Predicting Hotel Booking Cancellation in Portugal With Machine Learning

ABOUT THE PROJECT

Predicting Hotel Booking Cancellation in Portugal Project. Is a machine learning classification project that will try to predict whether a booking will be cancelled or a booking will not be cancelled using machine learning based on historical data.

The data for this project is from Hotel Booking Demand Dataset Sciencedirect. This data was acquired by extraction from hotel's Property management system from 2015 to 2017 from hotel in Region Algarve and Lisbon

Background Information

Hotel industry is one of the faster growing businesses of tourism sector, especially with the rise of giant OTA that make booking a hotel as easy as is ever been.

According to Portugal's National Institute of Statistic in 2017 hotel revenue rose approximately 18% to $3.6 billion

Growth of Hotel Industry in Portugal

$3.6 Billion

Hotel revenues in 2017 Source : link

20.6 million of Total Guest in 2017

(Portugal Population in 2017 : 10.31 million)

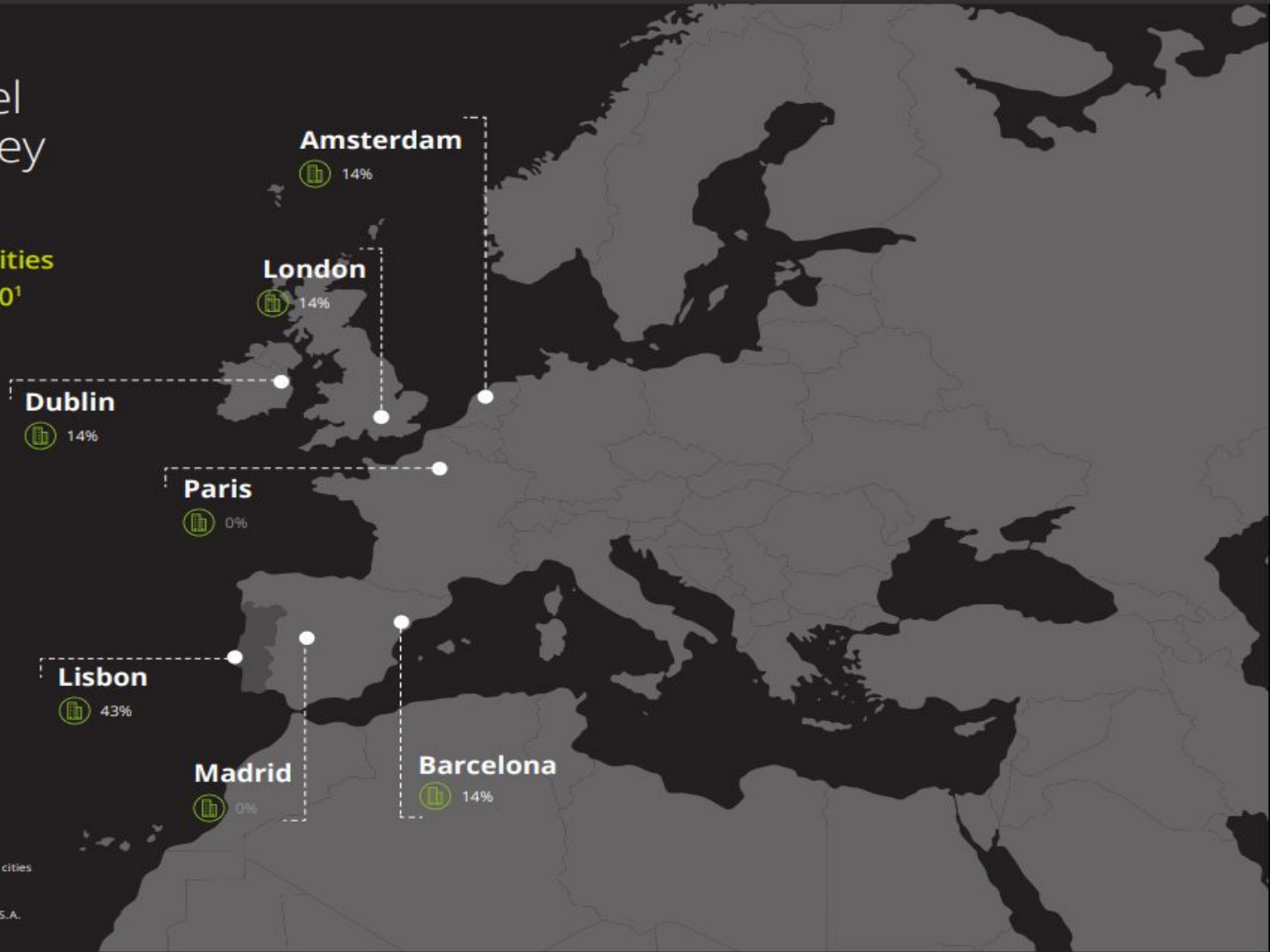
source : skift

Nominated as the most attractive European Cities (Lisbon) for hotel investment in 2020

source : Deloitte Hospitality Atlas 2019

Portuguese Hotel Investment Survey

Most attractive european cities for hotel investment in 2020[1]



Label:

Hotel investment

[1] Question with possibility of multiple answer.
19% of the participants also identified as more attractive other cities than those above mentioned.

Background Information
Problem

the growing trend of Hotel industry is beneficial for hotels however it comes with it's problem too.

One of the problem is the rising rate of cancellation in the hotel industry

Cancellation rate rose from under 33% in 2014 to 40% in 2018

source : Phocuswire

Problem Statement

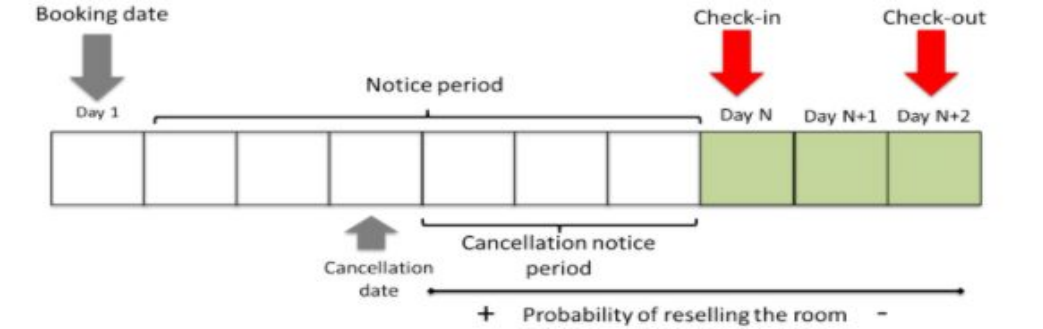
With the increase trend of cancellation from year to year, some hotel have think that high cancellation in hotel is the new norm of the industry which is a completely wrong approach.

One out of four hotel guests are cancelling hotel booking ahead of a stay. This cancellation trend has effect the hotel not being able to accurately forecast occupancy within their revenue management.

