Predicting Hotel Booking Cancellation

Lee Sheau Wei
Summative Capstone Project

Introduction

Target Audience

Hotel Management

Background

Hotel Booking :

i) Direct booking
ii) Online booking (OTA)

Majority bookings made through Online Travel Agencies (OTAs) such as Booking.com, Hotels.com and Agoda

→ getting popular and expanding quickly worldwide
→ OTA → free cancellation policy

Problem

Rising popularity of OTAs
→ Higher Booking Cancellation rate

Impact of higher cancellation rate :

i) Harder to accurately forecast
ii) Non-optimized occupancy
iii) Revenue loss
iv) Lowering prices last minute, reducing profit margin

Objective

- Use a real life hotel booking dataset to gain better insights, and get a full picture of its behaviour.
- Explore different Machine Learning techniques to predict Hotel Booking Cancellation.
- Build the best Machine Learning Model to predict the Hotel Booking Cancellation as accurate as possible, in order to manage their business accordingly, and increase their revenue.
- Identifying the most important features to predict and have a direct impact on Hotel Booking Cancellation.

Methodology

Data Set

→ City hotel & Resort Hotel

→ 119390 unique rows & 36 features

Baseline Model

Logistic Regression

Multiple Models

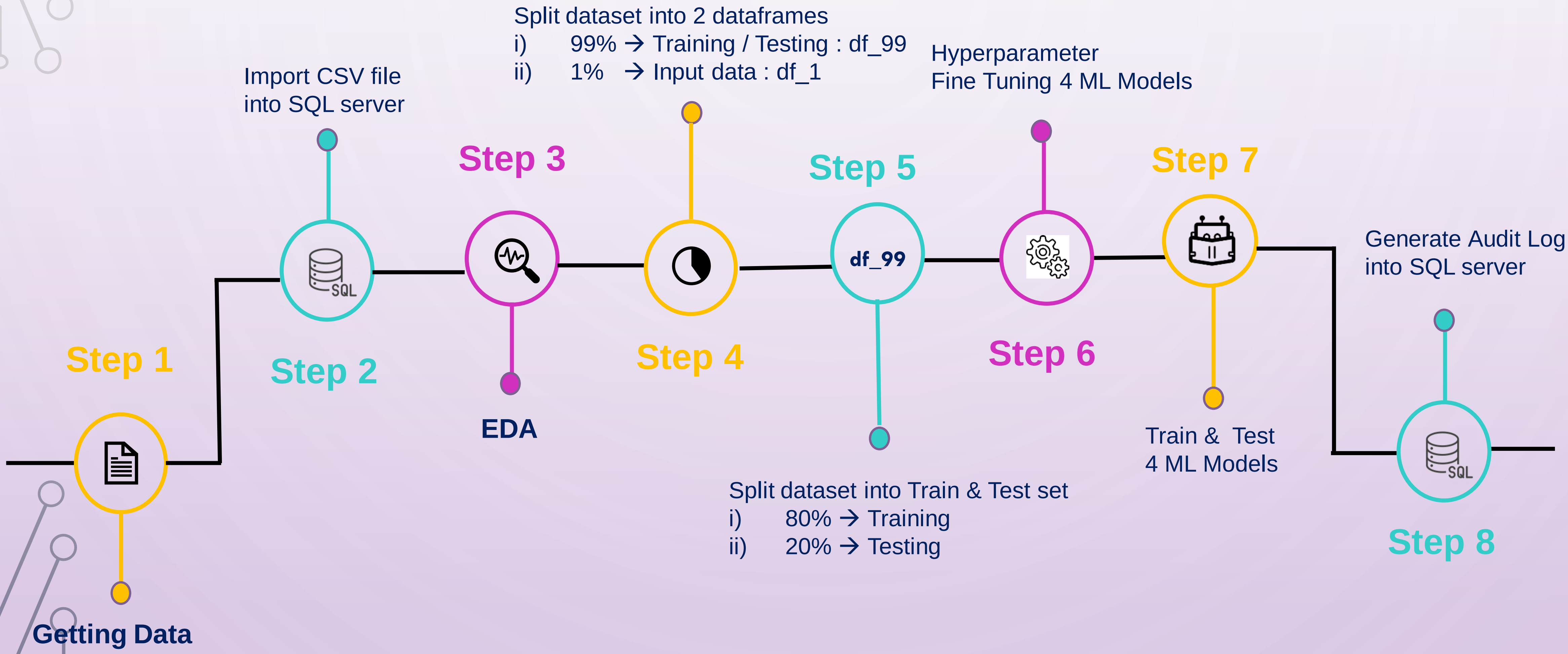
Random Forest

Multi Layer Perceptron

eXtreme Gradient Boost

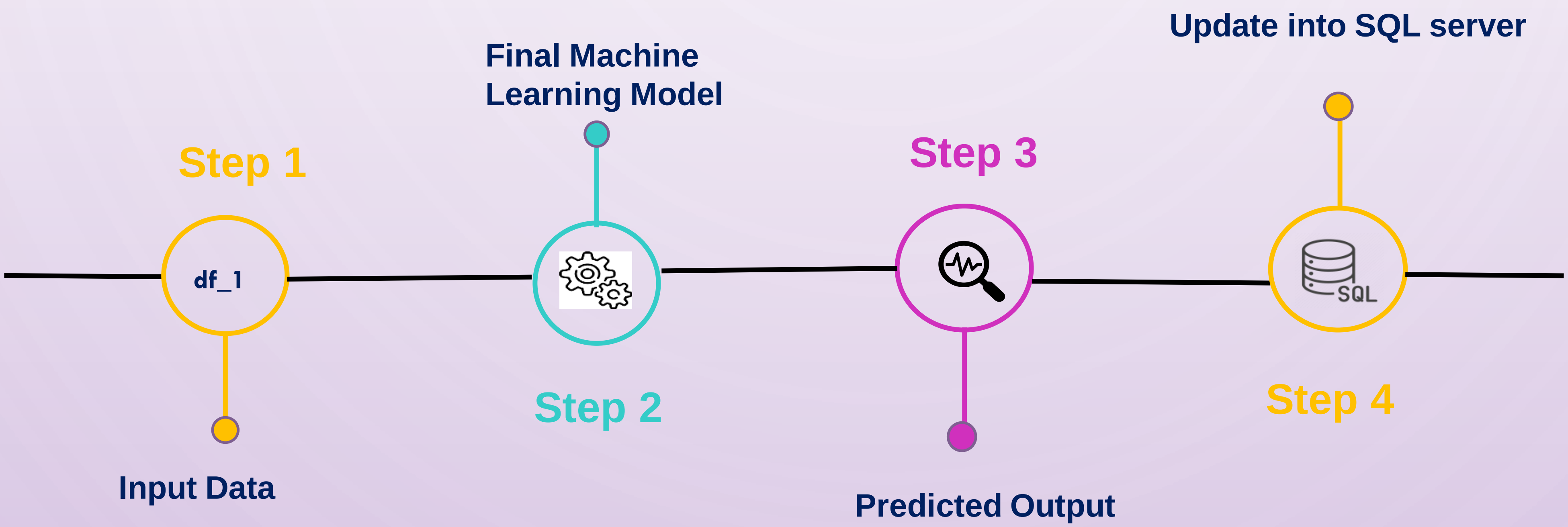
Methodology

For every ML model tested insert relevant audit information in a database table



Methodology

Write inputs & corresponding predicted output to a database table

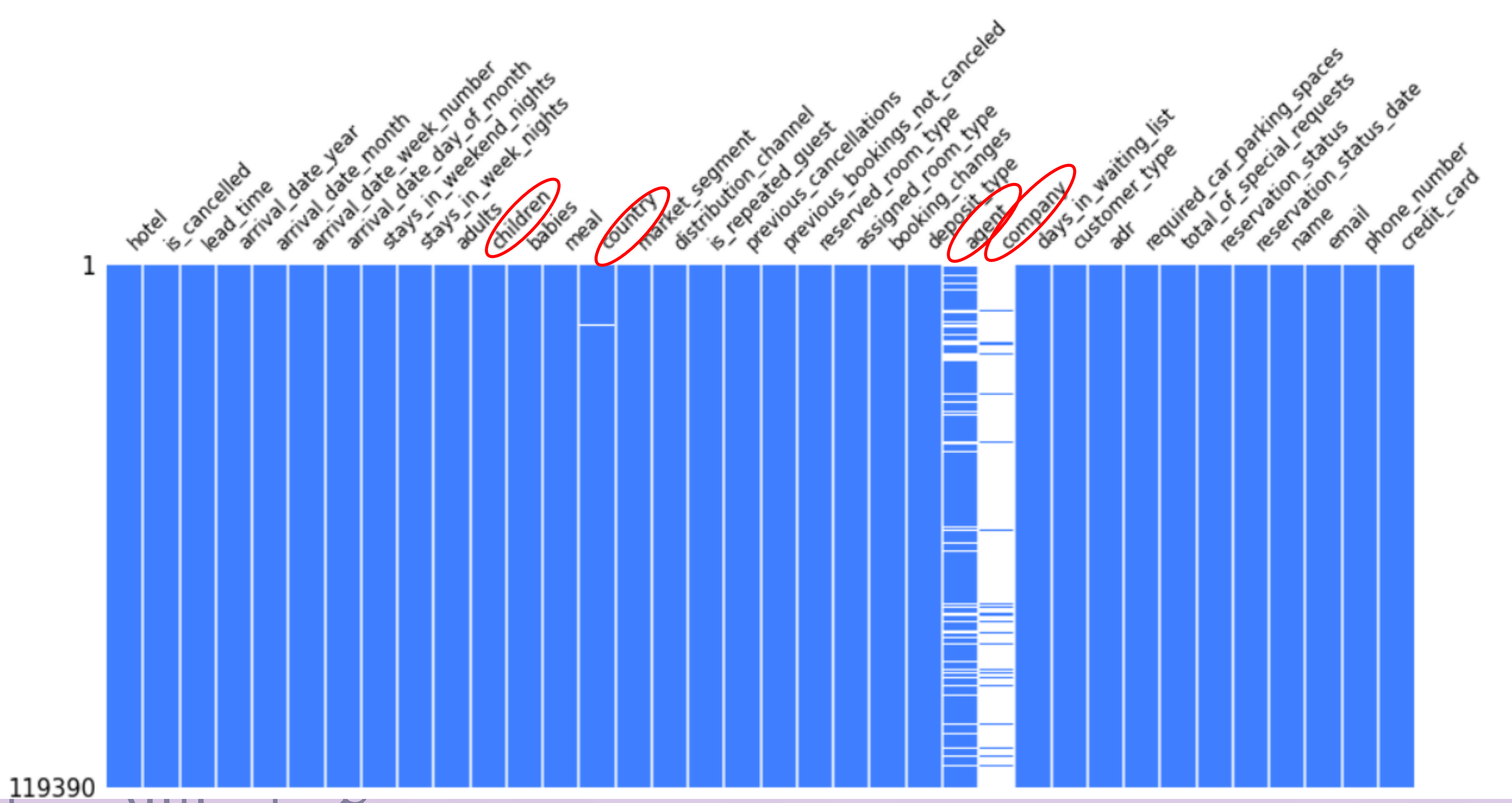


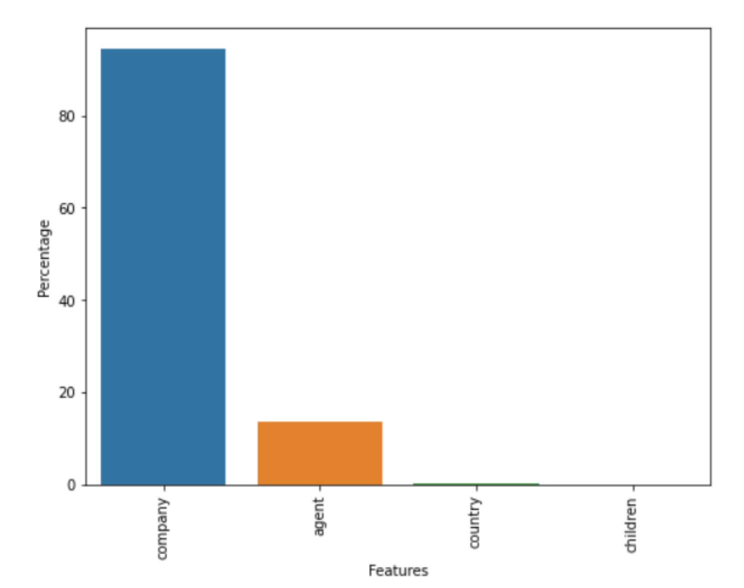
Features Description

	Features	Description					Features	Description
1	hotel	Types of Hotel	13	meal	Type of meal booked	25	company	ID of the company that made the booking
2	is_cancelled	Value indicating whether the booking was cancelled (0 : not cancelled; 1 : cancelled)	14	country	Country of origin. Categories are represe	26	days_in_waiting_list	Number of days the booking was in the waiting list before it was confirmed to the customer
