KALTEN
HOTEL

CAPSTONE PROJECT MODUL 3

HOTEL BOOKING DEMAND

Presentation by Imron Asofi | Data Science | 2023 | Purwadhika Digital Technology School

CONTENT

BUSINESS PROBLEM

Context

Problem Statement

Objective

Aproach

DATA UNDERSTANDING

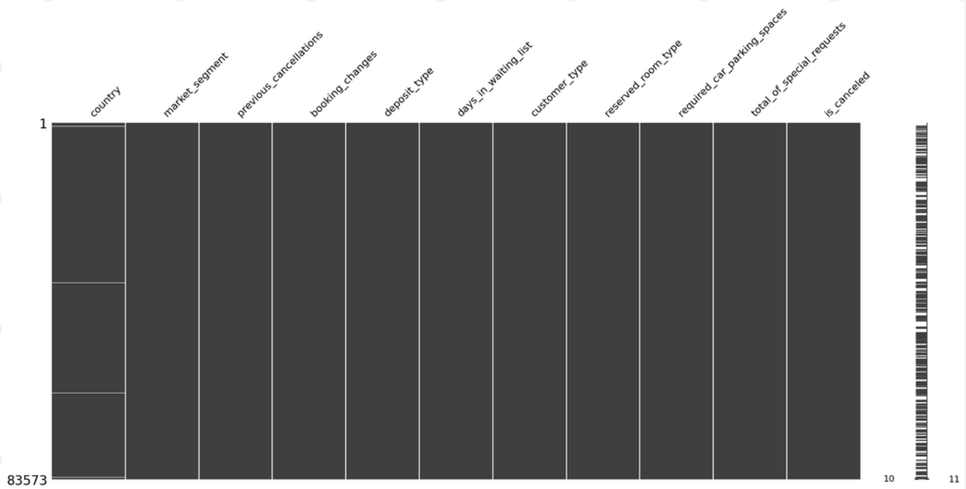
- Dataset menampilkan hotel booking dari 1 July 2015 - 31 Agustus 2017 dari 2 hotel di Portugal
- Pada dataset ini semua *value* pada kolom numerikal memiliki skala nilai yang mirip

Atribut	Deskripsi
country	Atribut ini mencatat negara asal pelanggan yang melakukan pemesanan hotel
market_segment	Atribut ini mengidentifikasi segmen pasar di mana pelanggan termasuk segmen pasar bisnis atau segmen pasar liburan. “TA” artinya “Travel Agents” & “TO” artinya “Tour Operators”
previous_cancellations	Atribut ini mencatat jumlah pembatalan sebelumnya yang dilakukan oleh pelanggan
booking_changes	Atribut ini mencatat jumlah perubahan yang dilakukan pada pemesanan, seperti perubahan tanggal atau jenis kamar
deposit_type	Atribut ini mengindikasikan jenis deposit yang dibayarkan oleh pelanggan, deposit dibagi menjadi 3 kategori: `No Deposit` (tidak melakukan deposit sama sekali), `Non Refund` (membuat deposit seharga total pembelian), `Refundable` (membuat deposit kurang dari total pembelian)
days_in_waiting_list	Atribut ini mencatat jumlah hari dalam daftar tunggu sebelum pemesanan dikonfirmasi
customer_type	Atribut ini menggambarkan jenis pelanggan (misalnya, transient, kontrak, grup) : `Contract` (Booking yang didasari oleh kontrak); `Group` (Booking dengan sistem grup); `Transient` (Booking untuk menginap dalam jangka waktu pendek & tidak termasuk dalam grup ataupun kontrak); `Transient-party` (Booking yang memiliki kelompok transient lainnya)
reserved_room_type	Atribut ini mencatat jenis kamar yang dipesan oleh pelanggan
required_car_parking_spaces	Atribut ini mencatat jumlah tempat parkir mobil yang diperlukan oleh pelanggan
total_of_special_requests	Menunjukkan total jumlah permintaan khusus yang dibuat oleh pelanggan (contoh: twin bed atau minta lantai atas)
is_canceled	Atribut target ini menandakan apakah pemesanan hotel dibatalkan atau tidak, (0 - tidak dibatalkan, 1 - dibatalkan)