This trend of cancellation also have causes hotel loss in opportunity cost (unsold room due to cancellation)

How's Cancellation Affecting Hotel

1. Loss of income in shape of unsold room (due to cancellation)



2. Lower RevPAR (Revenue Per Available Room) when selling cheaper at the last minute

Cancellation that's close to checkin leave a very little spot for hotel to maneuver and resell the room. on many occasion no alternative way but to lower the room price

Project Goals

1. The Goals of this project is to find out the characteristic of customers who cancelled and finding a pattern in cancelled booking by doing an exploratory data analysis

2. Building classification machine learning model to predict cancellation, Above 0.75

3. Build and Deploy web application / dashboard using flask from our machine learning algorithm, that can predict of cancellation based on user input

Business Question

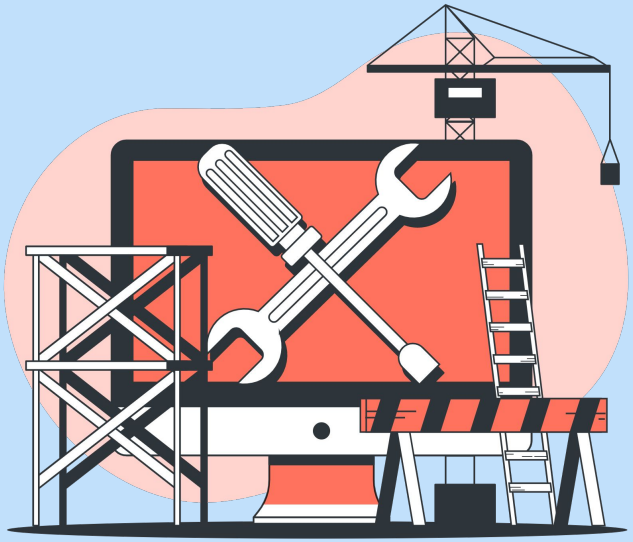
List of Question to help achieving the goal

- How Market Segment Of Booking Affecting Cancellation
- How's a lead time of a booking affecting cancellation
- How's different deposit type affecting cancellation of a booking
- How does cancellation rate of booking from portugal and booking that's made outside portugal
- What Are The Other Factors that affecting cancellation of booking
- What machine learning algorithm that has the highest accuracy when it comes predicting hotel booking cancellations

Project Limitation

This hotel booking cancellation project only applied for hotel bookings in Lisbon Region and Algarve Region both location are located in portugal.

predicting cancellation with this web application outside both region might have not so accurate result due to the location constraint, different pattern of cancellation



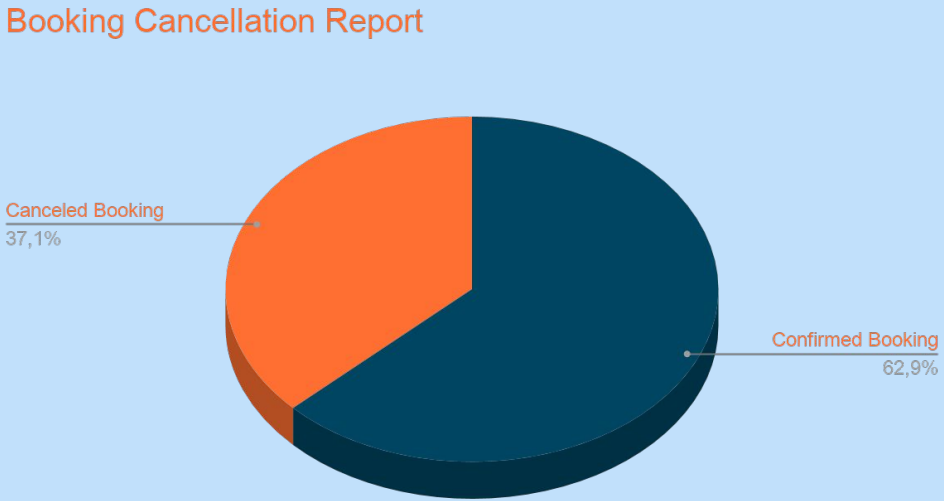
DATA ANALYSIS

TABLE OF CONTENTS

01

Cancellation Insight & Analysis

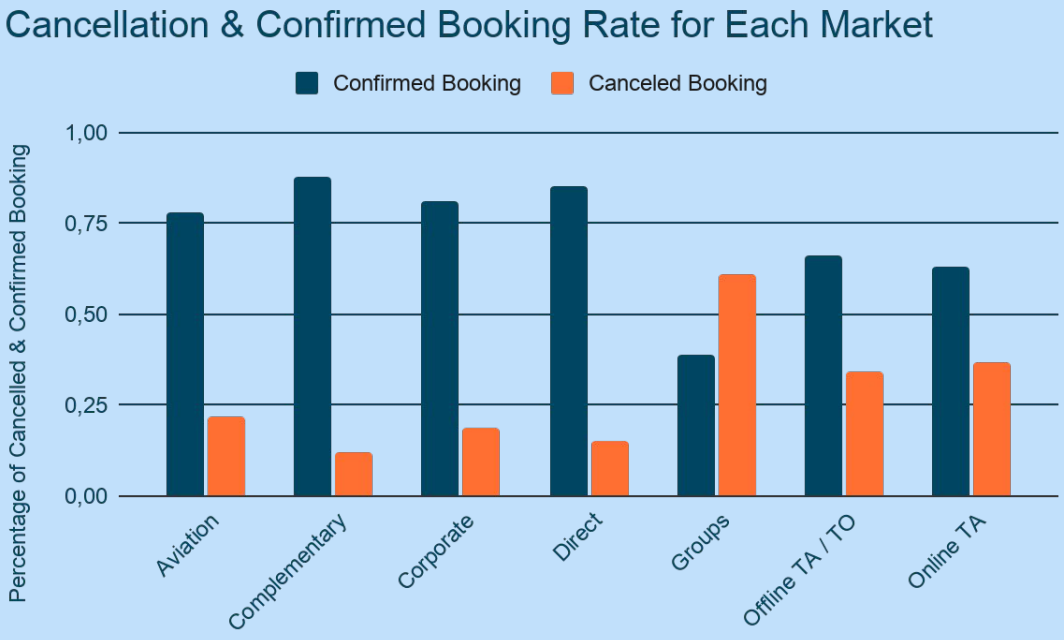
General Cancellation Report



Based on the dataset we see that percentage of cancellation in portugal from **July 2015 - August 2017 is 37.1%**, which is slightly below industry standard around that year at **41. 3%**