3	lead_time	Number of the days prior to arrival that the booking was placed in the hotel	15	market_segment	Market segmentation to which the booking was assigned. In categories the term "TA" means "Travel Agents" and "TO" means "Tour Operators"	27	customer_type	Type of customer. i) Contract : when the booking has an allotment or other type of contract associated to it ii) Group : when the booking is associated to a group iii) Transient : when the booking is not part of a group or contract, and is not associated to other transient booking iv) Transient party : when the booking is transient, but is associated to at least other transient booking
4	arrival_date_year	Year of arrival date	16	distribution_channel	Name of the distribution channel used to make the booking. The term "TA" means "Travel Agents"	28	adr	Average daily rate
5	arrival_date_month	Month of arrival date with 12 categories : "January" to "December"	17	is_repeated_guest	Value indicating whether the customer was a repeated guest at the time of booking. (0 : No; 1 : yes)	29	required_car_parking_spaces	Number of car parking spaces required by the guest
6	arrival_date_week_number	Week number of the arrival date in the year (1 to 52)	18	previous_cancellations	Total of previous bookings that were cancelled by the guest	30	total_of_special_requests	Number of special request made (eg. sea view, twin bed or high floor)
7	arrival_date_day_of_month	Day of the month of the arrival date (1 to 31)	19	previous_bookings_not_canceled	Total of previous bookings that were not	31	reservation_status	Reservation last status i) Cancelled : booking was cancelled by the customer ii) Check-Out : customer has checked in but already departed iii) No-Show : customer did not check in and did inform the hotel of the reason why
8	stays_in_weekend_nights	Number of weekend nights (Saturday and Sunday) the guest stayed	20	reserved_room_type	Room type requested by the guest	32	reservation_status_date	Date at which the last status was set. This variable can be used in conjunction with the Reservation
9	stays_in_week_nights	Number of week nights (Monday through Friday) the guest stayed	21	assigned_room_type	Room type assigned to the booking	33	name	Name of the guest