DATA PREPROCESING

	dataFeatures	dataType	null	nullPct	unique	uniqueSample
0	country	object	351	0.42	162	[ZWE, GHA]
1	market_segment	object	0	0.00	8	[Corporate, Direct]
2	previous_cancellations	int64	0	0.00	15	[4, 13]
3	booking_changes	int64	0	0.00	19	[9, 16]
4	deposit_type	object	0	0.00	3	[No Deposit, Non Refund]
5	days_in_waiting_list	int64	0	0.00	115	[18, 236]
6	customer_type	object	0	0.00	4	[Contract, Transient-Party]
7	reserved_room_type	object	0	0.00	10	[D, F]
8	required_car_parking_spaces	int64	0	0.00	5	[3, 2]
9	total_of_special_requests	int64	0	0.00	6	[1, 2]
10	is_canceled	int64	0	0.00	2	[1, 0]

Pada mising value yang terdapat di kolom country disini dilakukan pengisian dengan 'Unknown'



DATA PREPROCESING

Binning

```
category_dict = {
    'Asia': ['CN', 'JPN', 'IND', 'SGP'],
    'Europe': ['FRA', 'PRT', 'NLD', 'ESP', 'LUX', 'DEU', 'ITA', 'CHE', 'GBR', 'SRB', 'POL', 'SWE', 'AUT', 'CZE',
               'RUS', 'ROU', 'DNK', 'NOR', 'FIN', 'UKR', 'HUN', 'EST', 'SVN', 'LTU', 'LVA', 'ALB', 'HRV', 'LVA',
               'MKD', 'GEO', 'BLR', 'MLT', 'GRC', 'CYP', 'MLT', 'PAN', 'GIB', 'AND'],
    'North America': ['USA', 'CAN', 'MEX'],
    'South America': ['BRA', 'COL', 'ARG', 'CHL', 'URY', 'PER'],
    'Others': ['AUS', 'ZAF', 'TUR', 'IDN', 'KOR', 'THA', 'SAU', 'EGY', 'Unknown']
}
```

```
continent
Europe           71398
Others            7658
South America     1841
North America     1529
Asia              1146
Name: count, dtype: int64
```

Duplicated

Pada dataset ini terdapat 73371 data duplicate

MODELING

ENCODING

'market_segment',
'deposit_type',
'customer_type',
'continent'

SCALING

'previous_cancellations',
'booking_changes',
'days_in_waiting_list',
'required_car_parking_spaces',
'total_of_special_requests'

BINARY

'reserved_room_type'

MODELING

BASE MODEL

Logistic Regression
K-Nearist Neighbors
Decision Tree Classifier

MODEL ALGORITHM

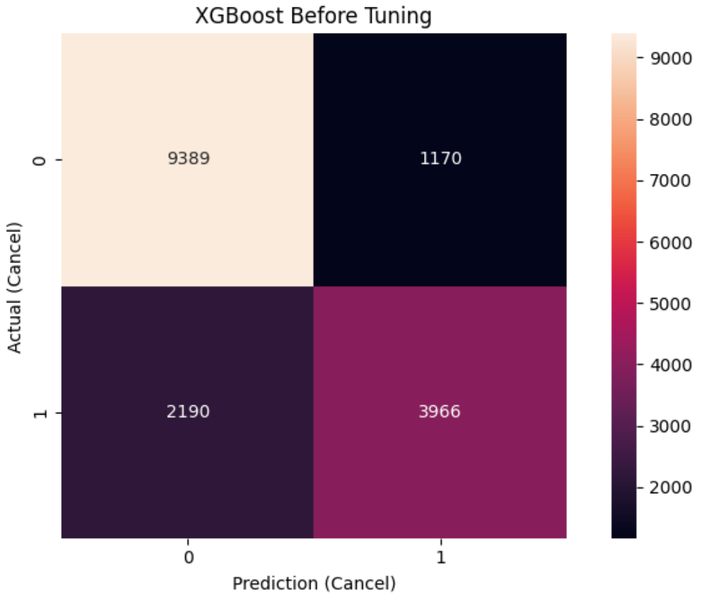
XGBClassifier
LGBMClassifier
CatBoostClassifier
RandomForestClassifier
GradientBoostingClassifier
BaggingClassifier
AdaBoostClassifier