source : Hotel management net

Cancellation Rate Per Market Segment

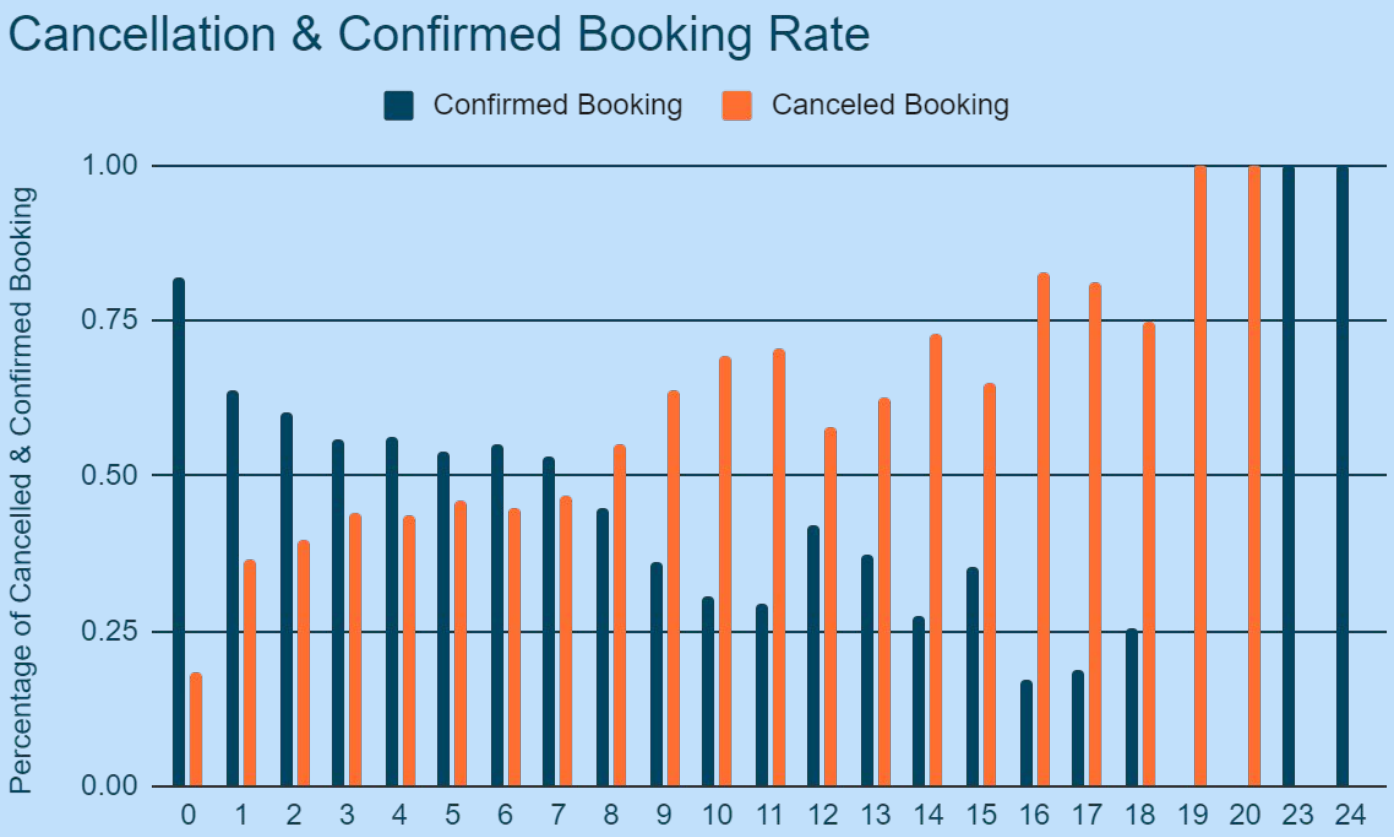


1. from our analysis we see that **corporate , Direct , and Aviation** has a cancellation rate around **18 - 22 %** of their booking

2. **Travel Agent (Online / Offline)** has a cancellation rate around **34 - 36 %**

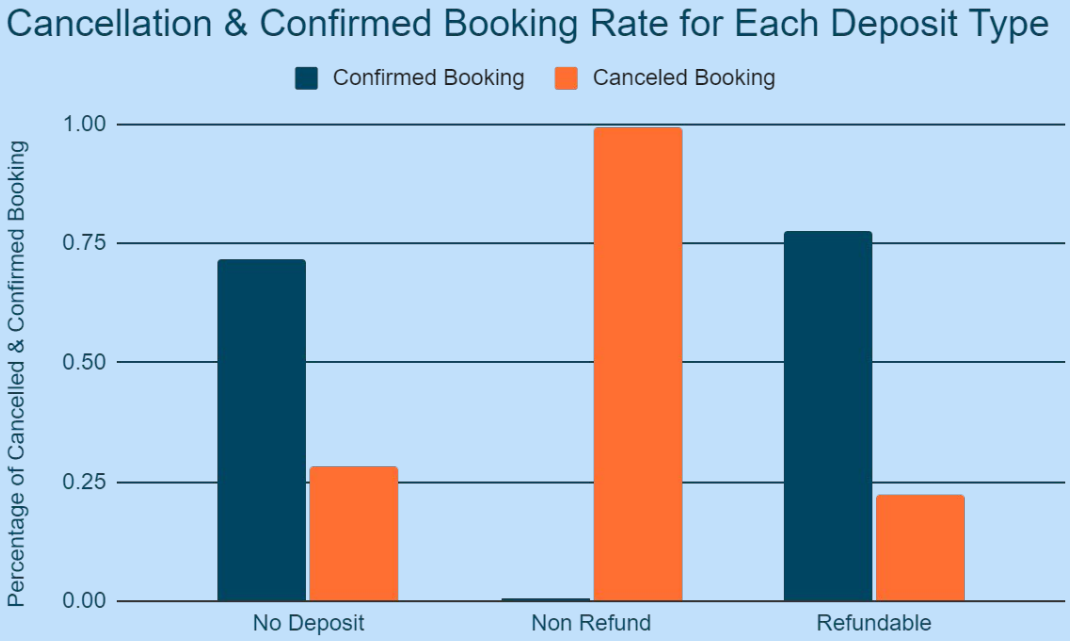
3. Lastly **Group** has the highest cancellation rate around **61 %**

Monthly Lead Time & Cancellation Rate



- Booking that has less than or equal **to 7 months** lead time have a higher confirmed booking rate **(>50%)** to it canceled rate
- Booking that has more than **7 months** have a higher cancellation rate compared to it's booking rate
- cancellation is positively correlated with lead time
  - **the higher the lead time the higher the cancellation rate**
  - **the shorter the lead time the less likely the booking will be cancelled**