10	adults	Number of adults	22	booking_changes	Number of amendments made to the booking (arrival or departure dates, number of persons, type of meal, ADR or reserved room type)	34	email	Email address of the guest
11	children	Number of children	23	deposit_type	Indication on if the customer made a deposit to guarantee the booking. This variable can assume 3 categories : i) No Deposit ii) Non Refund iii) Refundable	35	phone_number	Phone number of the guest
12	babies	Number of babies	24	agent	ID of the travel agency that made the booking	36	credit_card	Credit card of the guest

EDA Data Preparation

Checking Missing Values





company	94.306893
agent	13.686238
country	0.408744
children	0.003350

EDA Data Cleaning

1. Drop Company feature
2. Replace Missing values with 999 in Agent feature

agent

agent
999
999

3. Remove rows with missing value in Country feature
4. Replace Missing values with 0 in Children feature
5. Remove Irrelevant Features

name	email	phone-number	credit_card
Duane Mccormick	Mccormick_Duane39@comcast.net	480-793-6313	************5365
Kim Lloyd	Kim_Lloyd25@yandex.com	134-648-6325	************3187
Steven Logan	Steven_Logan@protonmail.com	859-131-1311	************3203
Ryan Gibson	Ryan_Gibson@att.com	866-551-7973	************1965

EDA Data Cleaning

6. Remove Adults & Children = 0

	hotel	adults	children	is_cancelled	stays_in_week_nights	stays_in_weekend_nights
2224	Resort Hotel	0	0.0	0	3	0
2409	Resort Hotel	0	0.0	0	0	0
3181	Resort Hotel	0	0.0	0	2	1
3684	Resort Hotel	0	0.0	0	4	1
3708	Resort Hotel	0	0.0	0	4	2
31765	Resort Hotel	0	0.0	0	8	2

7. Remove stays_in_week_nights & stays_in_weekend_nights = 0

	hotel	is_cancelled	adults	stays_in_week_nights	stays_in_weekend_nights
0	Resort Hotel	0	2	0	0
1	Resort Hotel	0	2	0	0
167	Resort Hotel	0	2	0	0
168	Resort Hotel	0	1	0	0
196	Resort Hotel	0	2	0	0
197	Resort Hotel	0	2	0	0
459	Resort Hotel	0	2	0	0

EDA Feature Engineering

1. Merge (Adults ,Children & Babies) features into a new feature
   → [total_guest]

2. Create a new feature consist of :
   i) stay_just_weekend
   ii) stay_just_weekday
   iii) stay_both_weekday_and_weekend

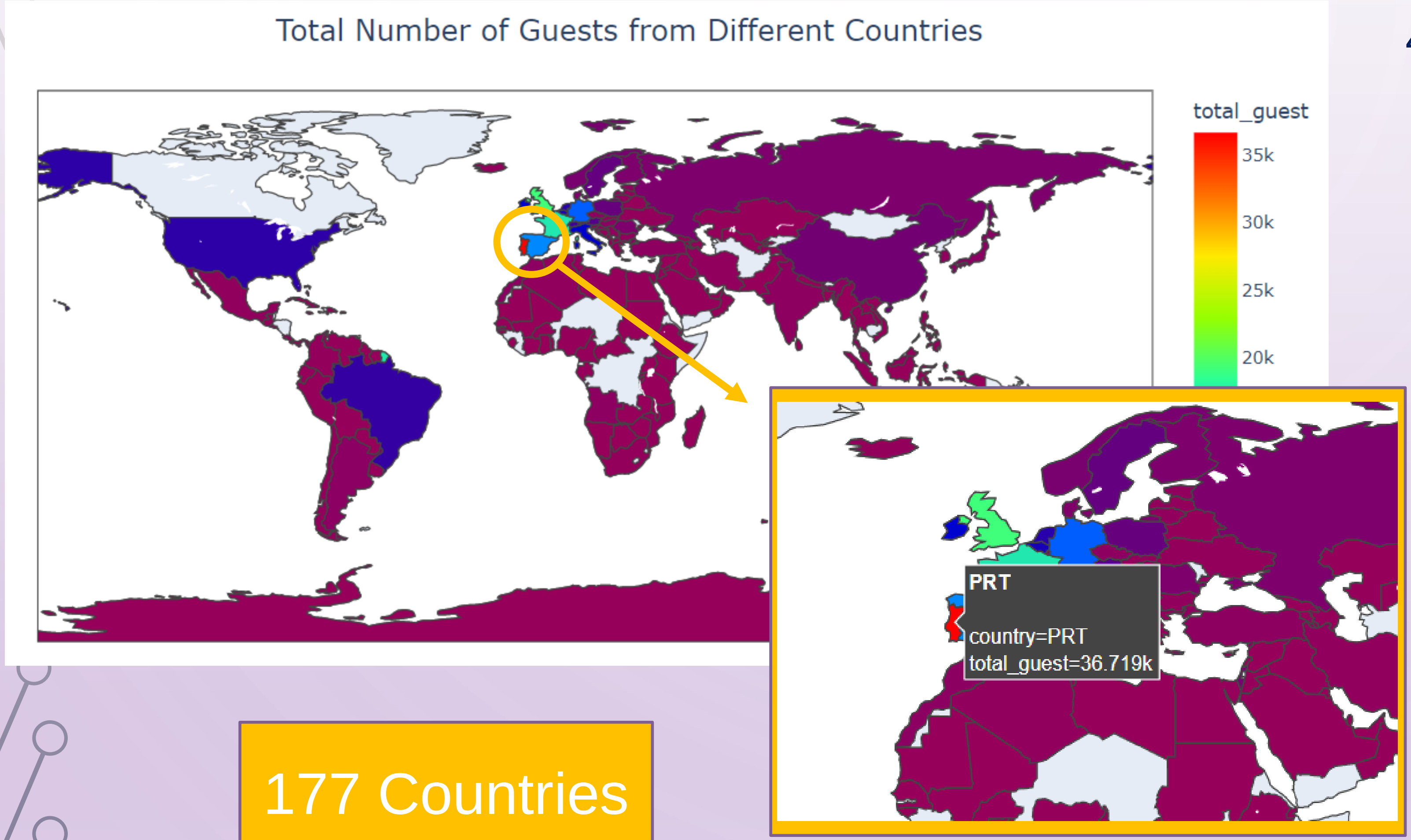
   → [weekend_or_weekday]

