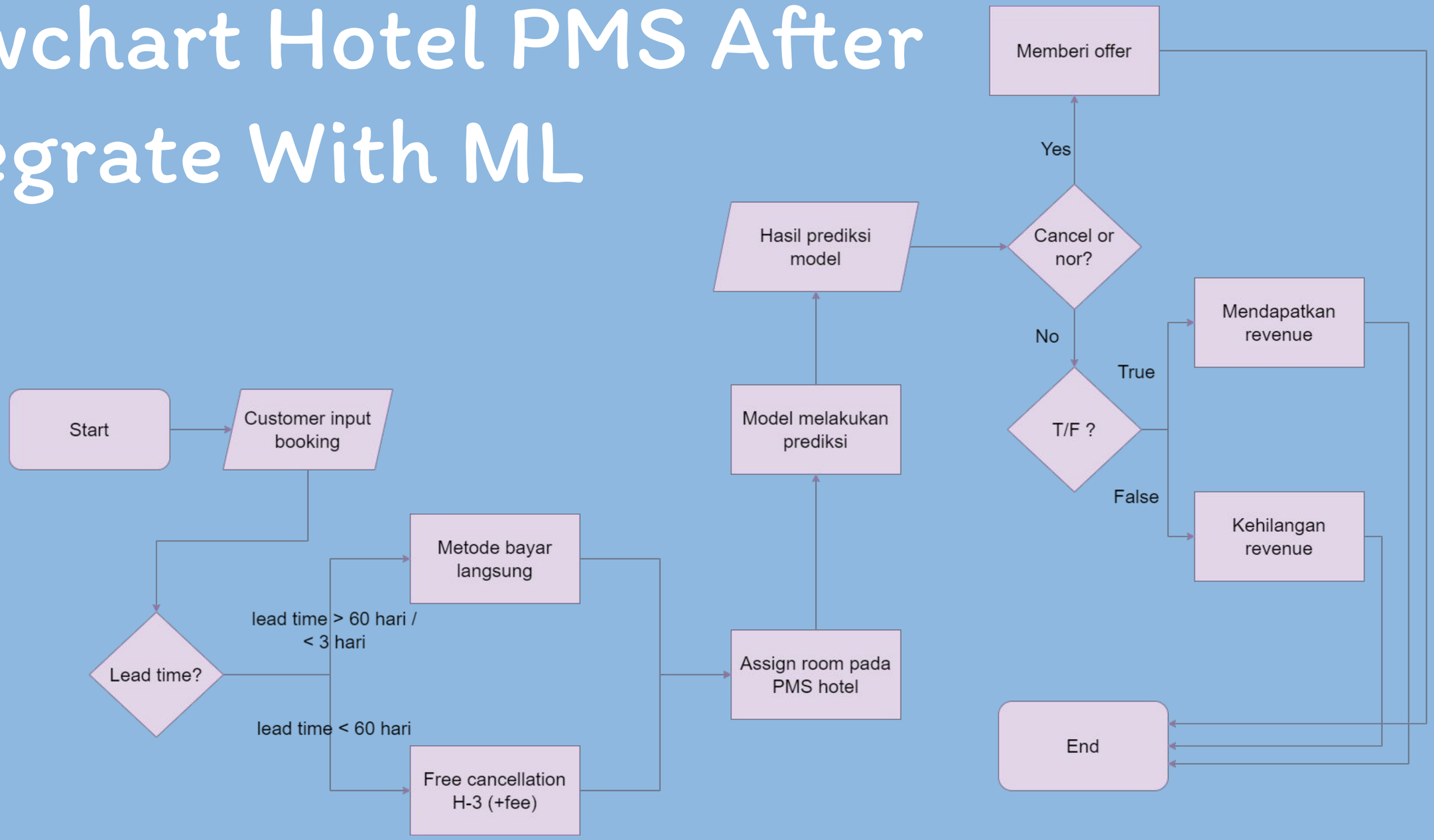
"reservasi dengan lead time lebih dari 60 hari memiliki kemungkinan 65% lebih tinggi untuk dibatalkan" - D-EDGE, The European No1 and World No3 hotel distribution technology provider in hospitality