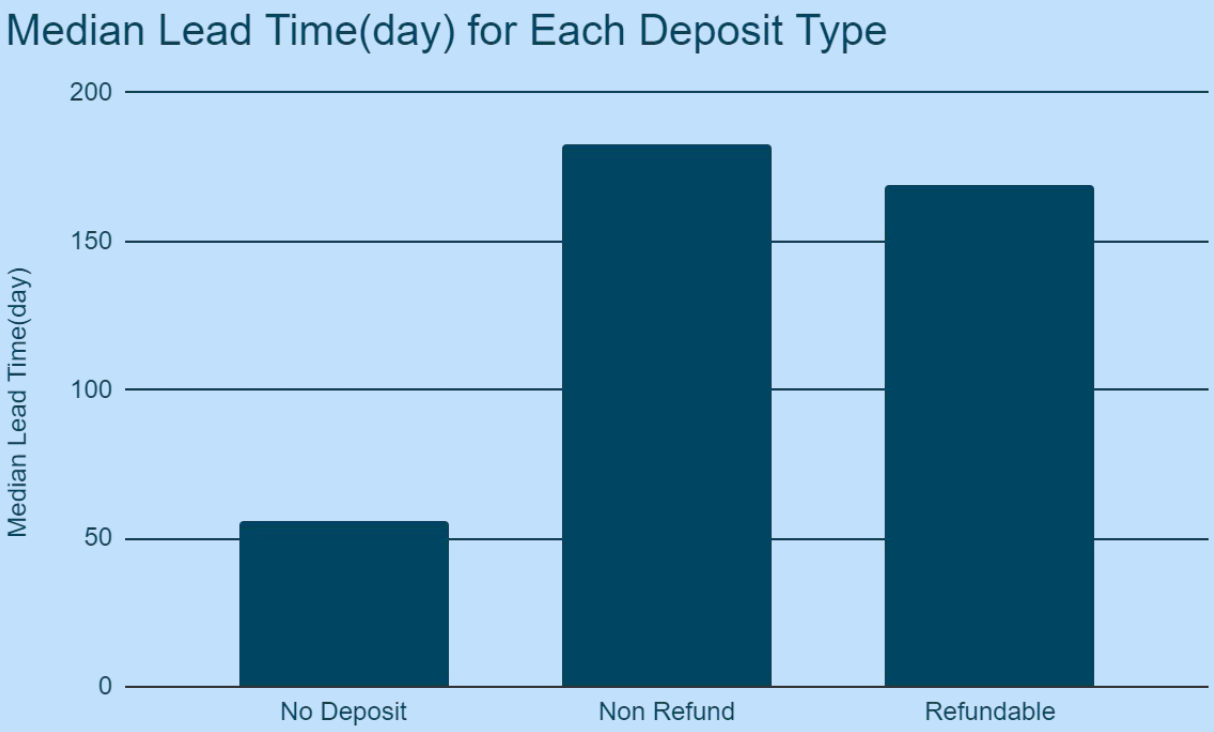
Deposit Type & Cancellation Rate



- **Non Refund Deposit type** has the highest cancellation rate among all deposit type **(99.4 %)**
- **No Deposit** has the second highest cancellation rate at **(28.3%)**
- **Refundable Deposit** has the lowest cancellation rate among all of the deposit, it cancellation rate is at **(22%)**

For the hotels this is nothing alarming since they don't lose revenue when no refund booking is canceled. but it's always a good practice to question something that's extraordinary happening. In this case we will take a closer look at the median lead time for each deposit type

Median Lead Time for Each Deposit Type

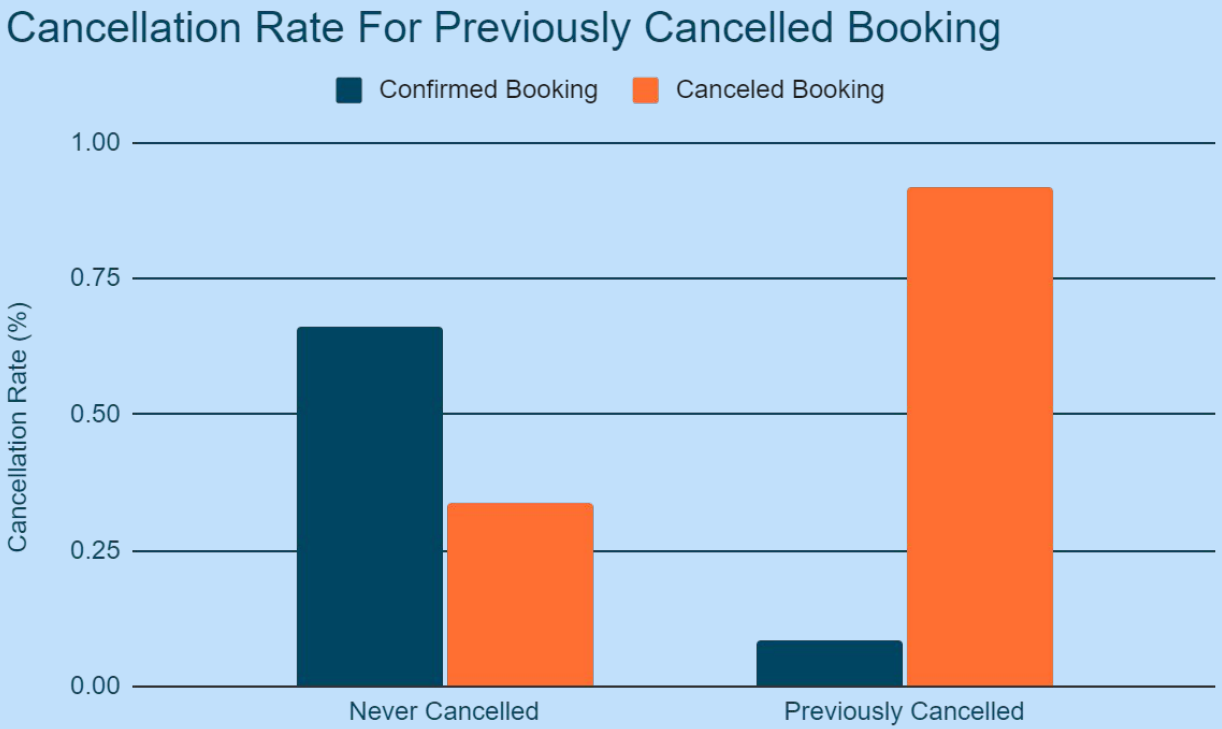


Based on the previous analysis we saw that non refund booking has the highest cancellation rate compared to other booking type,

and based on lead time and cancellation analysis we saw that the longer lead time has higher cancellation rate compared to the shorter one

- **Median Lead Time Non Refund 183 Days**
- **Median Lead Time Refundable 169 Days**
- **Median Lead Time for No deposit 56 Days**

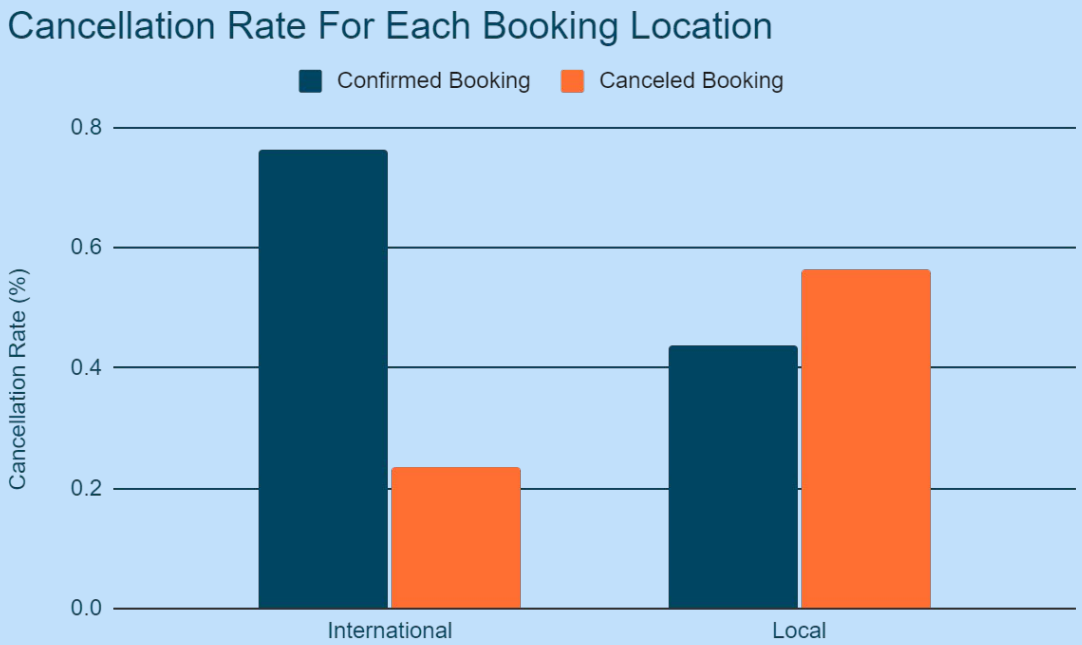
Previously Cancelled Booking & Cancellation Rate



