Welcome

HOTEL BOOKING PREDICTION

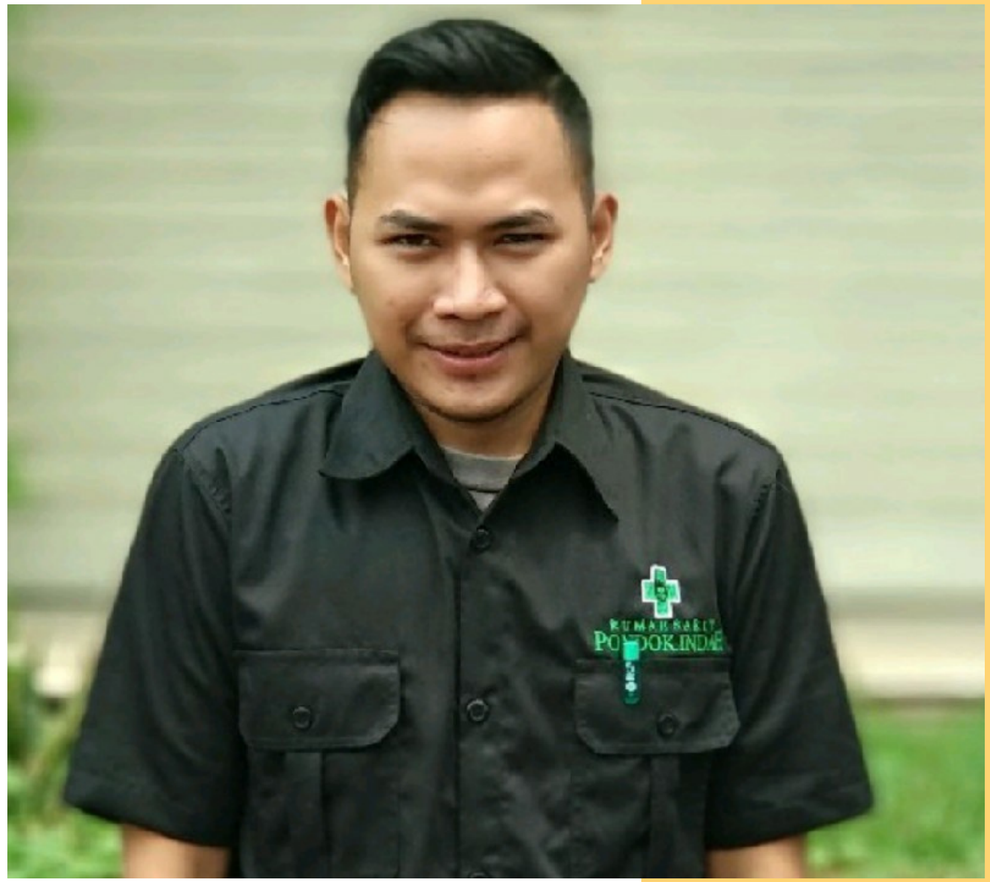
By Mahmud Riadhi

Introducing

ABOUT ME

I am a,

- CURRENTLY | Asst. Car Park Manager at PT. Securindo Packatama Ind.
- CURRENTLY | Graduated from a Data Science bootcamp at Dibimbing