Simulasi

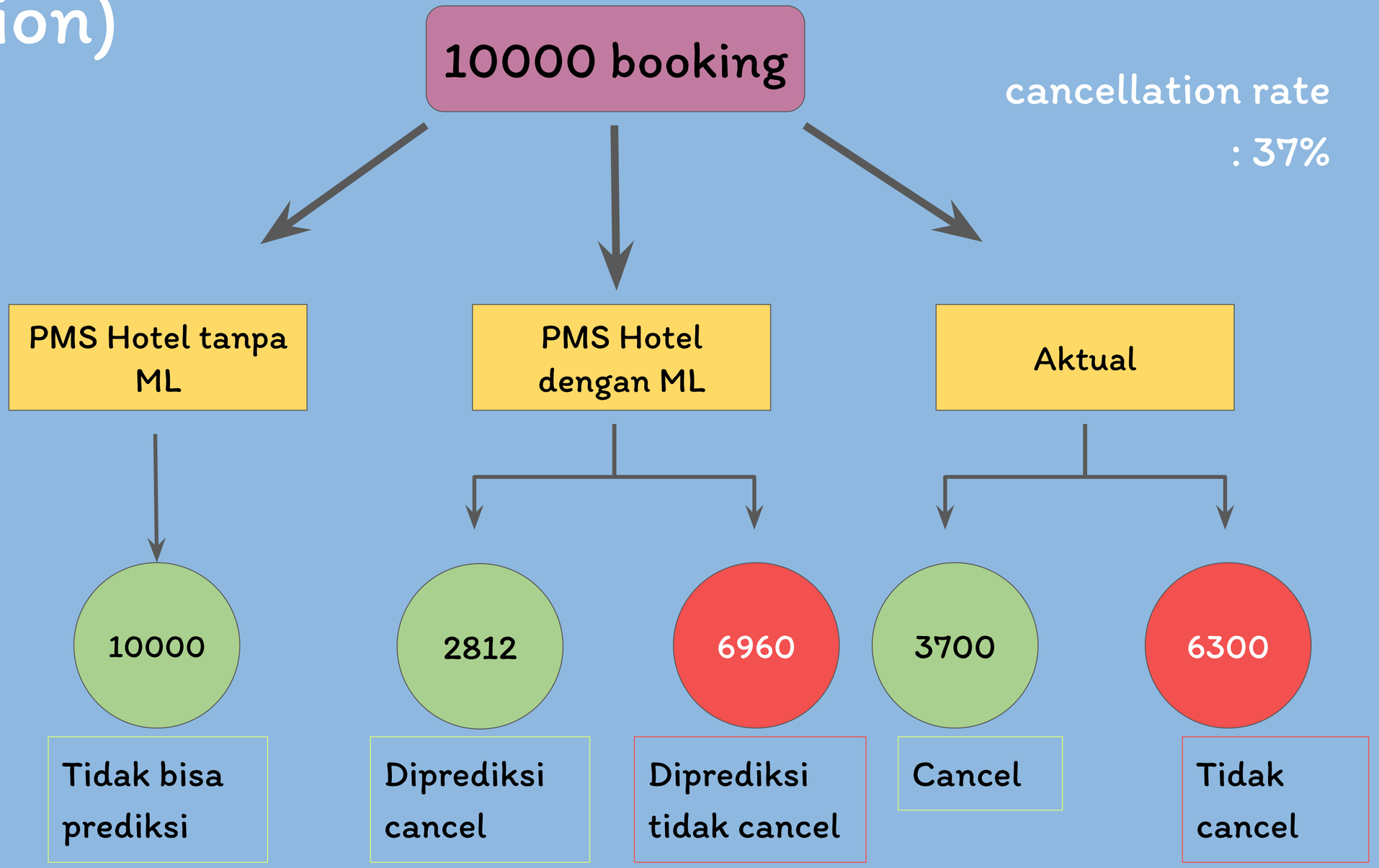
Flowchart Hotel PMS
Before ML



Flowchart Hotel PMS After Integrate With ML



Simulation Canceled Prediction (Assumption)



Business Simulation for Revenue

Room Type	Room Price	% Reserved Room	% Cancel	Room Qty	Room Cancel	Room not Cancel (Check In)	Revenue
A	18.000 €	72.03%	76.04%	7203	2813	4390	79.020.000 €
B	19.000 €	0.94%	0.83%	94	31	63	1.197.000 €
C	21.000 €	0.78%	0.70%	78	26	52	1.092.000 €
D	22.000 €	16.08%	13.80%	1608	511	1097	24.134.000 €
E	24.000 €	5.47%	4.33%	547	160	387	9.288.000 €
F	29.000 €	2.43%	1.99%	243	74	169	4.901.000 €
G	30.000 €	1.75%	1.73%	175	64	111	3.330.000 €
H	36.000 €	0.5%	0.55%	50	20	30	1.080.000 €
L	41.000 €	0.01%	0.00%	1	0	1	41.000 €
P	50.000 €	0.01%	0.03%	1	1	0	- €
				10000	3700	6300	
						Total Revenue	124.083.000 €

Room Type	Room Price	% Reserved Room	% Cancel	Room Qty	Room Cancel	Not Cancel*	Final Cancel	Room not Cancel (Check In)	Revenue
A	18.000 €	72.03%	76.04%	7203	2813	1550	1263	5940	106.920.000 €
B	19.000 €	0.94%	0.83%	94	31	20	11	83	1.577.000 €
C	21.000 €	0.78%	0.70%	78	26	18	8	70	1.470.000 €
D	22.000 €	16.08%	13.80%	1608	511	266	245	1363	29.986.000 €
E	24.000 €	5.47%	4.33%	547	160	90	70	477	11.448.000 €
F	29.000 €	2.43%	1.99%	243	74	40	34	209	6.061.000 €
G	30.000 €	1.75%	1.73%	175	64	36	28	147	4.410.000 €
H	36.000 €	0.5%	0.55%	50	20	15	5	45	1.620.000 €
L	41.000 €	0.01%	0.00%	1	0	0	0	1	41.000 €
P	50.000 €	0.01%	0.03%	1	1	0	1	0	- €
				10000	3700	2035	1665	8335	
								Total Revenue	163.533.000 €

Asumsi pemesanan yang berhasil diselamatkan dan tidak jadi cancel : 55%

Peningkatan revenue sebesar 39.450.000 €

(*) Pemesanan yang diprediksi cancel dan tidak jadi melakukan cancel setelah diberi offer treatment marketing (kelompok offer dapat dilihat di lampiran)