- Booking That's previously Canceled has **92%** cancellation rate
- Booking that's never been canceled before has cancellation rate of **34%**

this shows that booking that has been canceled before will more likely to be cancelled again

Booking Location & Cancellation Rate

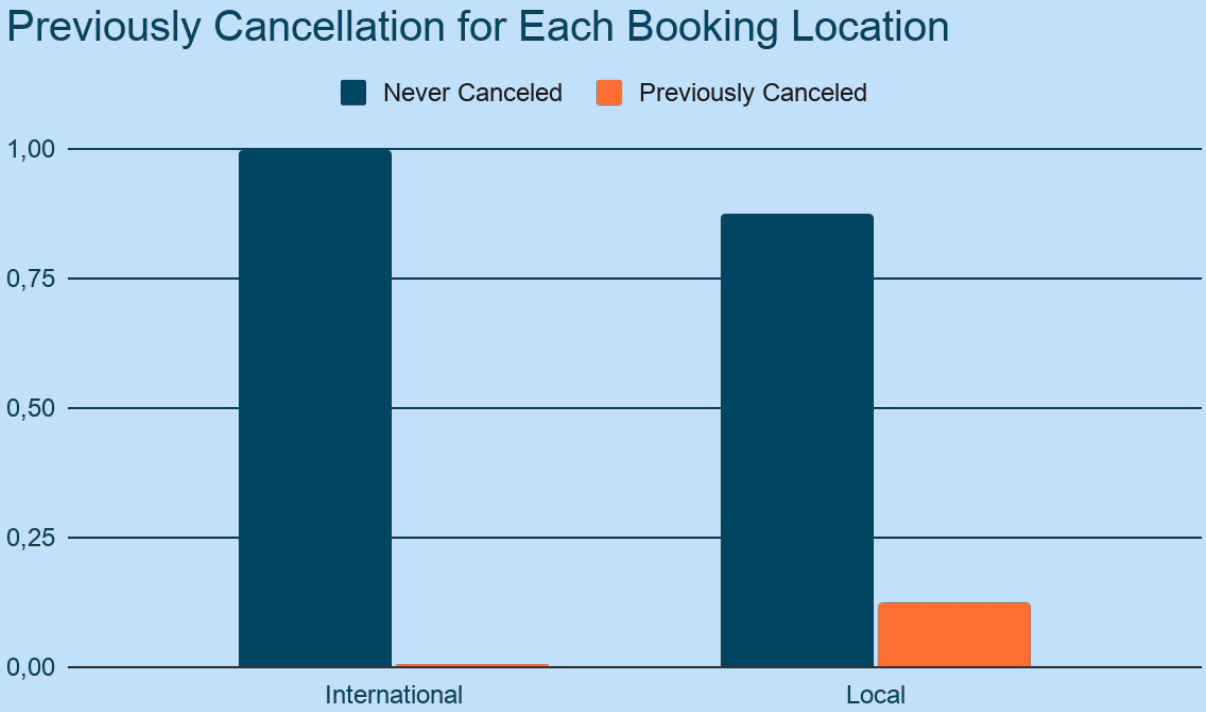


- **International Booking has 24% Cancellation Rate**
- **Local Booking has 56% Cancellation Rate**

This arise a question

Why does local booking are more likely to be canceled compared to international booking ?

Booking Location & Previously Cancellation

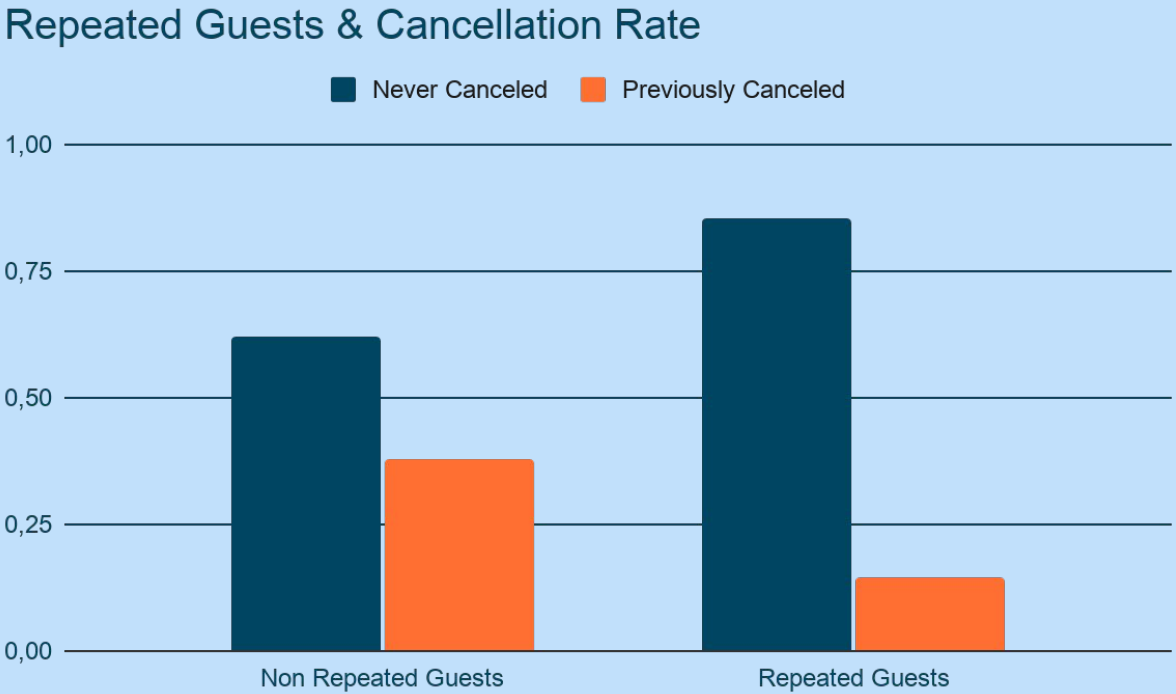


based on the previous analysis we saw that booking that has been canceled before has a cancellation rate of 92 %

- For international booking **99.5%** of international booking never been canceled before
- while for local booking **13%** of bookings has been canceled before

This higher number of previous cancellation definitely is one of the factors why local booking has a higher cancellation rate compared to international booking