Mahmud Riadhi

Data Enthusiast

OUTLINE

01

BUSINESS BACKGROUND AND OBJECTIVE

02

DATA PREPARATION AND FEATURE ENGINEERING

03

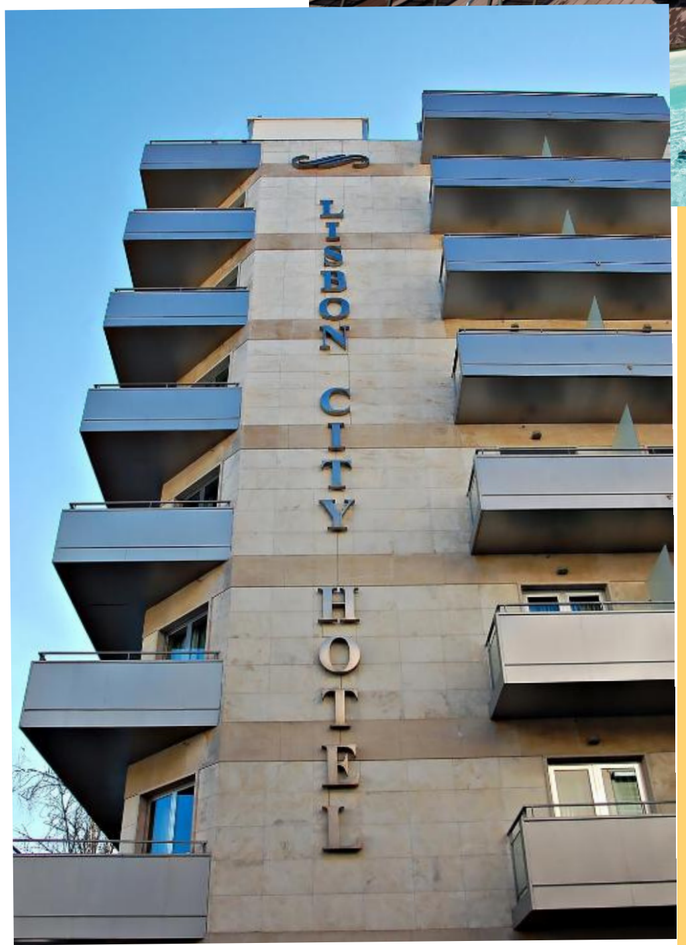
MODELLING AND EVALUATION

04

CONCLUSION AND RECOMMENDATION

01

BUSINESS BACKGROUND AND OBJECTIVE



INTRODUCTION AND PROBLEMS

In the midst of the hectic nature of managing a property, bookings can be overwhelming. Many bookings are absolutely unavoidable as guest plans change, hotels have better models to predict whether guests will actually come or not. This can help the hotel to plan things like personal needs and meals. Or provide additional facilities on the type of room booked so as to help the hotel proactively avoid this situation and generate revenue for the hotel.

OBJECTIVES

- Provides a variety of analysis and descriptive insights
- Knowing what factors cause booking cancellations and reducing losses for guests who cancel bookings
- Develop a model for the prediction of order cancellations and describe the model and its outputs

02

DATA PREPARATION AND FEATURE ENGINEERING

Dataset Information

Data Cleansing

Data Pre-Processing and Analysis

DATASET INFORMATION

Numerical

- lead_time
- arrival_date_year
- arrival_date_week_number
- arrival_date_day_of_month
- stays_in_weekend_nights
- stays_in_week_nights
- adults
- babies
- is_repeated_guest
- previous_cancellations
- previous_bookings_not_canceled
- booking_changes
- agent
- company
- days_in_waiting_list
- adr
- required_car_parking_spaces
- total_of_special_requests

Categorical

- hotel
- arrival_date_month
- meal
- country
- market_segment
- distribution_channel
- reserved_room_type
- assigned_room_type
- deposit_type
- customer_type
- reservation_status
- reservation_status_date

119390 rows
31 features

1 target

is_canceled

DATA CLEANSING - MISSING VALUES

	Value	Value %
hotel	0	0.000000
is_canceled	0	0.000000
lead_time	0	0.000000
arrival_date_year	0	0.000000
arrival_date_month	0	0.000000
arrival_date_week_number	0	0.000000
arrival_date_day_of_month	0	0.000000
stays_in_weekend_nights	0	0.000000
stays_in_week_nights	0	0.000000
adults	0	0.000000
children	4	0.003350
babies	0	0.000000
meal	0	0.000000
country	488	0.408744
market_segment	0	0.000000
distribution_channel	0	0.000000
is_repeated_guest	0	0.000000

	Value	Value %
previous_cancellations	0	0.000000
previous_bookings_not_canceled	0	0.000000
reserved_room_type	0	0.000000
assigned_room_type	0	0.000000
booking_changes	0	0.000000
deposit_type	0	0.000000
agent	16340	13.686.238
company	112593	94.306.893
days_in_waiting_list	0	0.000000
customer_type	0	0.000000
adr	0	0.000000
required_car_parking_spaces	0	0.000000
total_of_special_requests	0	0.000000
reservation_status	0	0.000000
reservation_status_date	0	0.000000

MISSING VALUES HANDLING

	Value	Value %
children	4	0.003350
country	488	0.408744
agent	16340	13.686.238
company	112593	94.306.893

children
country
→ Simple drop, because the missing values are small from the proportion (<5%)

agent
→ Imputed by '0', because most likely without using agency

company
→ Drop column, because missing values are more than proportion (>50%)