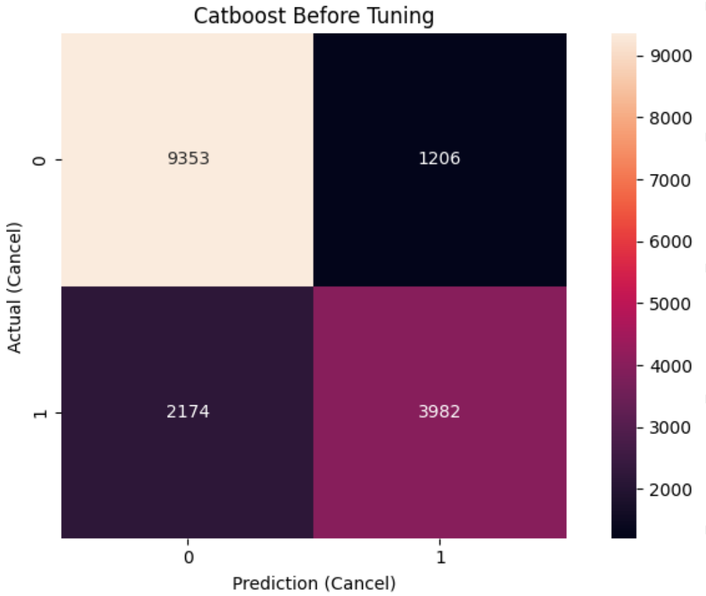
RESULT

	model	resample	fit_time	score_time	accuracy	precision	recall	f1
15	xgb	ros	1.418720	0.059105	0.766635	0.764307	0.650136	0.702612
16	cat	ros	21.896415	0.106110	0.766454	0.762635	0.651030	0.702421
35	xgb	smote	1.821457	0.056010	0.766301	0.761336	0.652288	0.702302
36	cat	smote	18.931268	0.107609	0.766440	0.771353	0.644206	0.702060
17	lgbm	ros	0.415263	0.060745	0.766302	0.766532	0.647577	0.702049
7	lgbm	none	0.327244	0.064264	0.765982	0.757135	0.654726	0.701910
25	xgb	rus	1.159836	0.062458	0.765787	0.754468	0.656634	0.701860
37	lgbm	smote	0.723131	0.057908	0.766266	0.770110	0.644734	0.701852
33	rf	smote	4.188086	0.213543	0.765678	0.751388	0.658949	0.701848
26	cat	rus	16.131141	0.115548	0.765561	0.752568	0.657649	0.701602

Hasil model 2 terbaik dihasilkan oleh XGBoost dan Catboost, kemudian data akan di oversampling untuk mencari hasil prediksi yang lebih baik