3. Drop features :
   Adults, Children, Babies,
   stays_in_week_nights &
   stays_in_weekend_nights

stays_in_week_nights VS stays_in_weekend_nights

stays_in_weekend_nights	0	1	2	3	4	5	6	7	8	9	10	12	13	14	16	Total
stays_in_week_nights																
0	0	4535	2355	0	0	0	0	0	0	0	0	0	0	0	0	6890
1	16247	7300	6526	0	0	0	0	0	0	0	0	0	0	0	0	30073
2	17865	8956	6741	0	0	0	0	0	0	0	0	0	0	0	0	33562
3	11531	6135	4520	0	0	0	0	0	0	0	0	0	0	0	0	22186
4	4477	2404	2653	0	0	0	0	0	0	0	0	0	0	0	0	9534
5	828	1186	8627	308	94	0	0	0	0	0	0	0	0	0	0	11043
6	0	0	843	298	345	0	0	0	0	0	0	0	0	0	0	1486
7	0	0	445	396	181	0	0	0	0	0	0	0	0	0	0	1022
8	0	0	390	130	132	0	0	0	0	0	0	0	0	0	0	652
9	0	0	81	61	86	0	0	0	0	0	0	0	0	0	0	228

EDA Feature Engineering



177 Countries

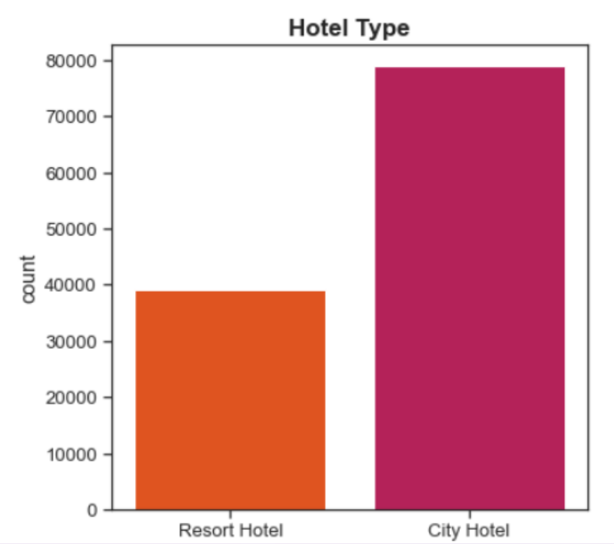
4. Create a new feature consist of :
   i) Europe
   ii) NorthAmerica
   iii) MiddleEast
   iv) South America
   v) APAC

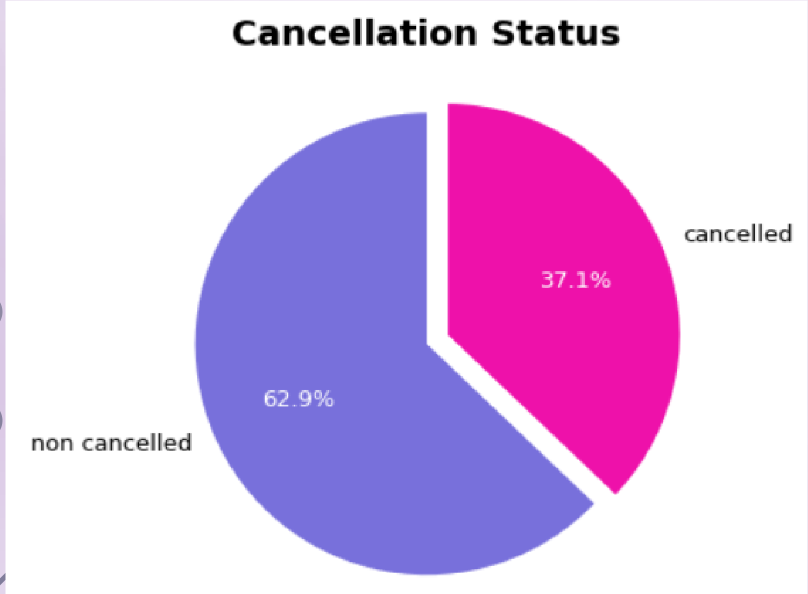
→ [continent]