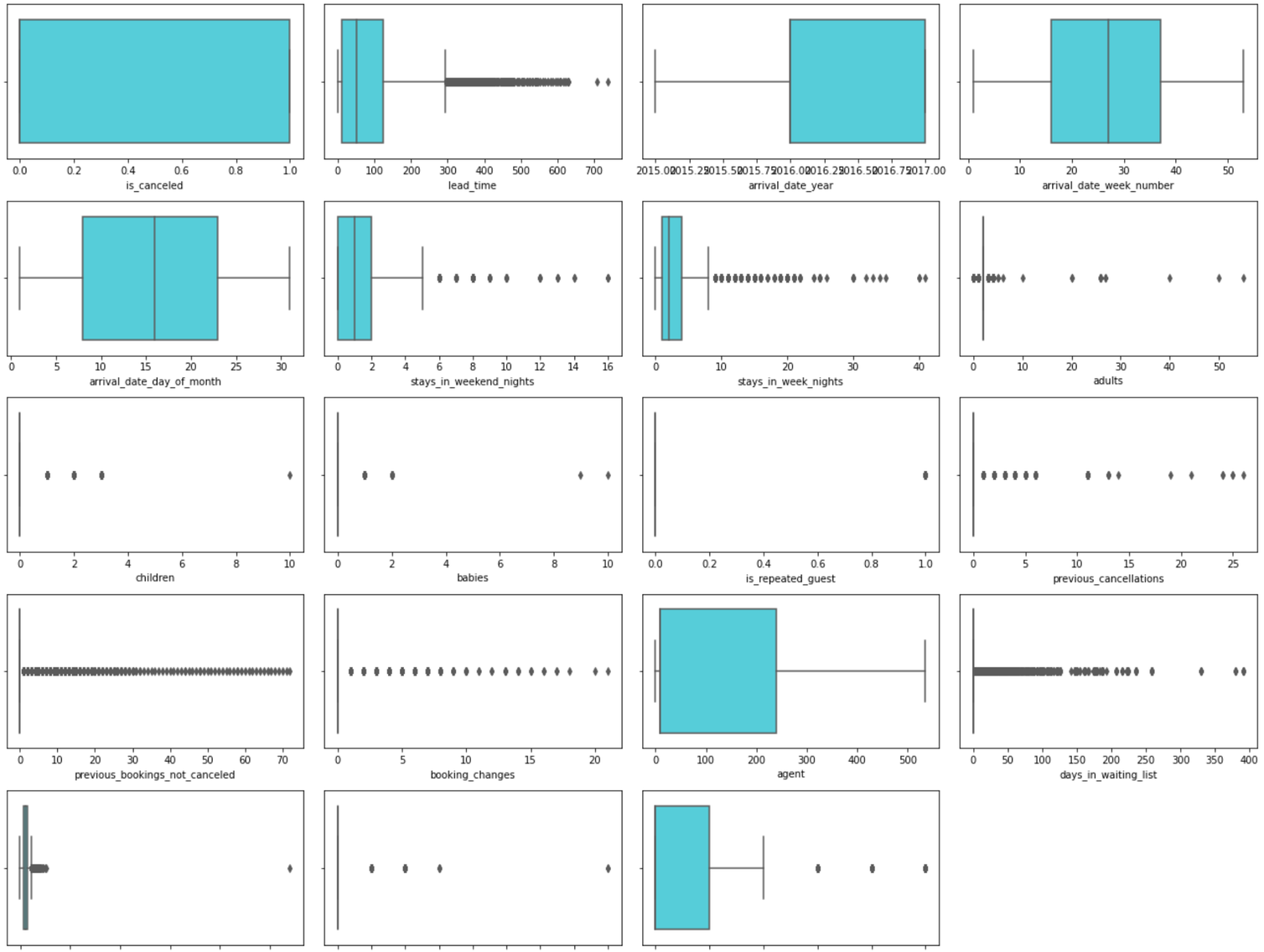
DUPLICATED VALUES HANDLING

31965 → Remove duplicate data

```
Data Frame Dimension Before Duplicate Removal: (118898, 31)
Data Frame Dimension After Duplicate Removal: (86933, 31)
```