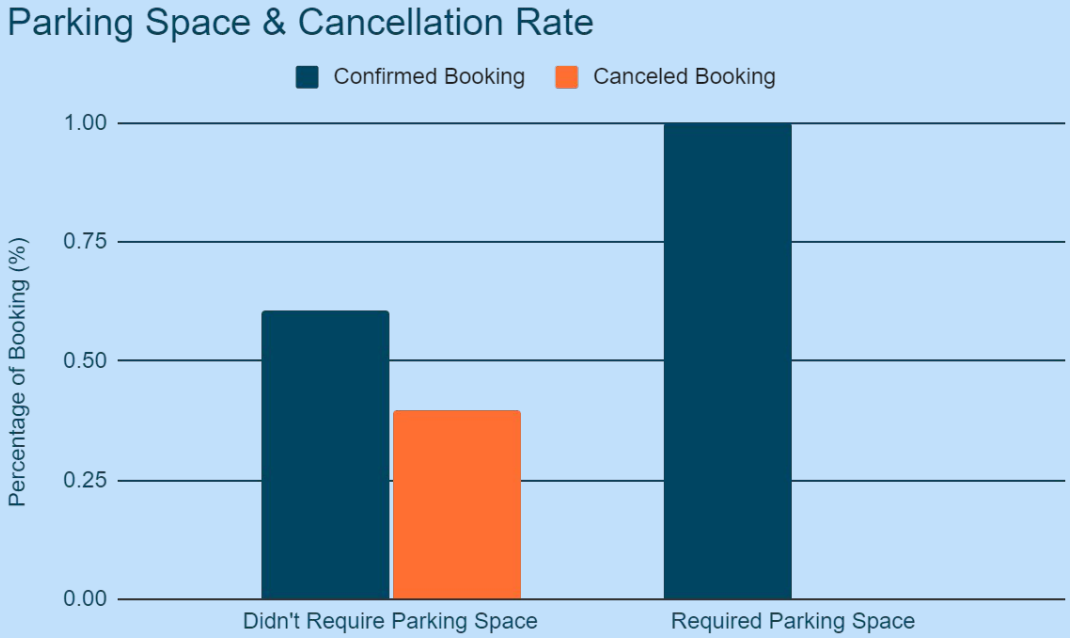
Repeated Guests & Cancellation Rate



- Repeated guests cancellation rate is around **14 %**
- Non Repeated guests are more than 2 x more likely to canceled their booking compared to repeated guest **(34%)**

in conclusion :
repeated guests are more likely to confirmed the booking compared to non repeated guests

Parking Space & Cancellation



From July 2015 to August 2017 There are **7407 (6.2 %)** out of total booking that required a parking space .

- Out of **7407 bookings** that required a parking space **there's not a single booking that's canceled (0 Cancellation for booking that required a parking space)**

This conclude booking that required a parking space will most likely be confirmed

02

EDA
Recommendation

EDA Recommendation

Only Non Refundable Deposit For Group Booking

From the analysis we see that group booking has the highest cancellation rate.

With this policy hotel wouldn't suffer any loss of revenue or loss of potential revenue due to cancelled group booking

Maximum Lead Time for Booking

We see a pattern of booking that has more than **210 days** lead time are more likely to be canceled

setting up a maximum lead time means users wouldn't be able to make booking that's too far in advance

Combination of Restriction

Setting up maximum lead time for booking might have resulted in hotel visibility to the potential guest search.

Combining deposit type lead time might help to get the hotel more exposure without higher risk of losing revenue. **(ex. Non Refundable Deposit only for booking > 210 days in advance)**

EDA Recommendation

Increase Direct Booking Market Segment

From this dataset we see that direct booking has the least cancellation rate at **15%** (*outside complimentary)*

with only being **10 % of the total booking**, the increase of direct booking might lead to less cancellation

strategy to increase direct booking

Stricter Cancellation Policy For Previously Canceled Booking

Booking that's previously canceled has cancellation rate of **92%**.

looking at this pattern we see that booking that's previously canceled are most likely to be canceled again.

to protect hotel from losing revenue due to this kind of cancellation, **hotels need to set booking payment in advance / non refundable deposit,** for booking that has been canceled before

Attracting Customers Who Drives

6% of total booking required a parking space **(7407 bookings)**, and out of **7407** bookings none of them were canceled**.**

That's around **10%** of total confirmed booking. Hotel could promote to attract customer that drives

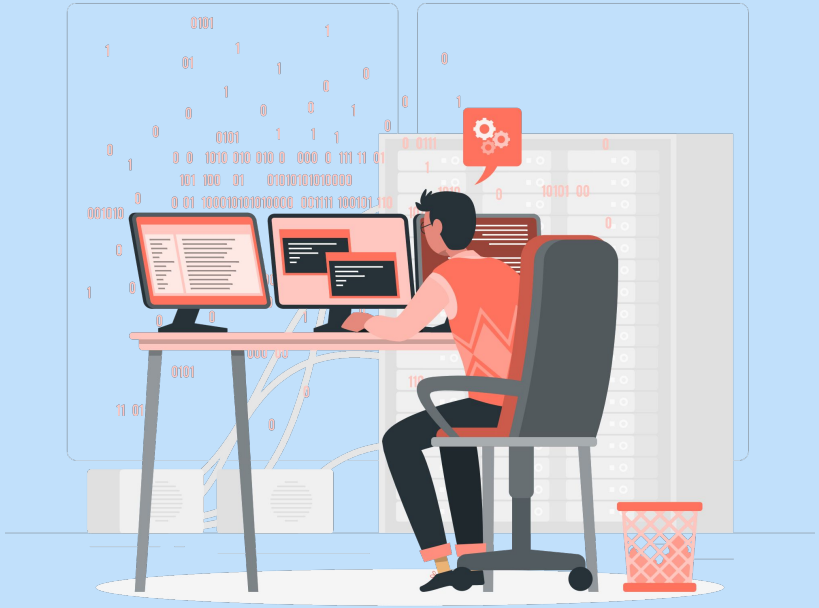
03

Machine Learning

Machine Learning Algorithm

Algorithm Used:

1. Logistic Regression
2. KNeighbors Classifier
3. Decision Tree Classifier
4. Random Forest Classifier
5. XGB Classifier



Base Model Evaluation Matrix

Evaluation Matrix

	Logistic Regression	KNN	Decision Tree	Random Forest	XGB
Accucary	0.804135	0.835399	0.831491	0.840862	0.848973
Recall	0.615210	0.742605	0.747592	0.763119	0.748272
Precision	0.810875	0.799219	0.787206	0.798695	0.827837
F1 Score	0.699620	0.769873	0.766888	0.780502	0.786046