Business Recomm Simulation - Wireframe UI/UX

Market Segment - Online TA

- Get Point
- -47% discount today

Strategi diskon dari OTA

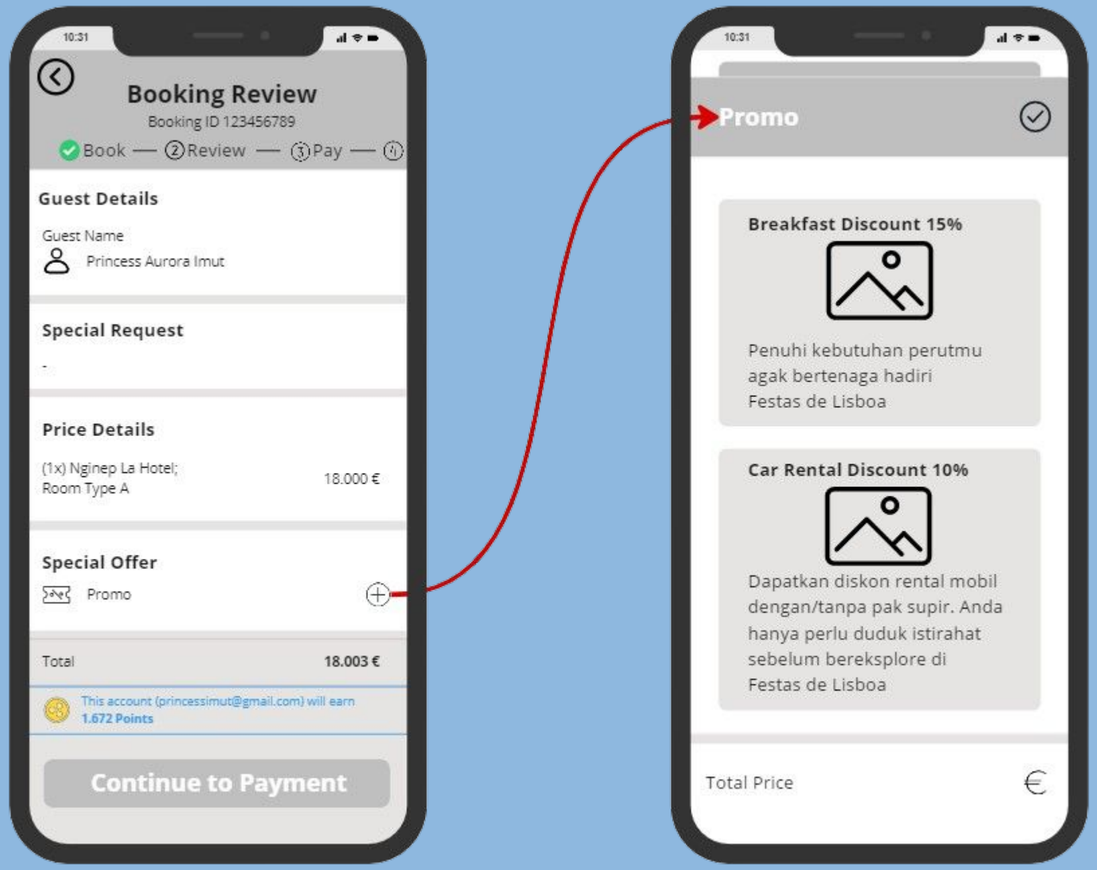
Deposit Type - Non Refund

- Lead Time > 60 hari
- Lead Time < 3 hari

Deposit Type - No Deposit

- Lead Time < 60 hari
- S&K berlalu:)

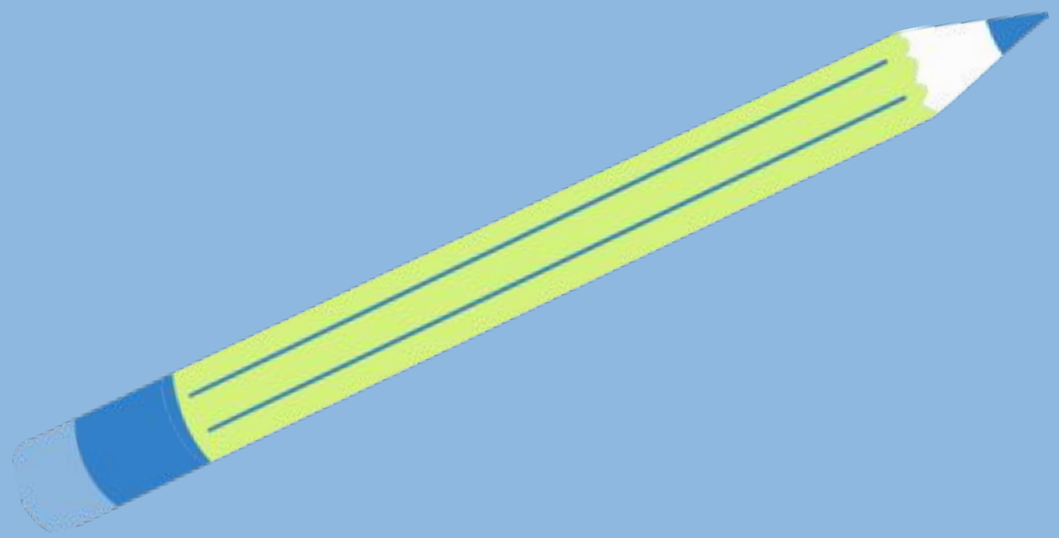
Business Recomm Simulation - Wireframe UI/UX



Origin Type - Local

- Promo muncul setelah OTA memastikan pesanan dilakukan oleh warga **local (Sudah mengisi data diri)**
- Promo disesuaikan dengan **event** di Portugal & promo khusus warga lokal lainnya

Business Recommendation (tambahan)



- Menambah fitur (travel purpose, age, price) untuk meningkatkan performa ML
- Memberi rekomendasi untuk memberi harga kamar terbaik pada lead time 24-26 hari sebelum puncak event/festival di Portugal (dari sumber eksternal diketahui kebanyakan traveller di Spain negara tetangga Portugal memesan hotel 25 hari sebelum kedatangan)
- Menawarkan offer yang tepat sesuai segmentasi yang disusun dari insight dan modeling (kategori offer dapat dilihat di lampiran)

Referensi

Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel Booking Demand Datasets. *Data in Brief*, *22*, 41–49. https://doi.org/10.1016/J.DIB.2018.11.126

Antonio, N., Almeida, A. de, & Nunes, L. (2017). Predicting Hotel Booking Cancellations to Decrease Uncertainty and Increase Revenue. *Tourism & Management Studies*, *13*(2), 25–39. https://doi.org/10.18089/tms.2017.13203

Antonio, N., de Almeida, A., & Nunes, L. (2017). Predicting Hotel bookings Cancellation With a Machine Learning Classification Model. *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, *2017-December*, 1049–1054. https://doi.org/10.1109/ICMLA.2017.00-11