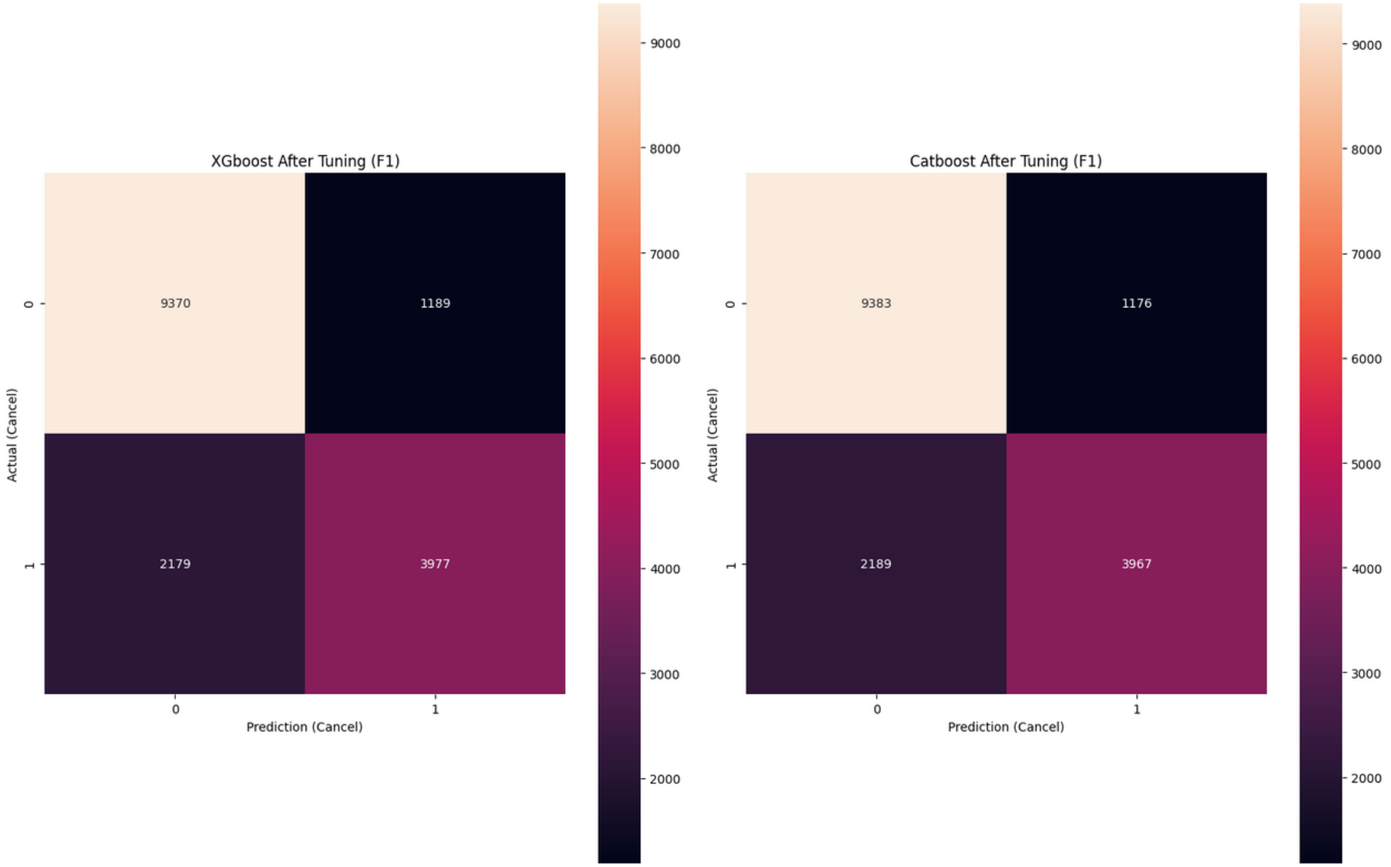
RESULT





	Model	Accuracy	Precision	Recall	F1
0	XGboost ROS Before Tuning	0.798983	0.772196	0.644250	0.702444
1	Catboost ROS Before Tuning	0.797786	0.767540	0.646849	0.702045

RESULT



	Model	Accuracy	Precision	Recall	F1
0	xgb_tuning_f1	0.798504	0.769841	0.646036	0.702526
1	cat_tuning_f1	0.798684	0.771340	0.644412	0.702186

RESULT

```
XGBoost - F1 Best Parameters: {'model__learning_rate': 0.1, 'model__max_depth': 5, 'model__n_estimators': 300}
XGBoost - F1 Best Score: 0.7027607154921389

Catboost - F1 Best Parameters: {'model__depth': 5, 'model__iterations': 300, 'model__learning_rate': 0.1}
Catboost - F1 Best Score: 0.7026586076082937
```