DATA PRE-PROCESSING AND ANALYSIS

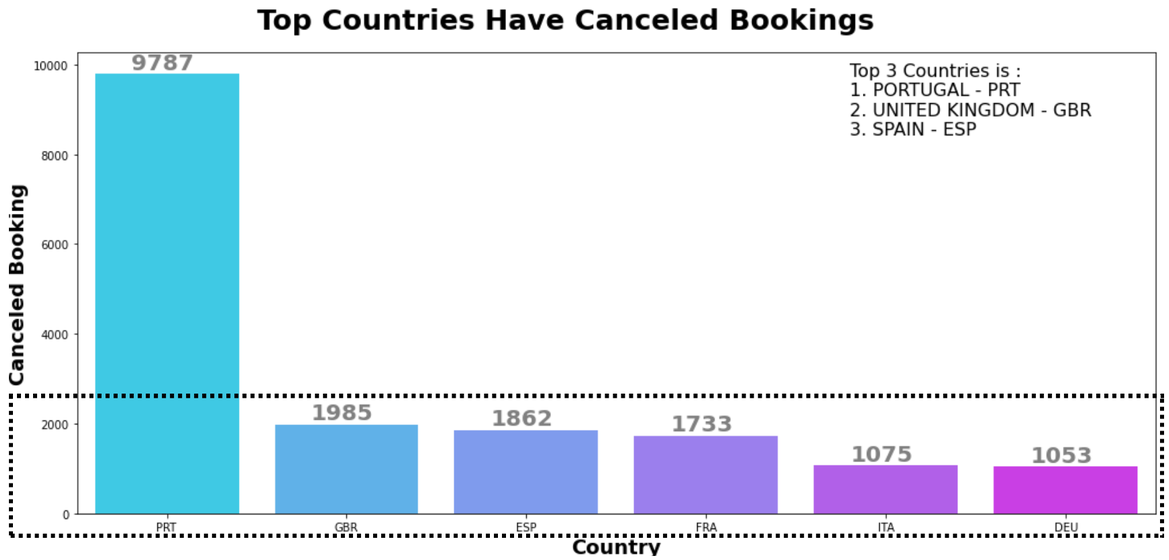
Boxplot Numerical Data

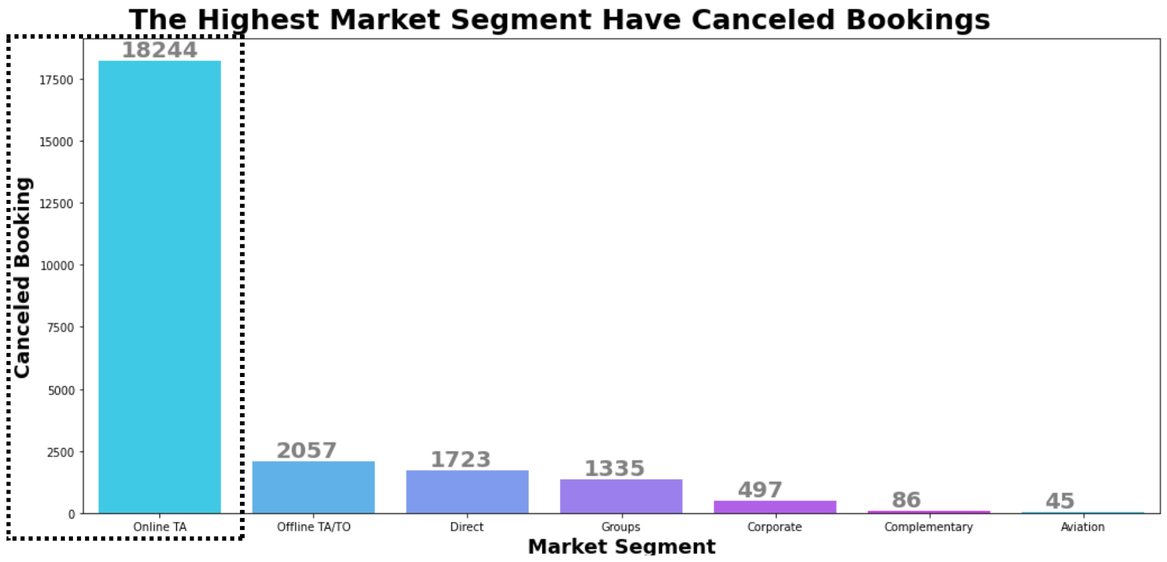


Many columns have multiple outliers. However, the outliers are still "normal" (not too extreme). That is. no need to handle it.

DATA PRE-PROCESSING AND ANALYSIS

Deep Dive Exploration





Insight

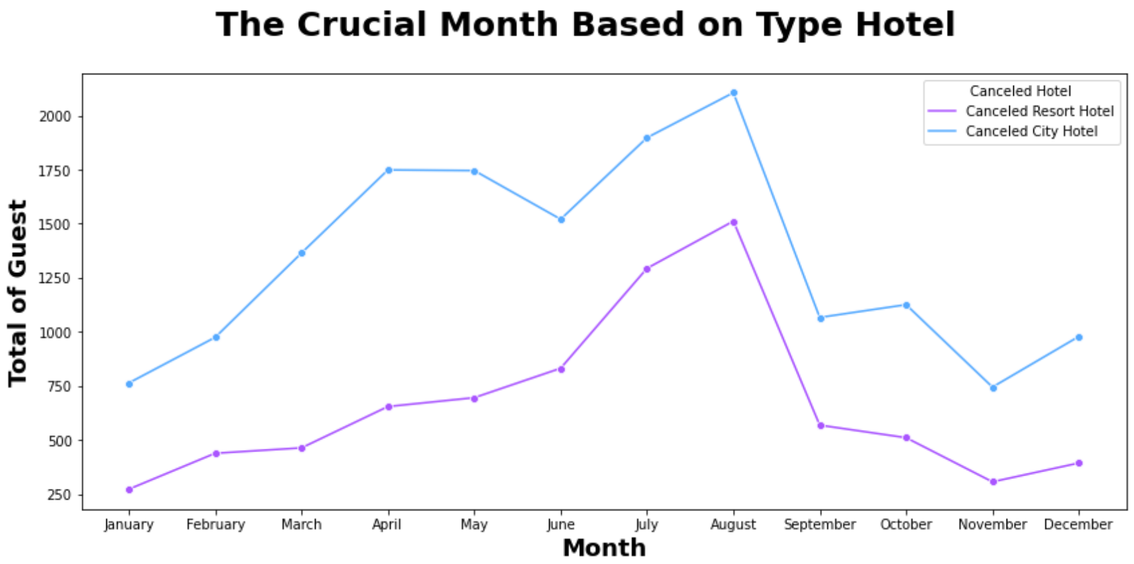
Most of the guests who canceled the booking were from the countries of origin of the two hotels which were located in Portugal and the bookings were made online by these guests. Likewise with other countries such as United Condoms, Spain, France and other countries, we can see that not a few have canceled bookings at the two hotels.

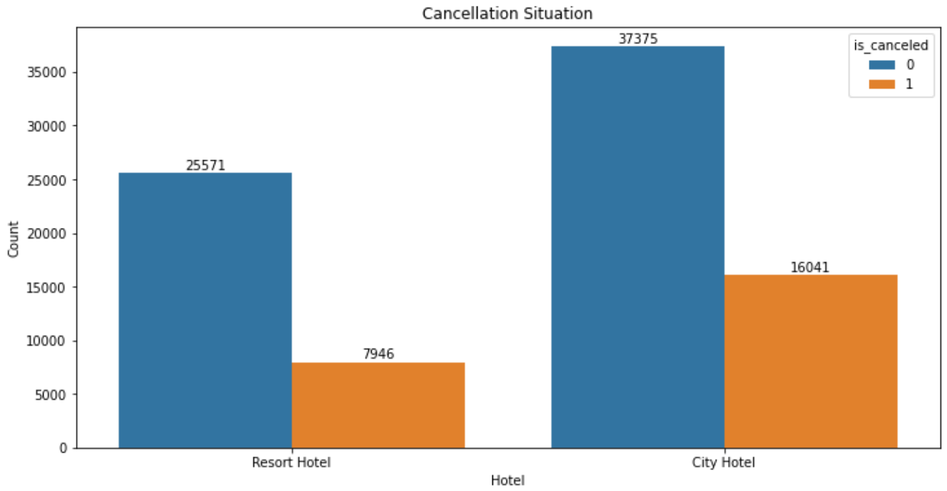
Recommendation

Need to improve the marketing strategy and promote the hotel more, especially on social media because most customers make online bookings. So that other countries can see the promotions that are being held at the hotel.