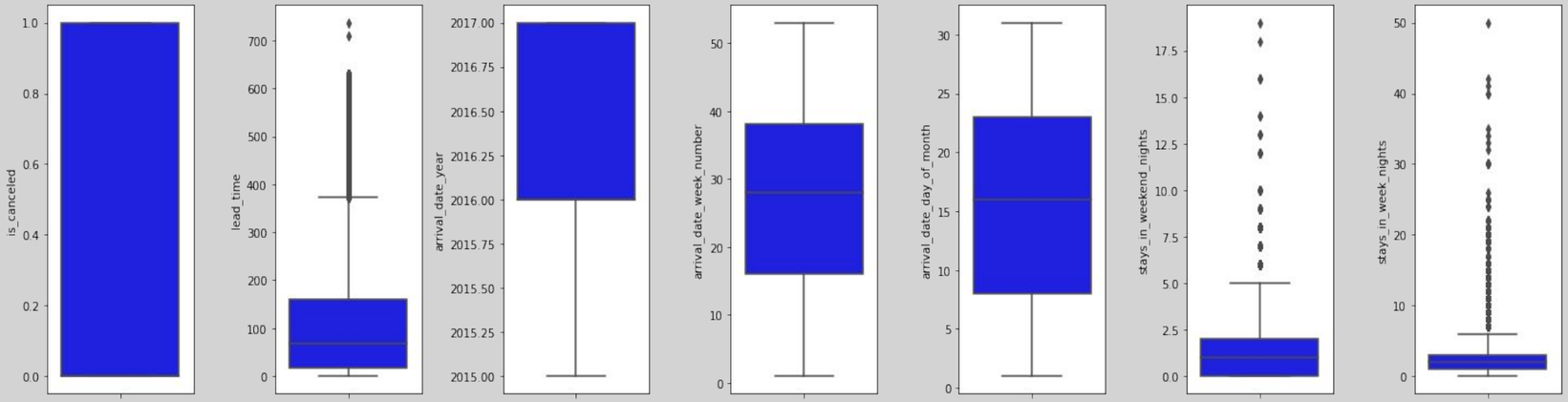
Boost Hotel bookings by optimizing your Booking Lead Time, (March 10, 2020). https://blog.experience-hotel.com/boost-hotel-bookings-by-optimizing-your-booking-lead-time/

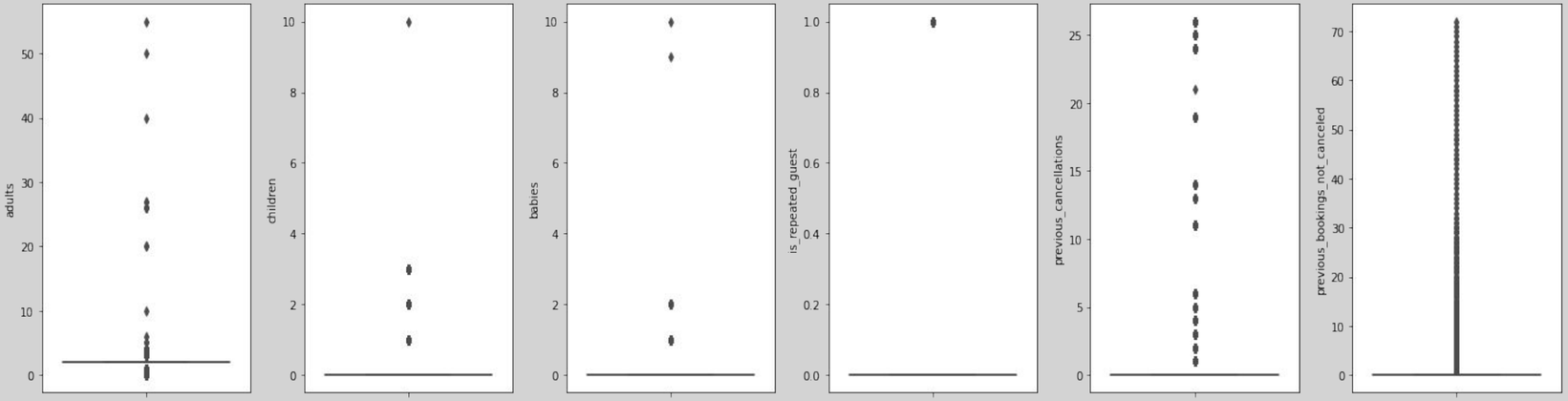
How Online Hotel Distribution Changing Is In Europe, (October 4, 2019). https://www.d-edge.com/how-online-hotel-distribution-is-changing-in-europe/

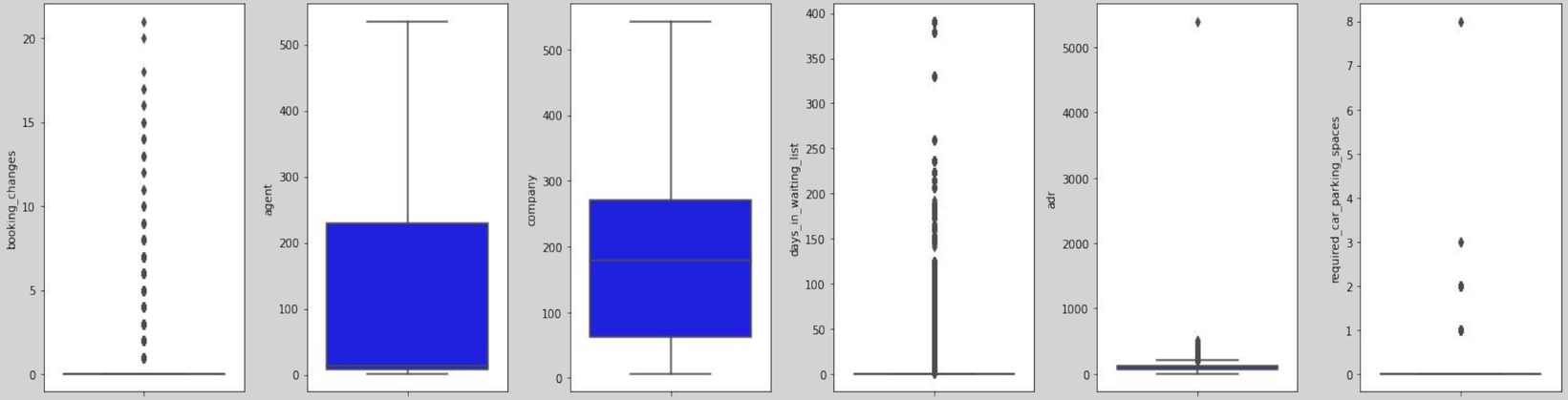
Sánchez-Medina, A. J., & C-Sánchez, E. (2020). Using Machine Learning and Big Data for Efficient Forecasting of Hotel Booking Cancellations. *International Journal of Hospitality Management*, *89*, 102546. https://doi.org/10.1016/J.IJHM.2020.102546