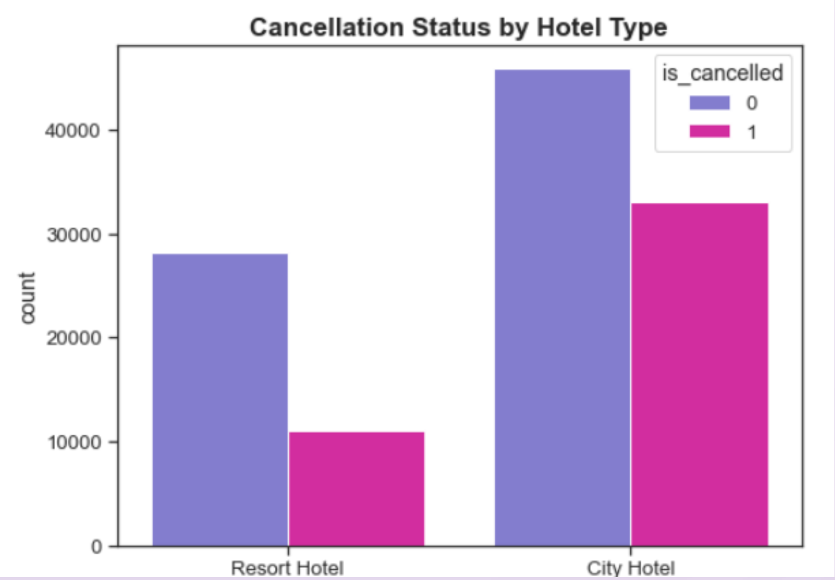
5. Drop Country feature

EDA Data Visualization

Binary Target Variable is_cancelled		
	0	non cancelled
	1	cancelled

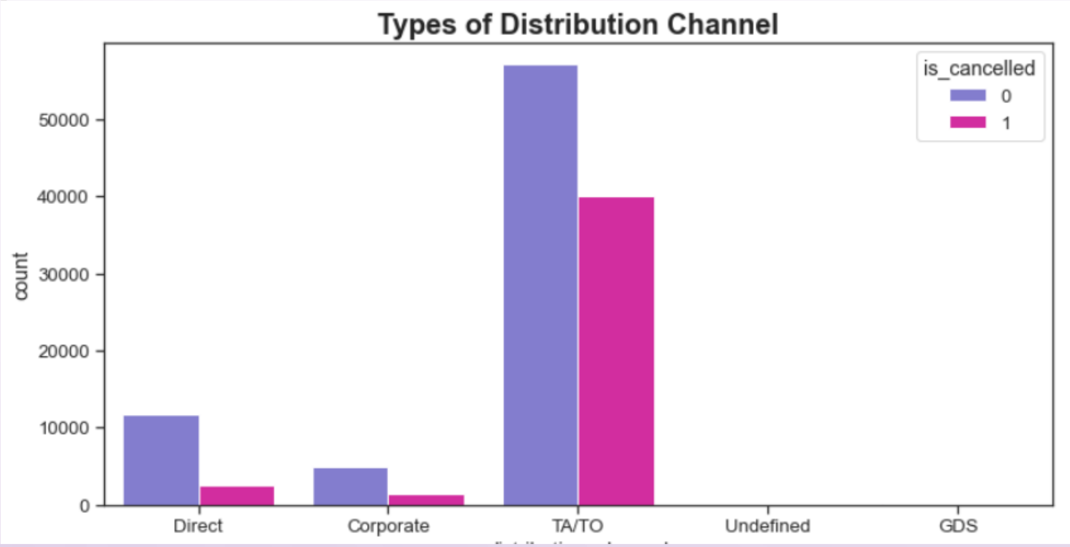


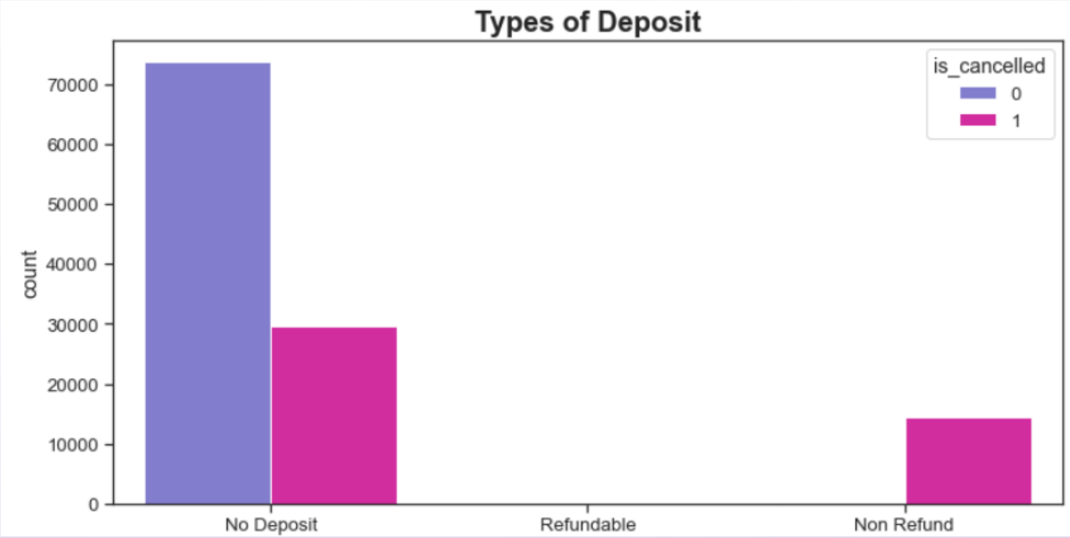


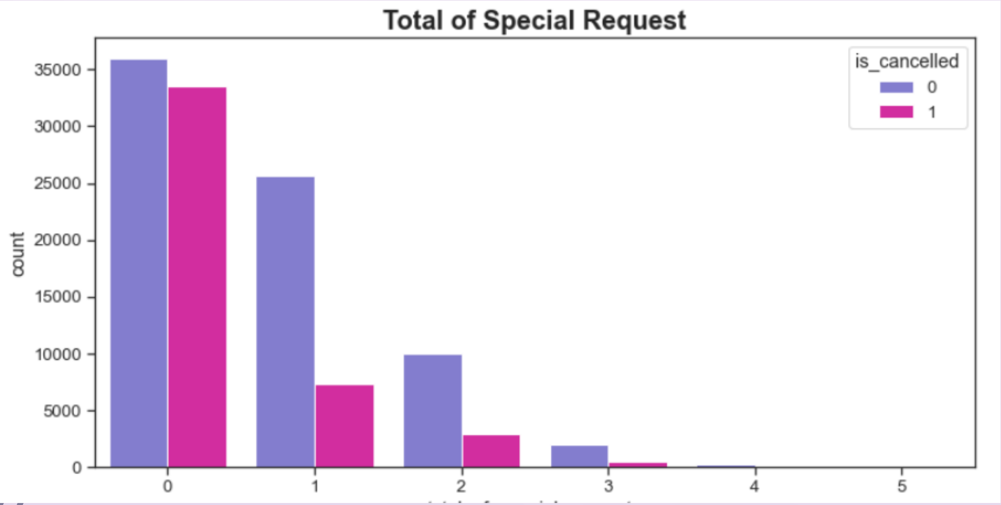


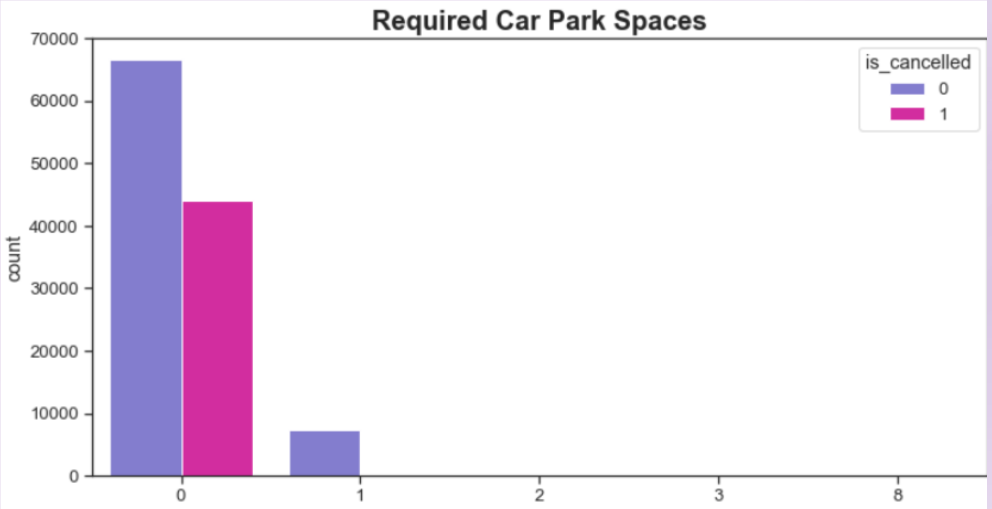
EDA Data Visualization

Target : is_cancelled VS independent variables