Accuracy Train & Test

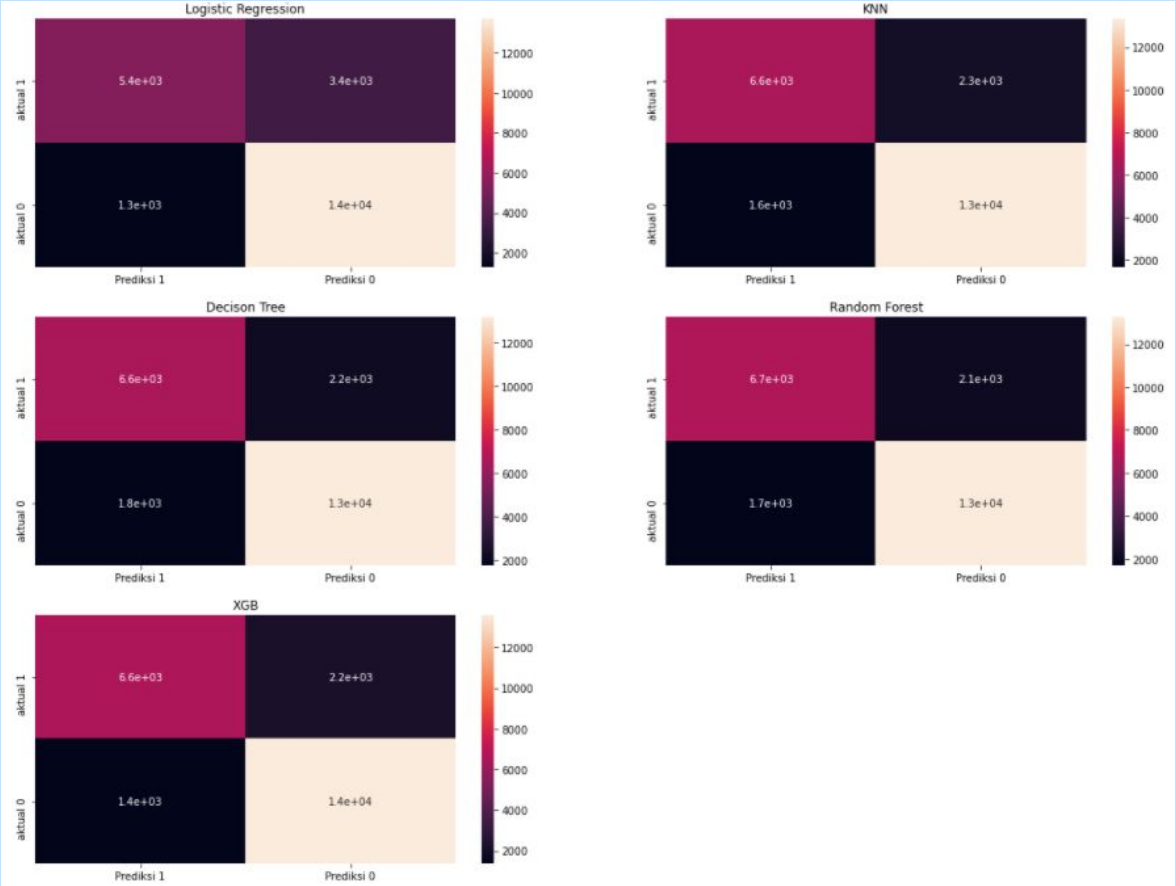
	training	testing
Logreg	0.802412	0.804135
KNN	0.879628	0.835399
Decision Tree	0.941978	0.831491
Random Forest	0.941957	0.840862
XGB	0.859478	0.848973

- **KNN, Decision Tree, Random Forest** algorithms have overfitting condition
- **XGB Classifier** has the best accuracy Score in the base model

Accuracy as The Primary Evaluation Matrix

- First because the data is somehow balance between the target class **(63%) Confirmed, 37% Canceled**
- **Every Class is equally important**

Base Model Confusion Matrix



Tuned Model Evaluation Matrix

Evaluation Matrix

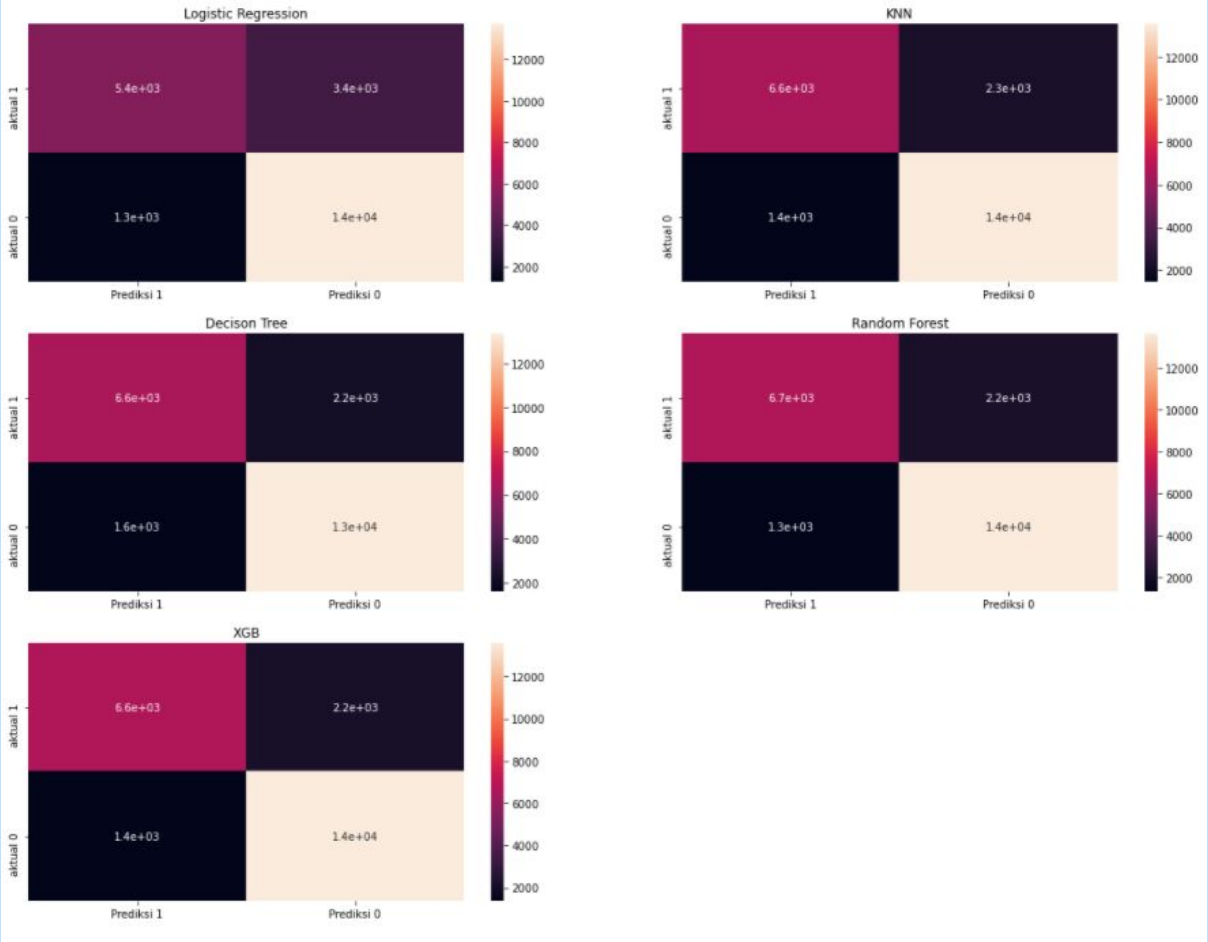
	Logistic Regression	KNN	Decision Tree	Random Forest	XGB
Accucary	0.803925	0.844350	0.839265	0.852250	0.850485
Recall	0.613397	0.744078	0.749405	0.753825	0.752465
Precision	0.811759	0.819498	0.803793	0.831895	0.828529
F1 Score	0.698773	0.779969	0.775647	0.790938	0.788667

Accuracy Train & Test

	training	testing
Logreg	0.802591	0.803925
KNN	0.838835	0.844350
Decision Tree	0.836114	0.839265
Random Forest	0.852797	0.852250
XGB	0.851399	0.850485