SiteMinder. (n.d.). *How Far do Hotel Guests Book in Advance?* Retrieved November 20, 2022, from https://www.siteminder.com/r/hotel-distribution/hotel-revenue-management/hotel-guests-book-advance/

Lampiran

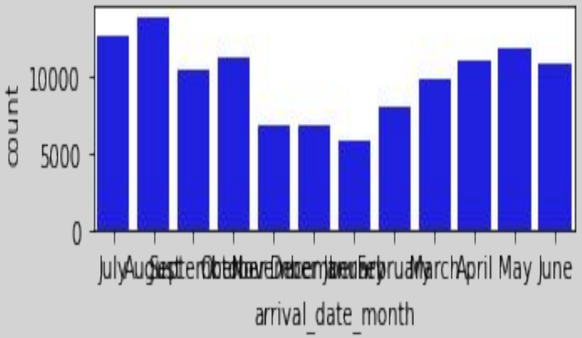

is_canceled
lead_time
arrival_date_year
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights

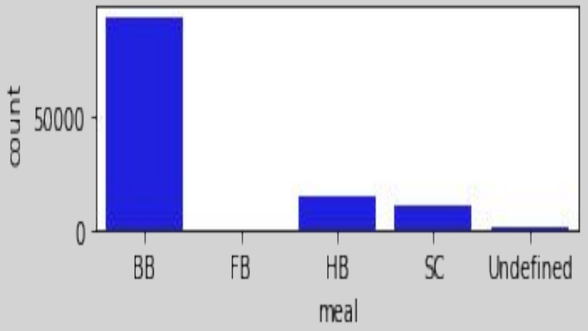

adults
children
babies
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled

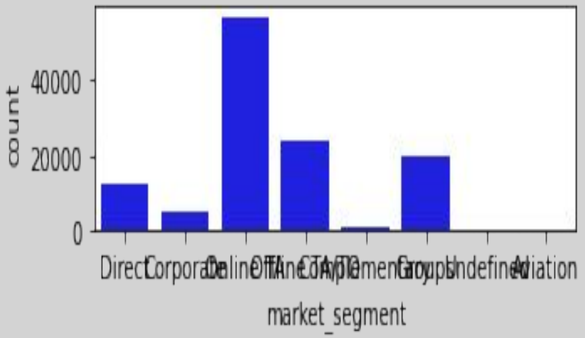

booking_changes
agent
company
days_in_waiting_list
adr
required_car_parking_spaces