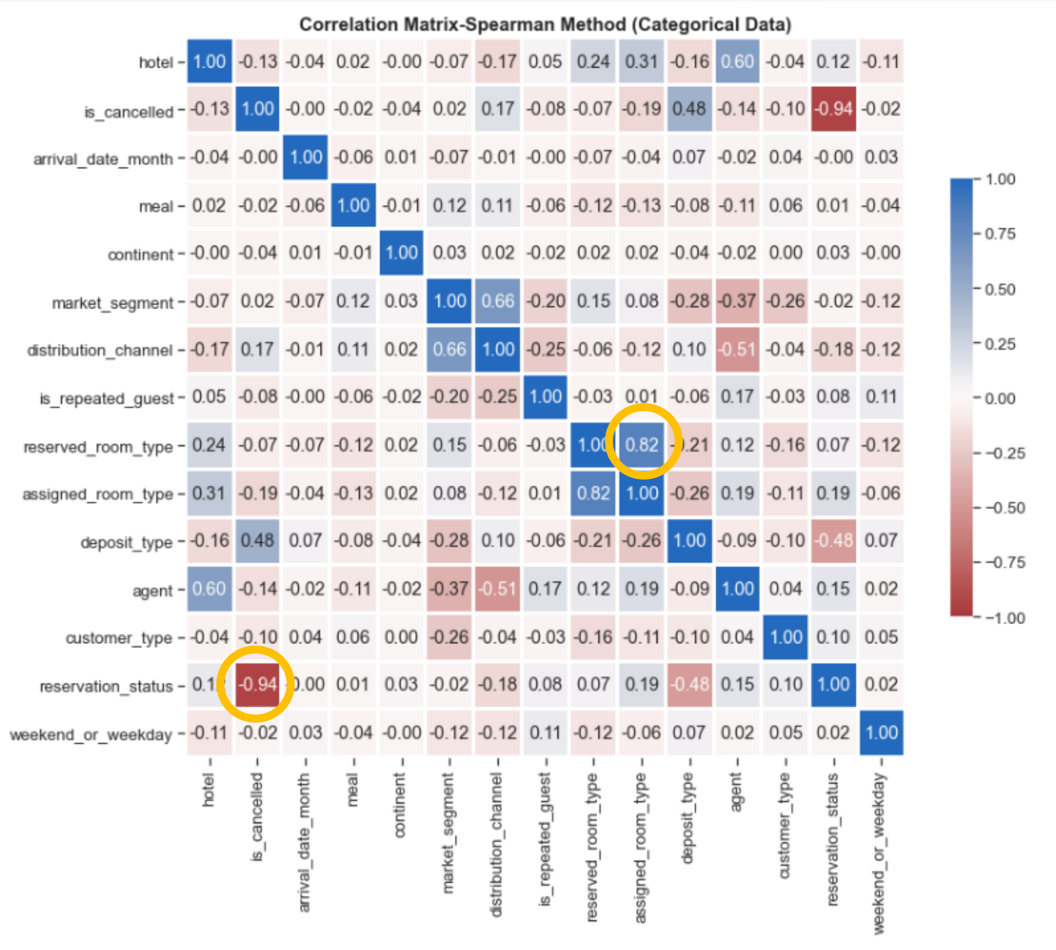






EDA Correlation Heatmap

Spearman’s Correlation (Categorical Data)



Highly Correlated Features

reservation_status VS is_cancelled : -0.94

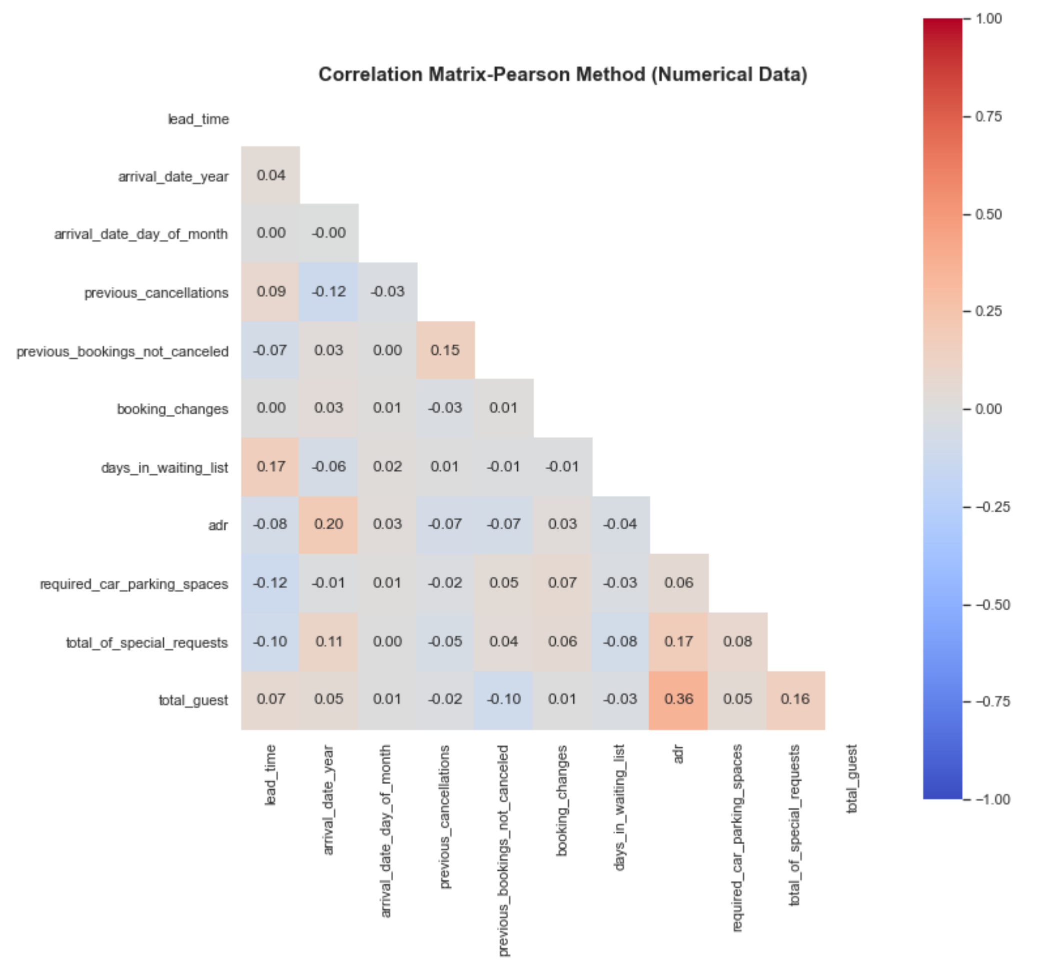
reservation_status	Canceled	Check-Out	No-Show	Total
is_cancelled				
0	0	73973	0	73973
1	42930	0	1189	44119
Total	42930	73973	1189	118092