- No More Overfitting Condition in All The Algorithms
- After Hyperparameter Tuning Random Forest Classifier Has The Highest Accuracy Among All Algorithms

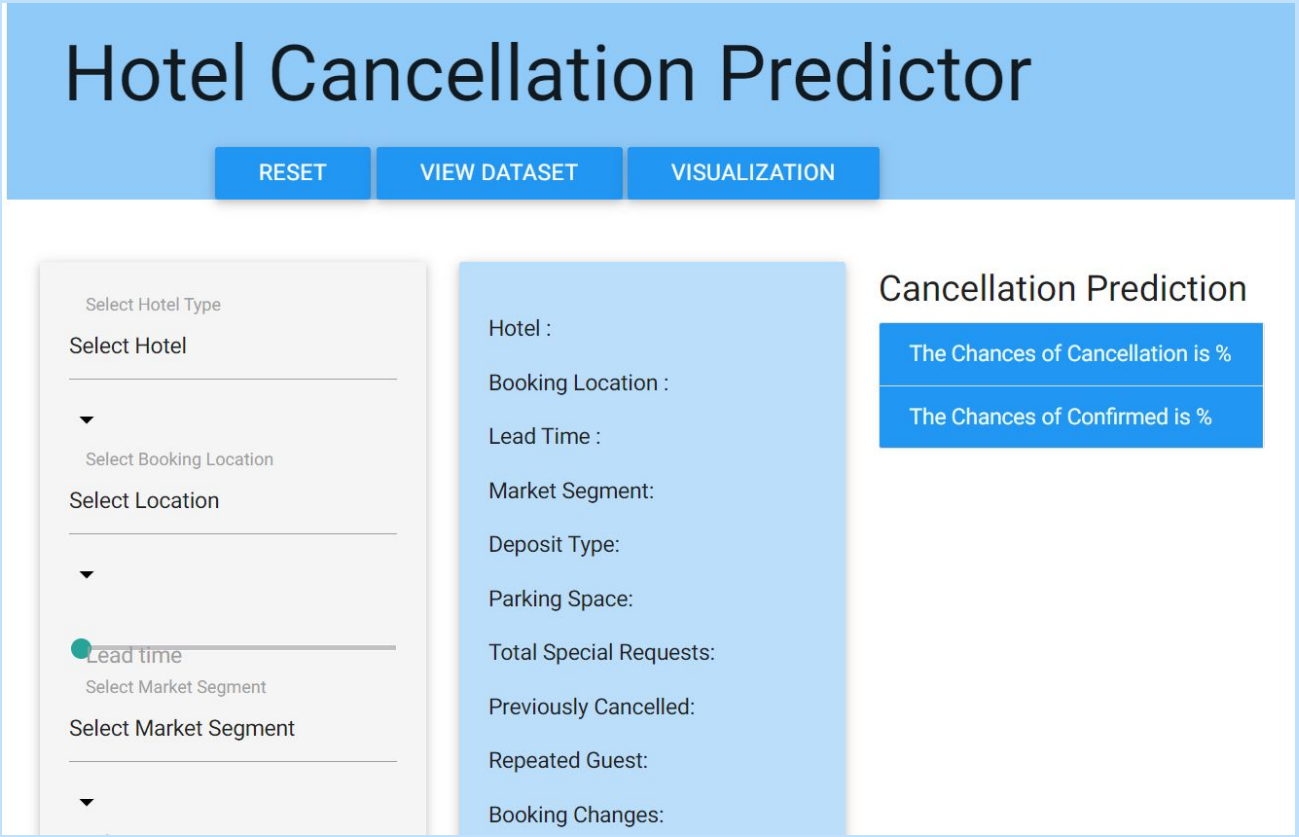
Tuned Model Confusion Matrix



04

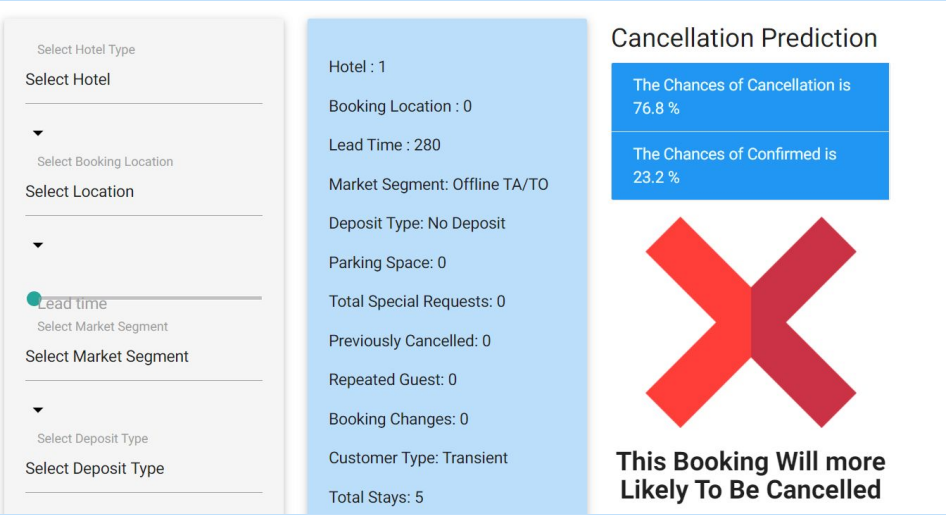
Machine Learning Dashboard

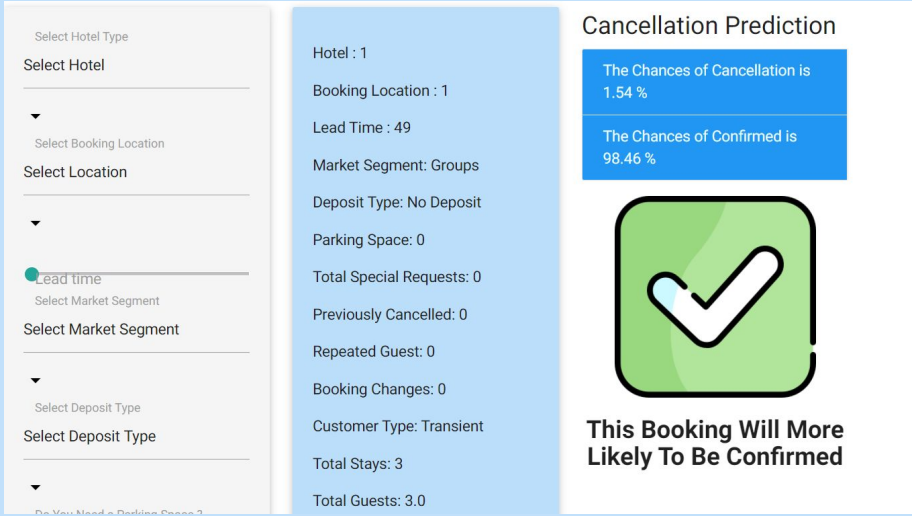
Dashboard



- The Dashboard have drop down option and sliders to adjust the user input
- The dashboard also have pages to see the dataset & visualization of the chart
- This Dashboard uses predict_proba instead of predict to show the user probability of a canceled / confirmed booking

Dashboard Predict Cancelled Result





Thank You!

For full notebook of the process
please go to my github page

Connect with me :


teguharia172@gmail.com