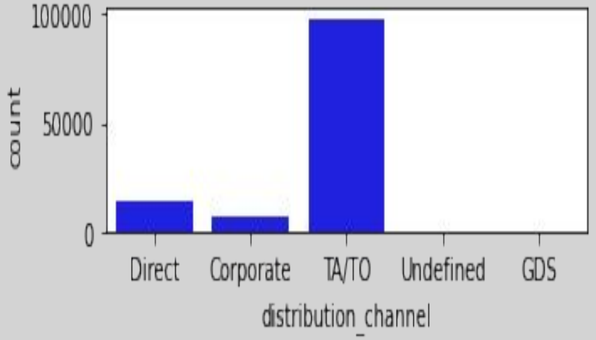
EDA - Boxplot

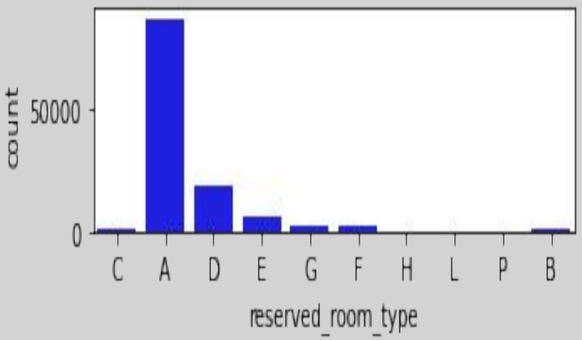
EDA - Countplot

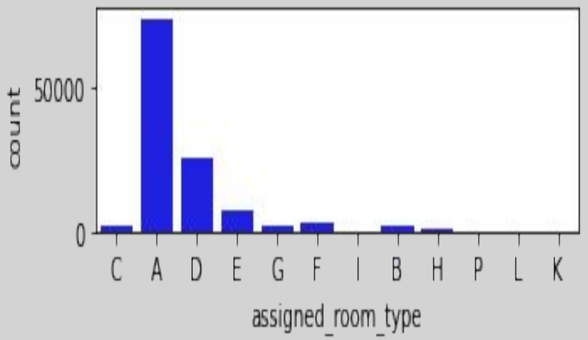


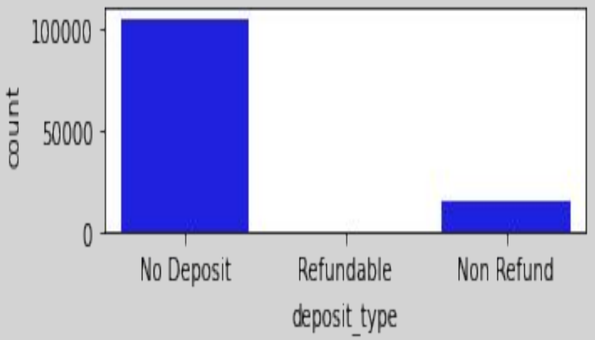


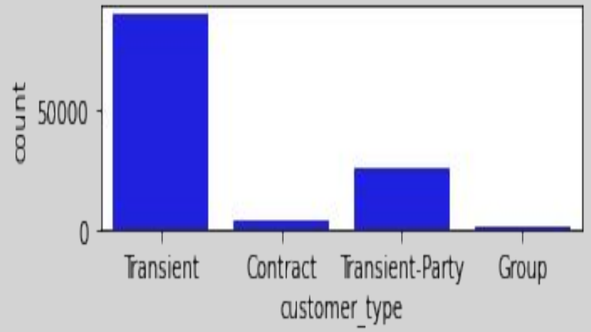


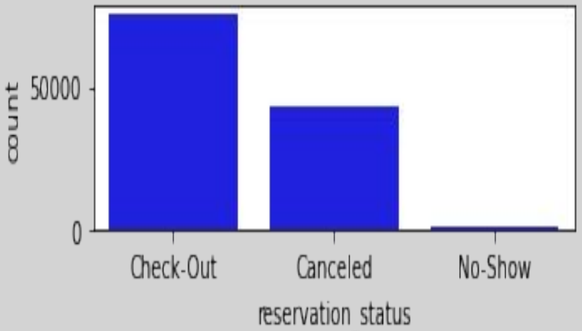












Feature Extraction

reserved_vs_assigned

```
#reserved_room_types dan assigned_room_types digabungkan menjadi reserved_vs_assigned (1 untuk reserved == assigned, 0 untuk reserved != assigned)
def reserved_assigned (reserved_room_type, assigned_room_type):
    if (reserved_room_type) == (assigned_room_type):
        reserved_assigned = '1'
    else:
      reserved_assigned = '0'
    return reserved_assigned

df_pre_new['reserved_vs_assigned'] = df_pre_new.apply(lambda x: reserved_assigned(x['reserved_room_type'], x['assigned_room_type']), axis = 1)
df_pre_new['reserved_vs_assigned'].describe()
```

season

```
#kolom season yang didapat dari kolom arrival_date_month
winter = ['December', 'January', 'February']
spring = ['March', 'April', 'May']
summer = ['June', 'July', 'August']
autumn = ['September', 'October', 'November']

def season(x):
  if x['arrival_date_month'] in winter:
    season = 'winter'
  elif x['arrival_date_month'] in spring:
    season = 'spring'
  elif x['arrival_date_month'] in summer:
    season = 'summer'
  else:
    season = 'autumn'
  return season

df_pre_new['season'] = df_pre_new.apply(lambda x: season(x), axis=1)
```

origin_type

```
#kolom turis local dan international di dapat dari kolom country
#PRT(portugal) = lokal karena datasetnya berasal dari portugal

def turis(x):
  if x['country'] == 'PRT':
    turis = 'local'
  else:
    turis = 'international'
  return turis

df_pre_new['origin_type'] = df_pre_new.apply(lambda x: turis(x), axis=1)
```

Modeling (2) - Featurwiz

Feature selection dengan featurewiz

Library **featurewiz** ini bekerja dengan menggunakan 2 tahapan yaitu tahap SULOV (Searching for uncorrelated list of variables) setelah itu fitur yang terpilih memasuki tahapan recursive XGBoost sebanyak 5 kali untuk mencari top X fitur pada setiap pengulangannya. Top X fitur pada setiap pengulangan ini akan digabungkan dan dihapus apabila ada duplikat, sehingga hasil akhirnya fitur akan menjadi lebih sederhana untuk menghindari overfitting

Feature Selection - Featurewiz

Dari

40 fitur

Menjadi

- df1 : 18 fitur
- df2 : 13 fitur
- df3 : 13 fitur

Fitur Modeling (2) - Featurwiz

Fitur df1 - 18 fitur

['deposit_type_No Deposit',
'market_segment_Online TA',
'total_of_special_requests_norm',
'reserved_vs_assigned',
'required_car_parking_spaces'
'origin_type',
'previous_cancellations',
'deposit_type_Refundable',
'lead_time_norm',
'booking_changes',
'customer_type_Transient',
'previous_bookings_not_cancele
d',
'meal_SC',
'market_segment_Groups',
'hotel',
'distribution_channel_TA/TO',
'meal_HB',
'meal_BB']

Fitur df2 - 18 fitur

['deposit_type_No Deposit',
'market_segment_Online TA',
'total_of_special_requests_norm',
'reserved_vs_assigned',
'required_car_parking_spaces',
'origin_type',
'previous_cancellations',
'deposit_type_Refundable',
'lead_time_norm',
'booking_changes',
'customer_type_Transient',
'previous_bookings_not_canceled'
'meal_SC',
'market_segment_Groups',
'hotel',
'distribution_channel_TA/TO',
'meal_HB',
'meal_BB']