DATA PRE-PROCESSING AND ANALYSIS

Deep Dive Exploration





Insight

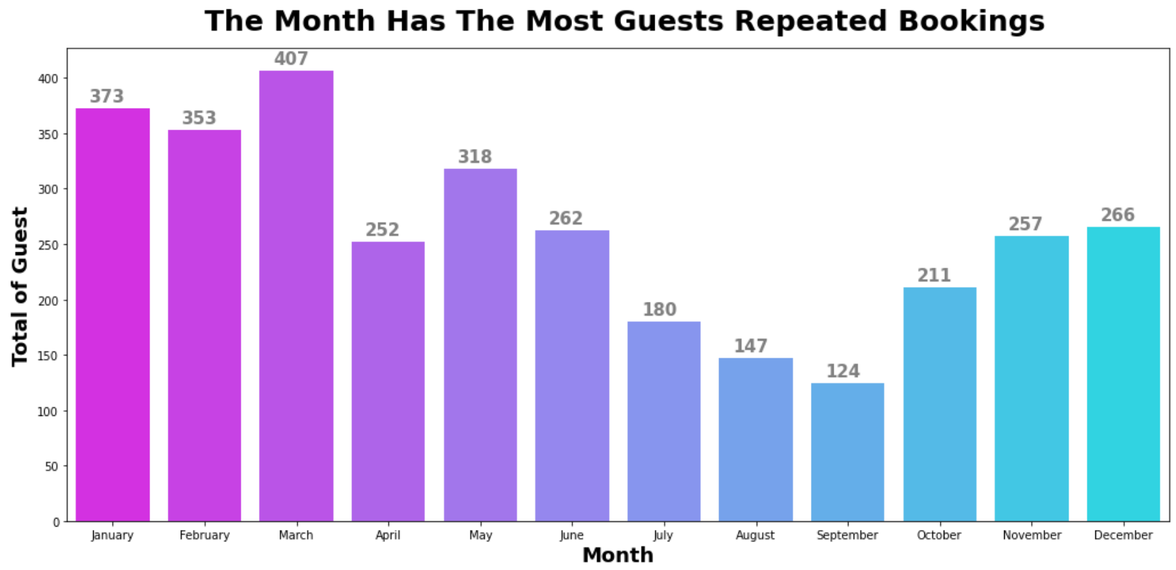
Every first month until the middle of the year, the two hotels always experience an increase for guests who cancel bookings and the highest is in August. However, after August until the end of the year there was a very significant decline for the guests.

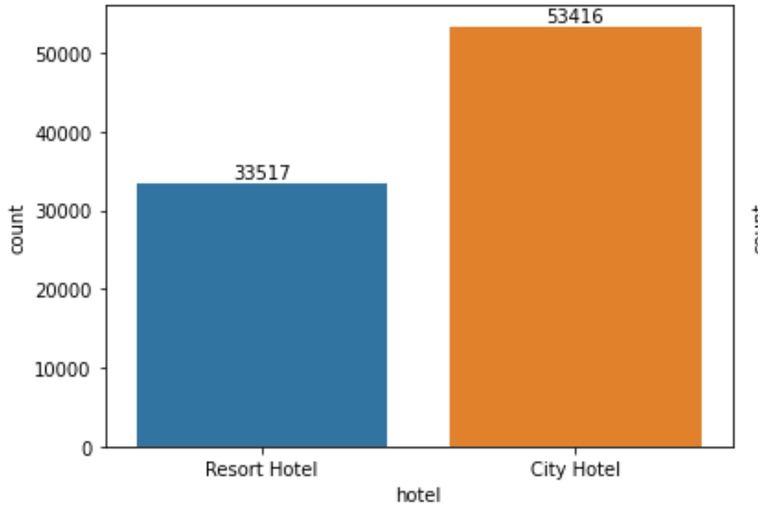
Recommendation

In June - August, the hotel can lower room rates to avoid many cancellations during the month. So that it can replace lost revenue in the previous month and the hotel can at the same time increase the percentage of bookings that are not canceled.

DATA PRE-PROCESSING AND ANALYSIS

Deep Dive Exploration





Insight

Based on the distribution in addition, hotel bookings made to repeat guests are not so much compared to the distribution of bookings at each hotel. This means that most guests visit the hotel only once.