XGB Classifier

F1 Score default : 0.702444

F1 Best Score: 0.702760

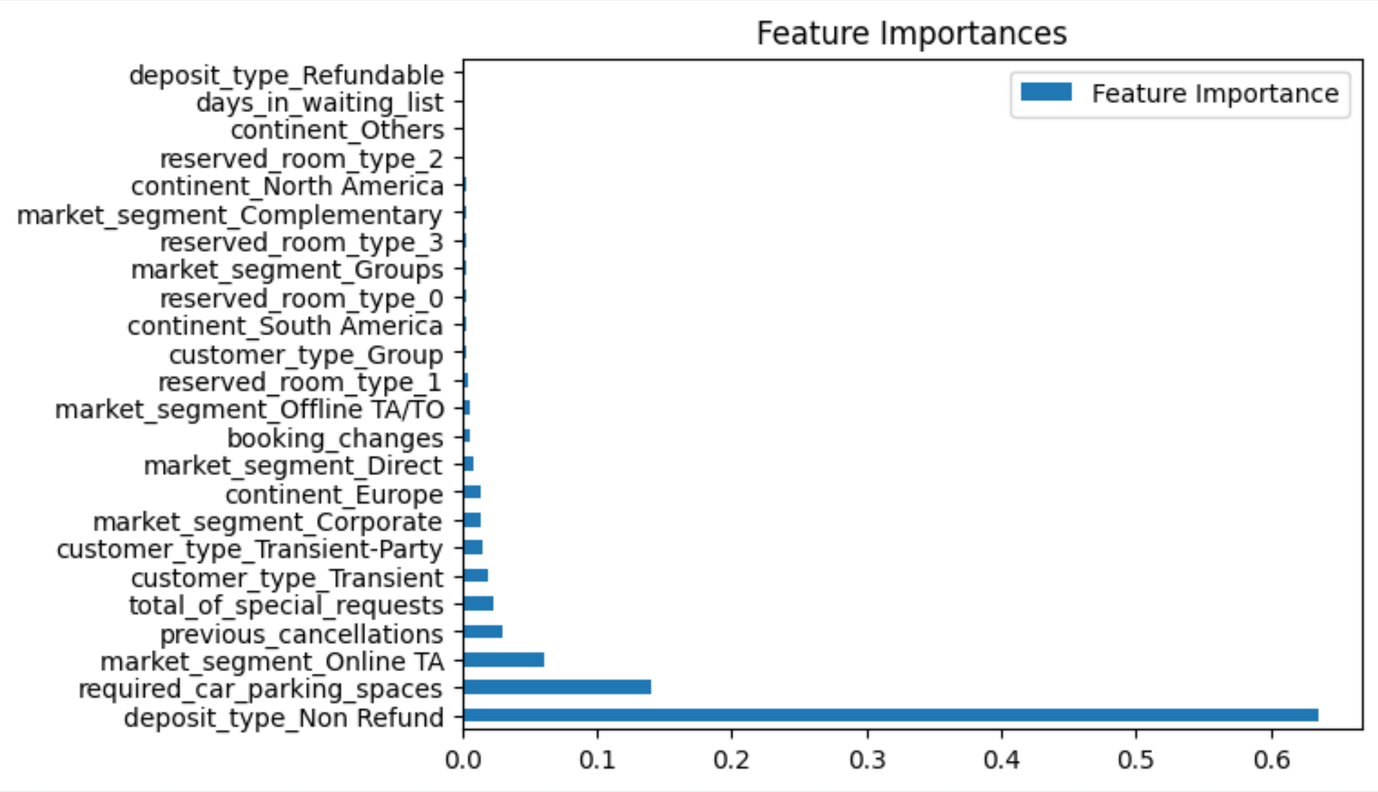
Best parameter

'model__n_estimators': 300

'model__max_depth': 5

'model__learning_rate': 0.1

RESULT


Feature Importances
Feature Importance
deposit_type_Refundable
days_in_waiting_list
continent_Others
reserved_room_type_2
continent_North America
market_segment_Complementary
reserved_room_type_3
market_segment_Groups
reserved_room_type_0
continent_South America
customer_type_Group
reserved_room_type_1
market_segment_Offline TA/TO
booking_changes
market_segment_Direct
continent_Europe
market_segment_Corporate
customer_type_Transient-Party
customer_type_Transient
total_of_special_requests
previous_cancellations
market_segment_Online TA
required_car_parking_spaces
deposit_type_Non Refund
0.0
0.1
0.2
0.3
0.4
0.5
0.6

CONCLUSION

Classification Report Tuned XGBoost :

	precision	recall	f1-score	support
0	0.81	0.89	0.85	10559
1	0.77	0.65	0.70	6156
accuracy			0.80	16715
macro avg	0.79	0.77	0.78	16715
weighted avg	0.80	0.80	0.79	16715

Dengan adanya analisa ini, kita dapat memprediksi pelanggan mana yang sekiranya akan booking cancel ataupun tidak. Pelaku industri perhotelan memiliki lebih banyak tools yang bisa digunakan dalam menghadapi hal ini.

RECOMMENDATIONS

1. **Penambahan Data. Dalam rangka meningkatkan performa model ini, langkah yang paling efektif adalah menambah jumlah data dan juga keberagaman data. Kebanyakan data numerikal pada dataset ini, memiliki nilai 0,**
2. **Membuat kontrak atau peraturan dengan Travel Agent baik yang offline maupun online, memungkinkan untuk merendahkan angka cancel booking,**
3. **Dengan adanya tools prediksi ini, pelaku industri perhotelan dapat mencegah cancel booking dengan memberikan diskon atau layanan lebih kepada pelanggan calon cancel booking**
4. **Dapat menyesuaikan ketentuan booking sesuai dari feature - feature prediksi cancel booking**

THANK YOU

Presentation by Imron Asofi