Fitur df3 - 13 fitur

['deposit_type_No Deposit',
'previous_cancellations',
'market_segment_Online TA',
'reserved_vs_assigned',
'deposit_type_Refundable',
'total_of_special_requests_norm'
'origin_type',
'customer_type_Transient',
'booking_changes',
'lead_time_norm',
'meal_SC',
'hotel',
'market_segment_Groups']

df1 - Featurewiz

	Logistic Regression	Knn	Decision Tree	GaussiaNB	Random Forest	XGBoost	CatBoost	AdaBoost
Accuracy (Test Set)	0.81	0.84	0.84	0.55	0.84	0.85	0.85	0.82
Precision (Test Set)	0.81	0.80	0.80	0.45	0.80	0.83	0.83	0.81
Recall (Test Set)	0.64	0.76	0.75	0.97	0.77	0.76	0.76	0.68
F1-Score (Test Set)	0.72	0.78	0.78	0.62	0.79	0.79	0.80	0.74
roc_auc (test-proba)	0.89	0.90	0.88	0.84	0.92	0.93	0.93	0.90
roc_auc (test-proba)	0.89	0.94	0.98	0.84	0.97	0.94	0.94	0.90
roc_auc (crossval train-mean)	0.89502	0.94134	0.97578	0.84283	0.97327	0.94052	0.94245	0.90512
roc_auc (crossval test-mean)	0.87715	0.80413	0.75539	0.83418	0.82737	0.87708	0.88005	0.88879
roc_auc (crossval train-std)	0.00180	0.00356	0.00221	0,00521	0.00203	0.00125	0.00161	0.00154
roc_auc (crossval test-std)	0.0278	0.0302	0.0543	0.02940	0.02763	0.02242	0.02501	0.0236

df2 - featurewiz

	Logistic Regression	Knn	Decision Tree	GaussiaNB	Random Forest	XGBoost	CatBoost	AdaBoost
Accuracy (Test Set)	0.81	0.84	0.84	0.55	0.84	0.85	0.85	0.82
Precision (Test Set)	0.81	0.80	0.80	0.45	0.80	0.83	0.83	0.81
Recall (Test Set)	0.64	0.76	0.75	0.97	0.77	0.76	0.76	0.68
F1-Score (Test Set)	0.72	0.78	0.78	0.62	0.79	0.79	0.80	0.74
roc_auc (test-proba)	0.89	0.90	0.88	0.84	0.92	0.93	0.93	0.90
roc_auc (test-proba)	0.89	0.94	0.98	0.84	0.97	0.94	0.94	0.90
roc_auc (crossval train-mean)	0.89502	0.94134	0.97578	0.84283	0.97327	0.94052	0.94245	0.90512
roc_auc (crossval test-mean)	0.87715	0.80413	0.75539	0.83418	0.82737	0.877084	0.88005	0.88879
roc_auc (crossval train-std)	0.00180	0.00356	0.00221	0.00521	0.00203	0.00125	0.00161	0.00154
roc_auc (crossval test-std)	0.02780	0.03020	0.05433	0.02940	0.02763	0.02242	0.02501	0.02362

Hasil Modeling (2) - featurwiz

df3 - featurewiz

	Logistic Regression	Knn	Decision Tree	GaussiaNB	Random Forest	XGBoost	CatBoost	AdaBoost
Accuracy (Test Set)	0.79	0.82	0.83	0.75	0.83	0.84	0.84	0.81
Precision (Test Set)	0.80	0.80	0.81	0.98	0.80	0.83	0.83	0.81
Recall (Test Set)	0.64	0.75	0.74	0.39	0.76	0.75	0.75	0.68
F1-Score (Test Set)	0.71	0.77	0.77	0.55	0.78	0.79	0.79	0.74
roc_auc (test-proba)	0.88	0.89	0.88	0.82	0.90	0.91	0.91	0.89
roc_auc (test-proba)	0.88	0.93	0.96	0.82	0.96	0.92	0.92	0.89
roc_auc (crossval train-mean)	0.87867	0.89735	0.96331	0.82141	0.96077	0.92749	0.92791	0.88081
roc_auc (crossval test-mean)	0.86821	0.78937	0.76307	0.80814	0.82115	0.87472	0.87823	0.87780
roc_auc (crossval train-std)	0.00469	0.01722	0.00171	0.00664	0.00154	0.00157	0.00135	0.00396
roc_auc (crossval test-std)	0.03302	0.03427	0.06023	0.05261	0.03803	0.03333	0.03035	0.02738