Recommendation

It needs to be targeted in such a way, to guests who have used the hotel by providing special benefits if the guest returns to stay again

DATA PRE-PROCESSING AND ANALYSIS

Feature Engineering

Label Encoding

Features	0	1
Hotel	Resort Hotel	City Hotel

Ordinal Encoding

Features	1	2	3	4
deposit_type	No Deposit	Non Refund	Refundable	-
reservation_status	No-Show	Canceled	Check-Out	-
meal	SC	BB	HB	FB
customer_type	Transient	Transient-Party	Contract	Group

DATA PRE-PROCESSING AND ANALYSIS

Feature Engineering

Frequency Encoding

Before Encoding	After Encoding
market_segment	%market_segment
distribution_channel	%distribution_channel
reserved_room_type	%reserved_room_type
assigned_room_type	%assigned_room_type

Datetime Feature

Before Transform	After Transform
reservation_ status_date	year
	month
	day

03

MODELLING AND EVALUATION

MODELLING PROCESS

Baseline Model

BASELINE MODEL

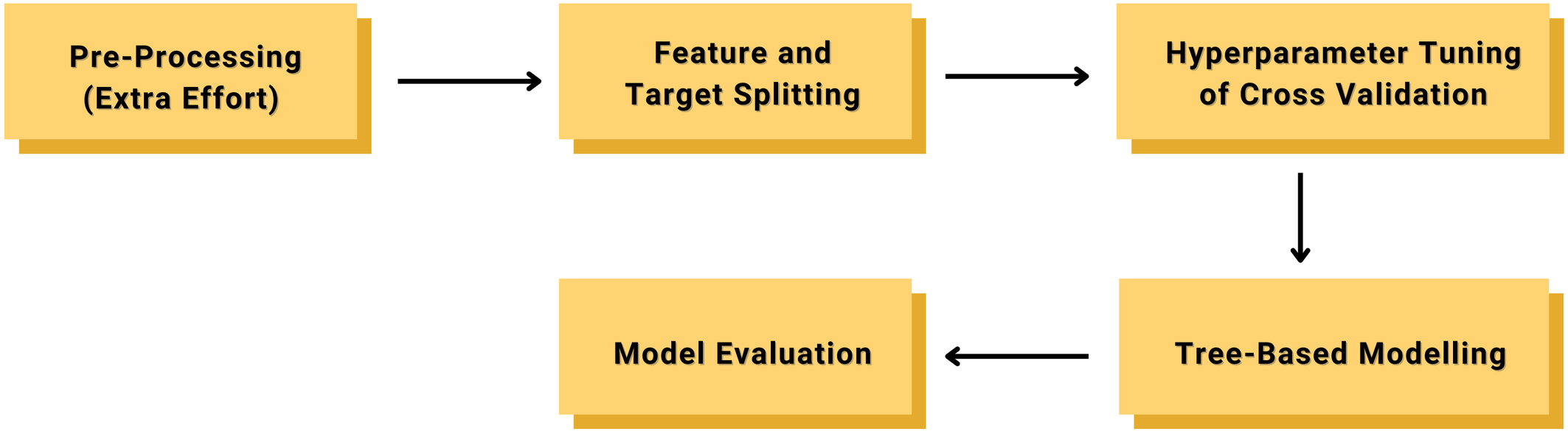
Result Model

MODEL	ACCURACY	AUC	PRECISION	RECALL	F1-SCORE
K-Near Neighbords	0.71	0.60	0.48	0.36	0.41
Logistic Regression	0.73	0.54	0.55	0.12	0.20
Decision Tree	0.75	0.69	0.56	0.56	0.56
Random Forest	0.81	0.72	0.72	0.53	0.61

- The dataset has been pre-processed (low effort) and divided into training data (80%), test data (20%)
- Random forest has the highest accuracy from other models

MODELLING PROCESS

Model Improvement



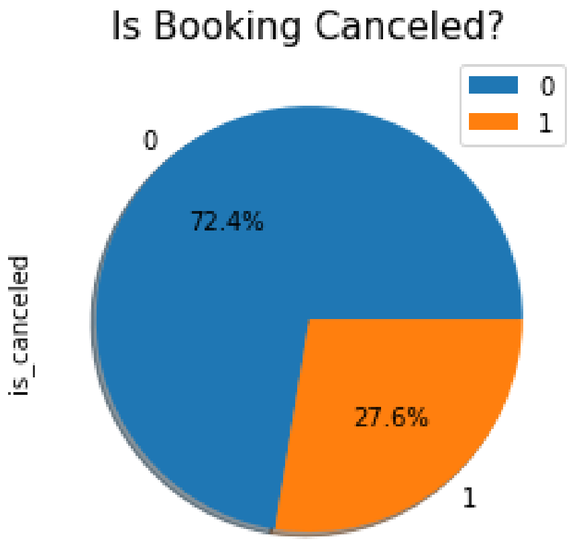
MODELLING PROCESS

Feature and Target Splitting

Data (dfhotel)