reserved_room_type VS assigned_room_type : 0.82

Drop features :
reservation_status
reservation_status_date
reserved_room_type

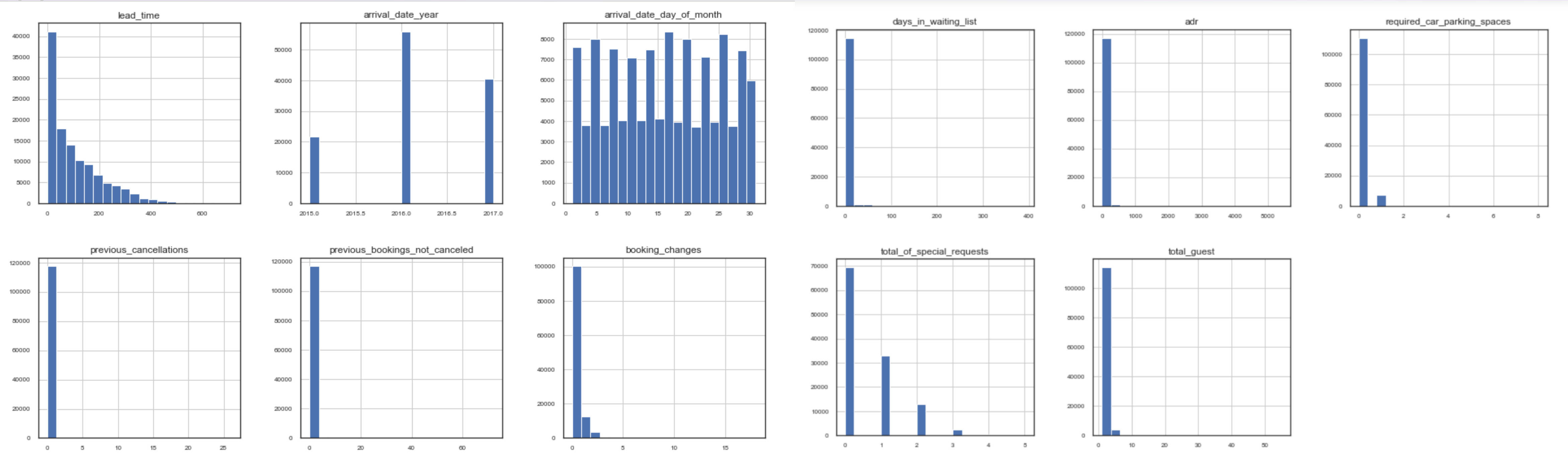
EDA Correlation Heatmap

Pearson’s Correlation (Numerical Data)