Hasil Modeling (2) - featurwiz

source code df 1 dapat dilihat disini

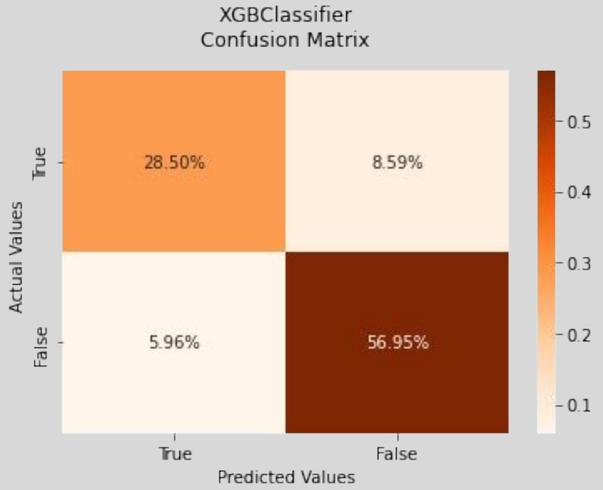
source code df 2 dapat dilihat disini

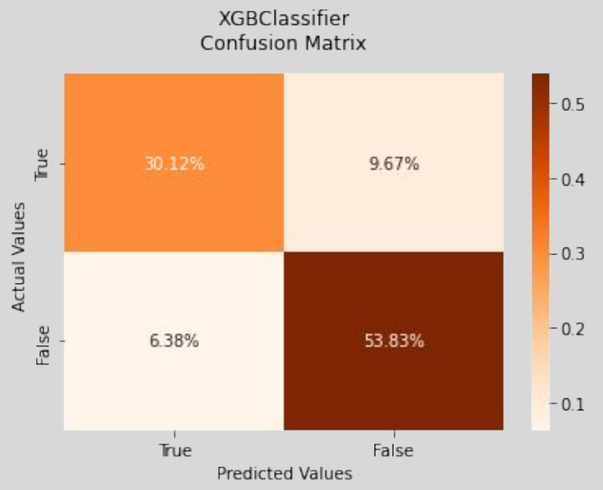
Hasil Modeling (2) - featurwiz

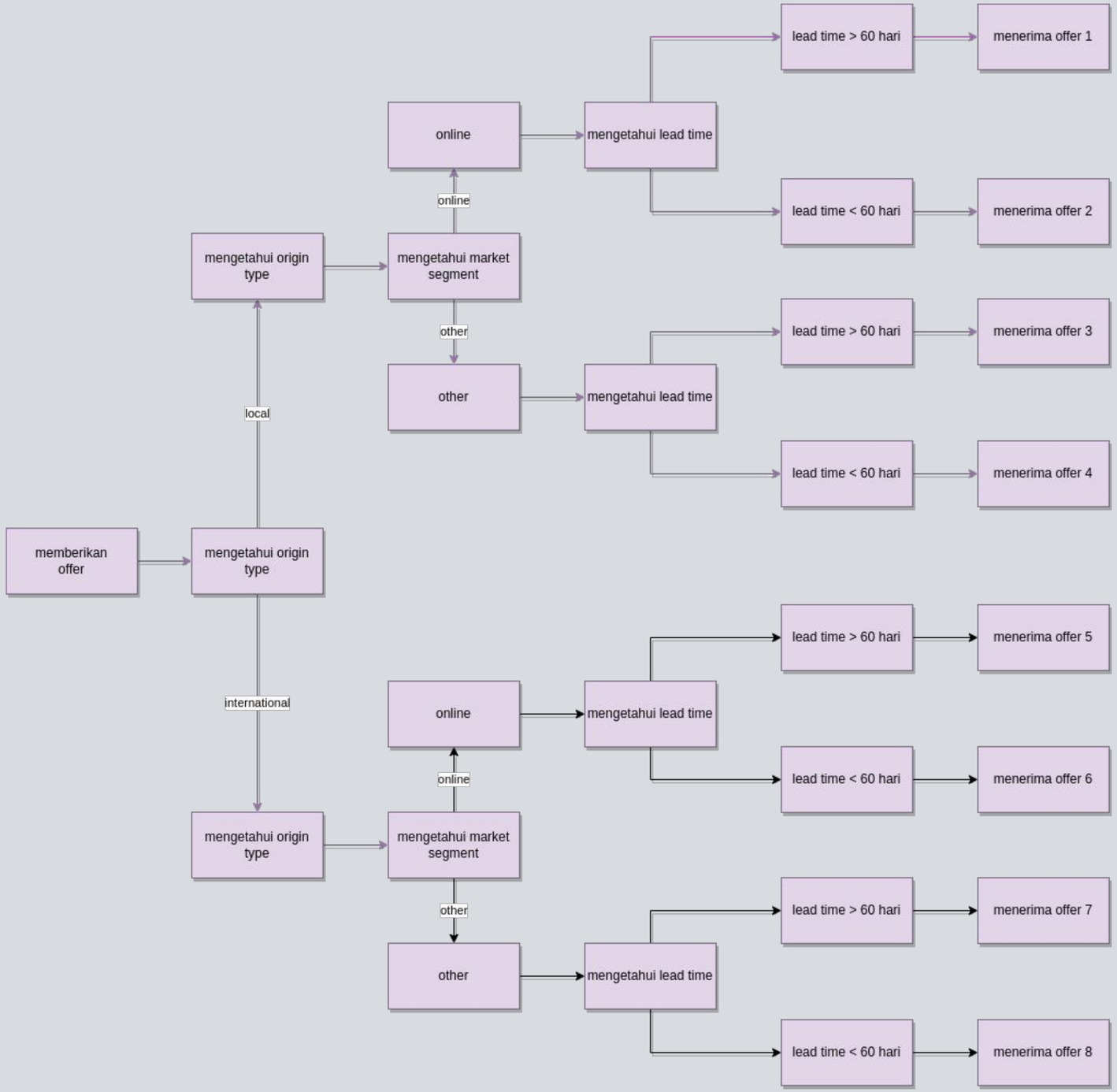
df1 - featurewiz	Hyperparameter Tuning	
	XGBoost	CatBoost
Accuracy (Test Set)	0.86	0.83
Precision (Test Set)	0.84	0.78
Recall (Test Set)	0.76	0.74
F1-Score (Test Set)	0.80	0.76
roc_auc (test-proba)	0.93	0.90
roc_auc (test-proba)	0.94	0.90
roc_auc (crossval train-mean)	0.94599	0.89911
roc_auc (crossval test-mean)	0.88311	0.87465
roc_auc (crossval train-std)	0.00170	0.00187
roc_auc (crossval test-std)	0.02651	0.00733

df2 - featurewiz	Hyperparameter Tuning	
	XGBoost	CatBoost
Accuracy (Test Set)	0.85	0.83
Precision (Test Set)	0.83	0.78
Recall (Test Set)	0.77	0.74
F1-Score (Test Set)	0.80	0.76
roc_auc (test-proba)	0.93	0.90
roc_auc (test-proba)	0.96	0.90
roc_auc (crossval train-mean)	0.95872	0.89911
roc_auc (crossval test-mean)	0.87105	0.87465
roc_auc (crossval train-std)	0.00196	0.00185
roc_auc (crossval test-std)	0.02628	0.00733

df3 - featurewiz	Hyperparameter Tuning	
	XGBoost	CatBoost
Accuracy (Test Set)	0.84	0.81
Precision (Test Set)	0.83	0.77
Recall (Test Set)	0.76	0.75
F1-Score (Test Set)	0.79	0.76
roc_auc (test-proba)	0.92	0.89
roc_auc (test-proba)	0.94	0.89
roc_auc (crossval train-mean)	0.94369	0.88968
roc_auc (crossval test-mean)	0.86489	0.87530
roc_auc (crossval train-std)	0.00099	0.00160
roc_auc (crossval test-std)	0.03353	0.01063







Kelompok Offer

[Kelompok offer selengkapnya dapat dilihat disini]

Thank you!

BlueCode