Features (All columns except 'is_canceled') → **X_train, X_test**

Target ('is_canceled') → **y_train, y_test**



Imbalanced data, it means not valid using ACCURACY. But, using AUC or F1-Score

MODEL IMPROVEMENT

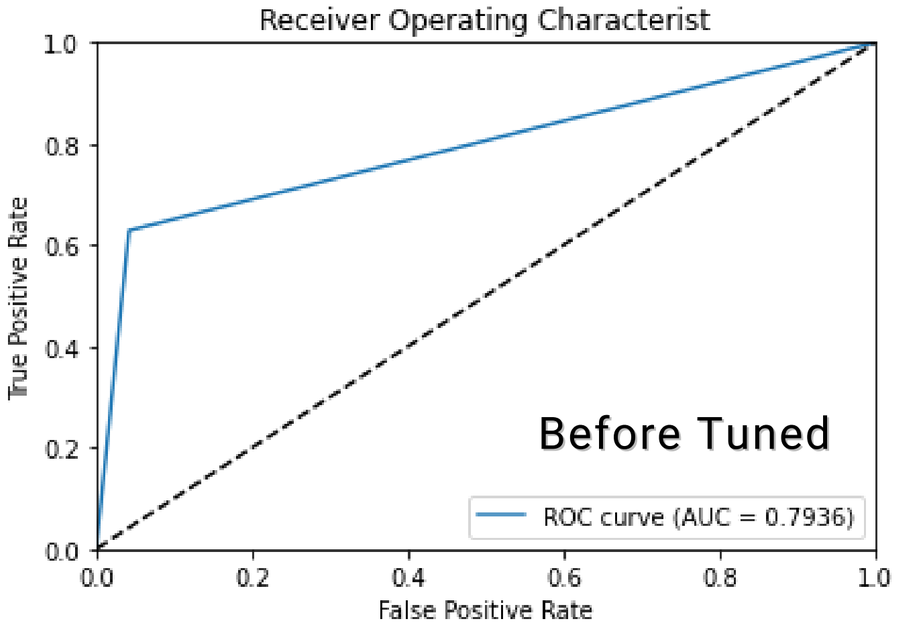
Result Model

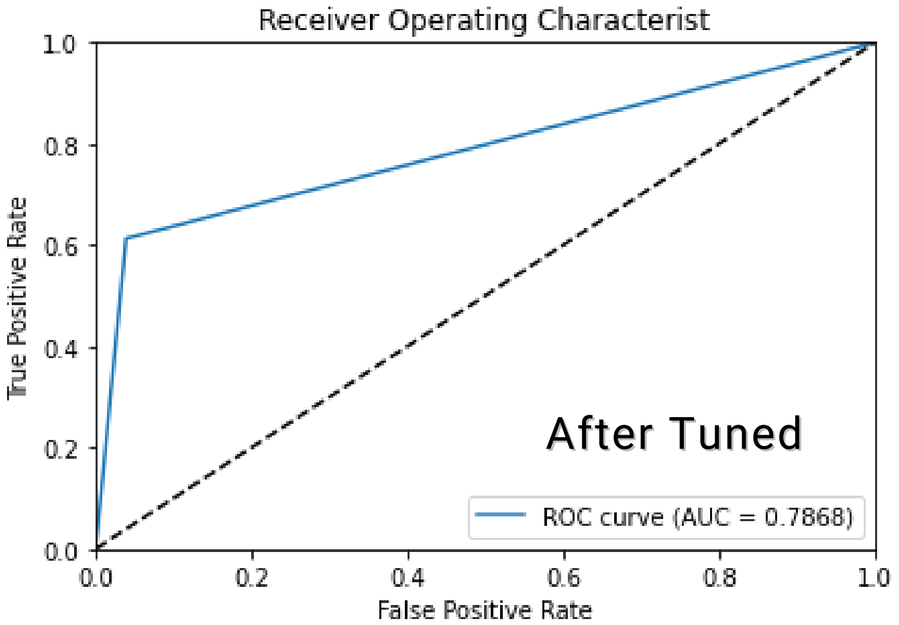
MODEL	ACCURACY	AUC	PRECISION	RECALL	F1-SCORE
K-Near Neighbords	0.81	0.73	0.70	0.54	0.61
Logistic Regression	0.80	0.68	0.73	0.43	0.54
Decision Tree	0.82	0.77	0.66	0.67	0.66
Random Forest	0.87	0.79	0.85	0.63	0.72

- The dataset has been pre-processed (extra effort) and divided into training data (80%), test data (20%)
- Random forest still has the highest AUC and F1-score of the other models

HYPERPARAMETER TUNING OF CROSS VALIDATION

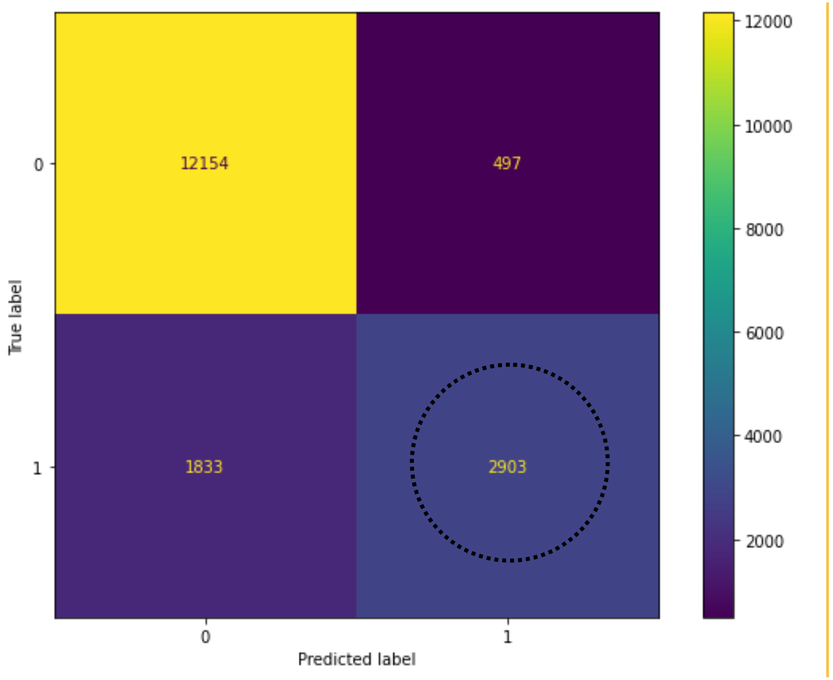
RANDOM FOREST	ACCURACY	AUC	PRECISION	RECALL	F1-SCORE
Before Tuned	0.87	0.79	0.85	0.63	0.72
After Tuned	0.87	0.79	0.85	0.61	0.71



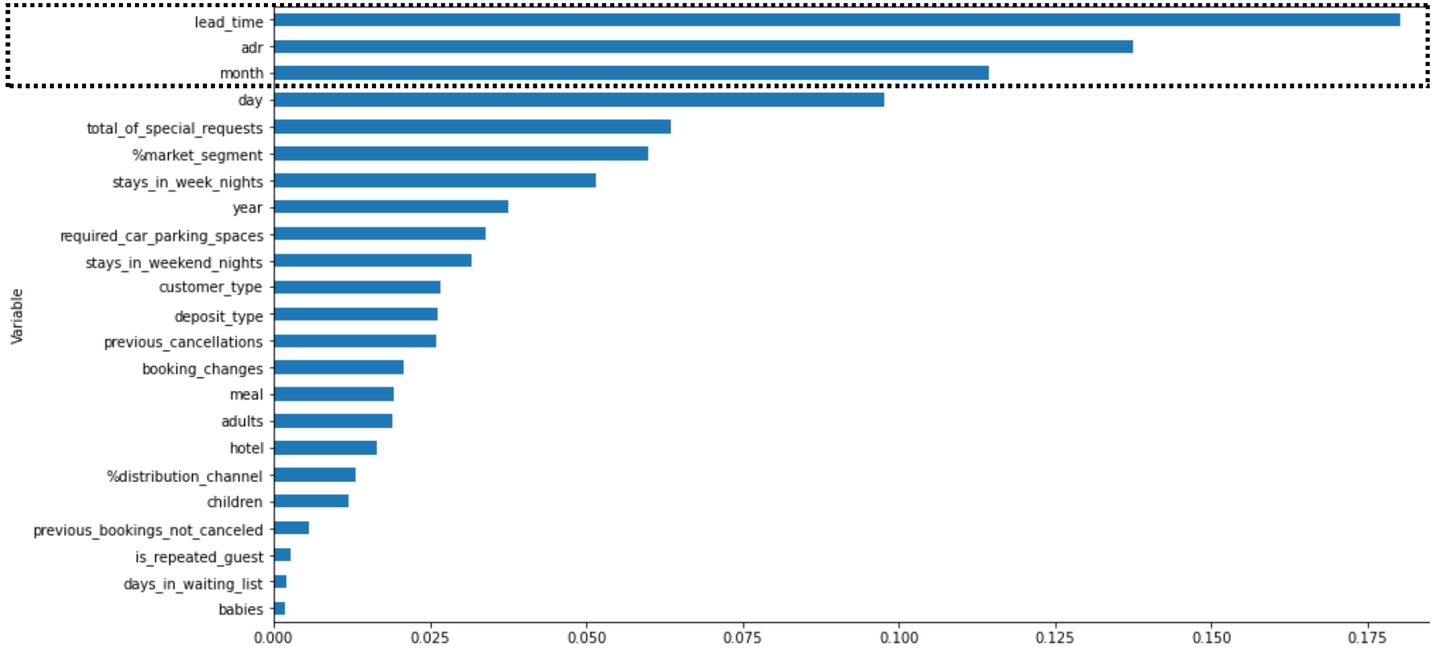


CONFUSION MATRIX AND FEATURES IMPORTANCE

Confusion Matrix



Features Importance



- True Positif in confusion matric : Model predict guest **canceled booking** and acctually they **canceled booking**
- **lead_time, adr and month** are the highest three factors that affect **canceled booking**

ESTIMATED POTENTIAL IMPACT

- City Hotel, average price per night 90 [EUR]
- Resort Hotel, average price per night 70 [EUR]

Model performance (AUC) 79%

- There are 100 guests booked at the Resort Hotel. However, canceling the order.
- The model can predict correctly as many as 79 guests canceled orders.
- Hotels, can strategize for 79 predicted guests cancelled bookings

So, with the model can reduce losses by ;

- City Hotel, 7110 [EUR] (90 * 79)
- Resort Hotel, 5530 [EUR] (70 * 79)

RECOMMENDATION

- Offer non-refundable and flexible rates
- Offer packages to be agreed upon so that there is no sudden cancellation of the order. Packages that link the cost of room, airfare and food
- Make a policy with a penalty if it is not appropriate to cancel a booking, but besides that the hotel must inform the guest in advance about the policy

CONCLUSION

- Random forest model can increase hotel income
- Target guests who will cancel hotel reservations so they can minimize booking cancellations for those guests

THANKS

Source : https://www.sciencedirect.com/science/article/pii/S2352340918315191
Script : https://colab.research.google.com/drive/1VN9jSWmPT8jXDgfFcEt0Sq7LYF7lfbns?usp=sharing
Github : https://github.com/MahmudRiadhi