EDA Feature Engineering

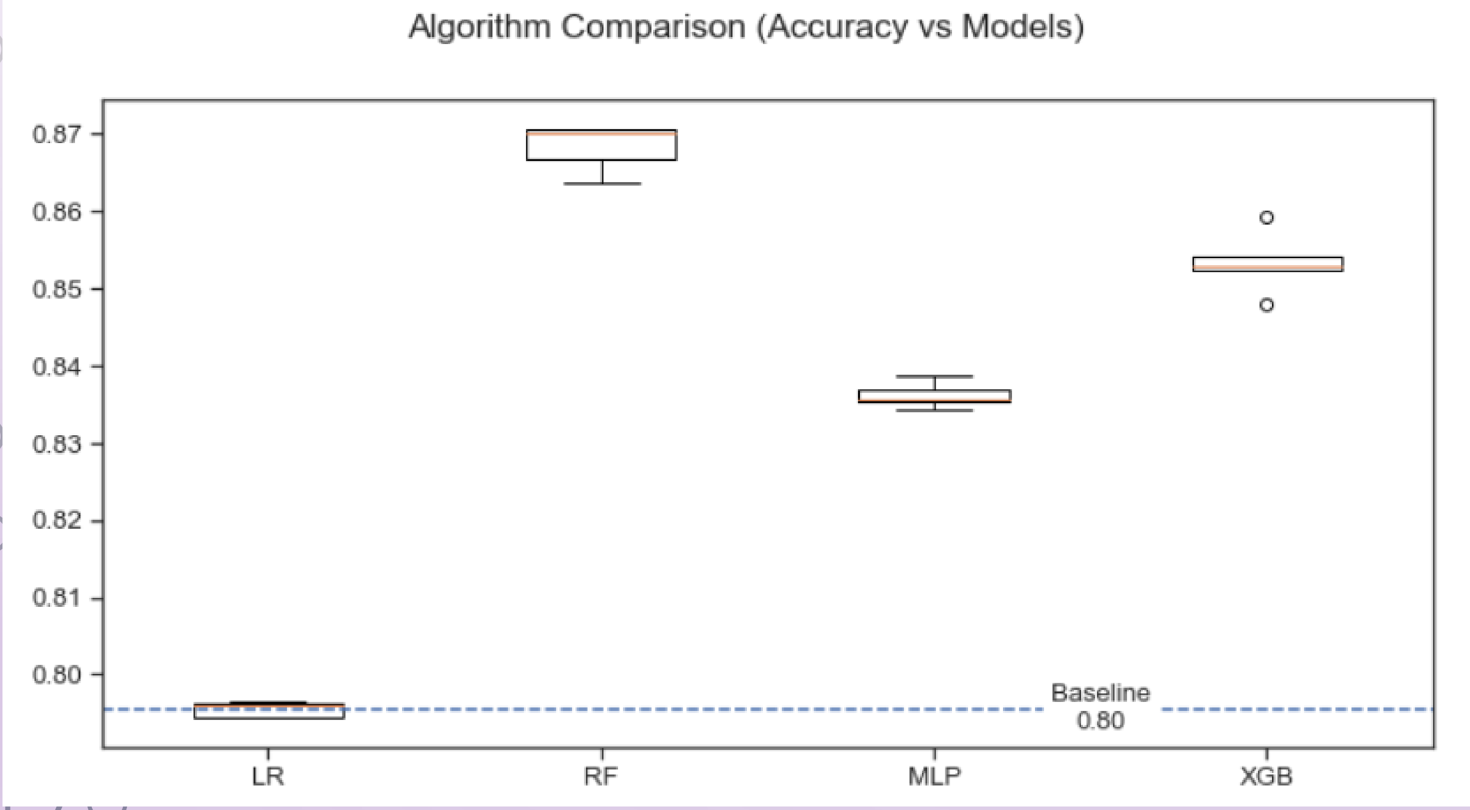
Distribution



Outliers & Skewness : Imputation & Square Root Transform method

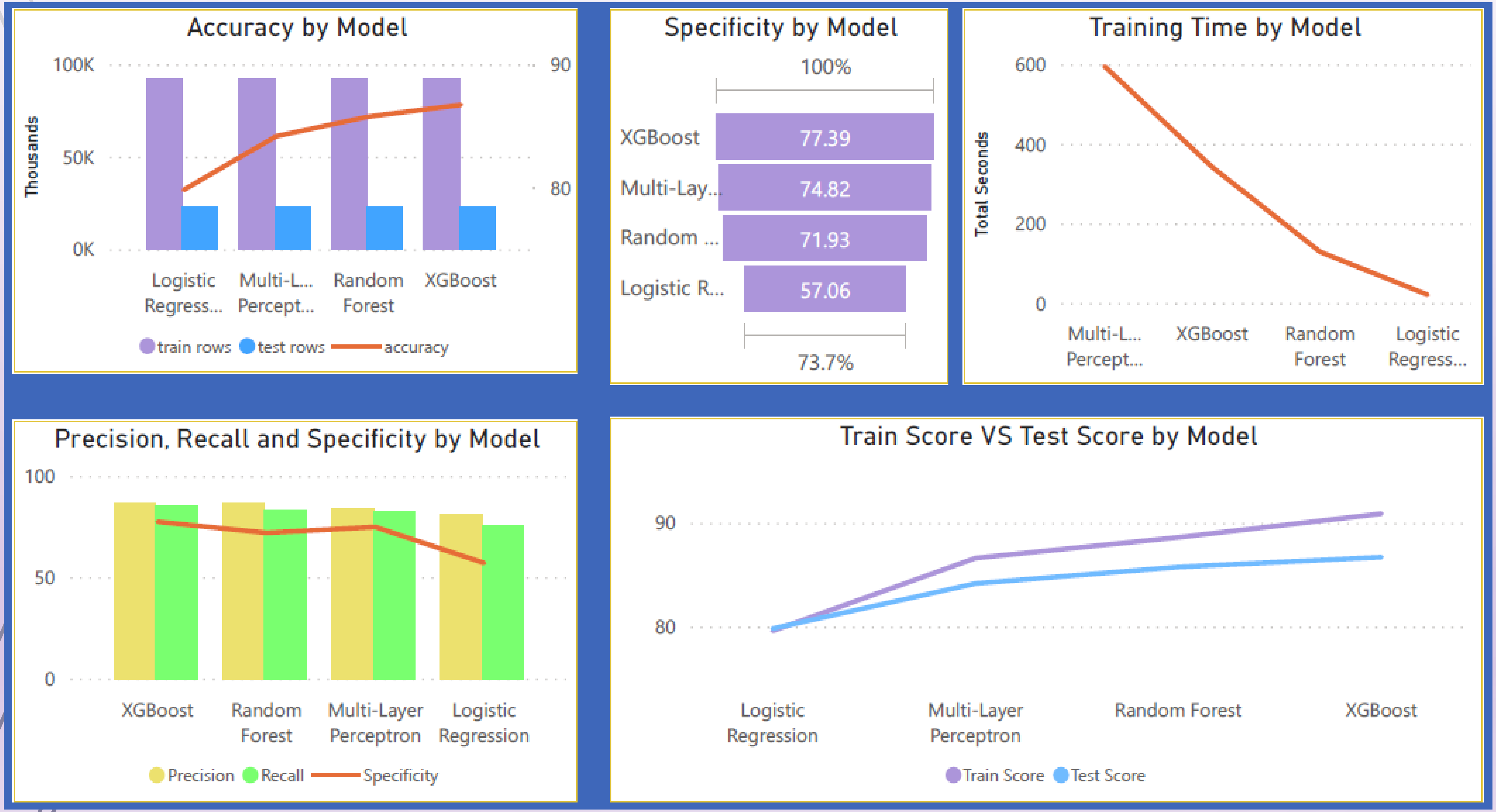
Machine Learning Models Training & Evaluation

Models Comparison (Basic Parameters)



Model	Accuracy (%)
Logistic Regression (LR)	79.55
Random Forest (RF)	86.44
Multilayer Perceptron (MLP)	83.62
eXtreme Gradient Boost (XGBoost)	85.33

Audit Log Power BI



Machine Learning Models Training & Evaluation

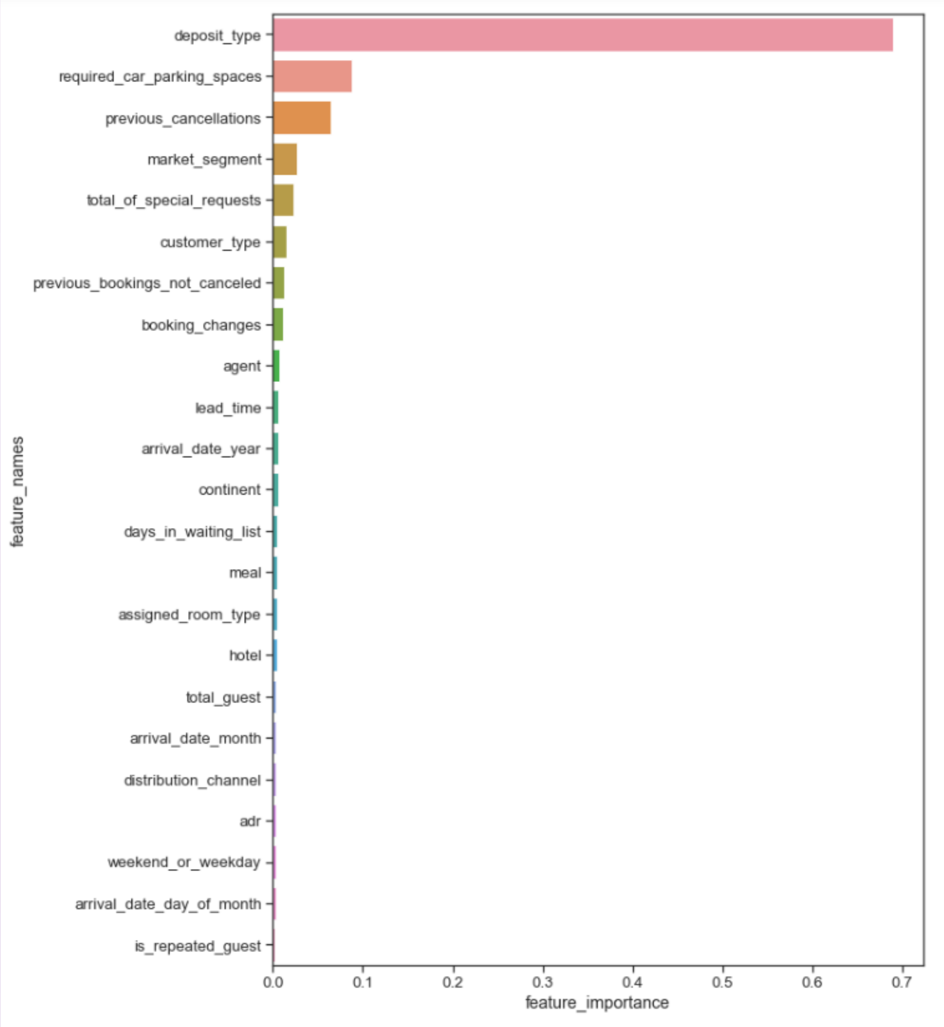
Models Comparison (Hyperparameter Fine Tuning)

Predict Target : is_cancelled : 1
(True Negative)

Model	Accuracy (%)	Specificity (%)
Logistic Regression (LR)	79.82	57.06
Random Forest (RF)	85.76	71.93
Multilayer Perceptron (MLP)	84.17	74.82
eXtreme Gradient Boost (XGBoost)	86.71	77.39

Feature Importance

Final Model : eXtreme Gradient Boost



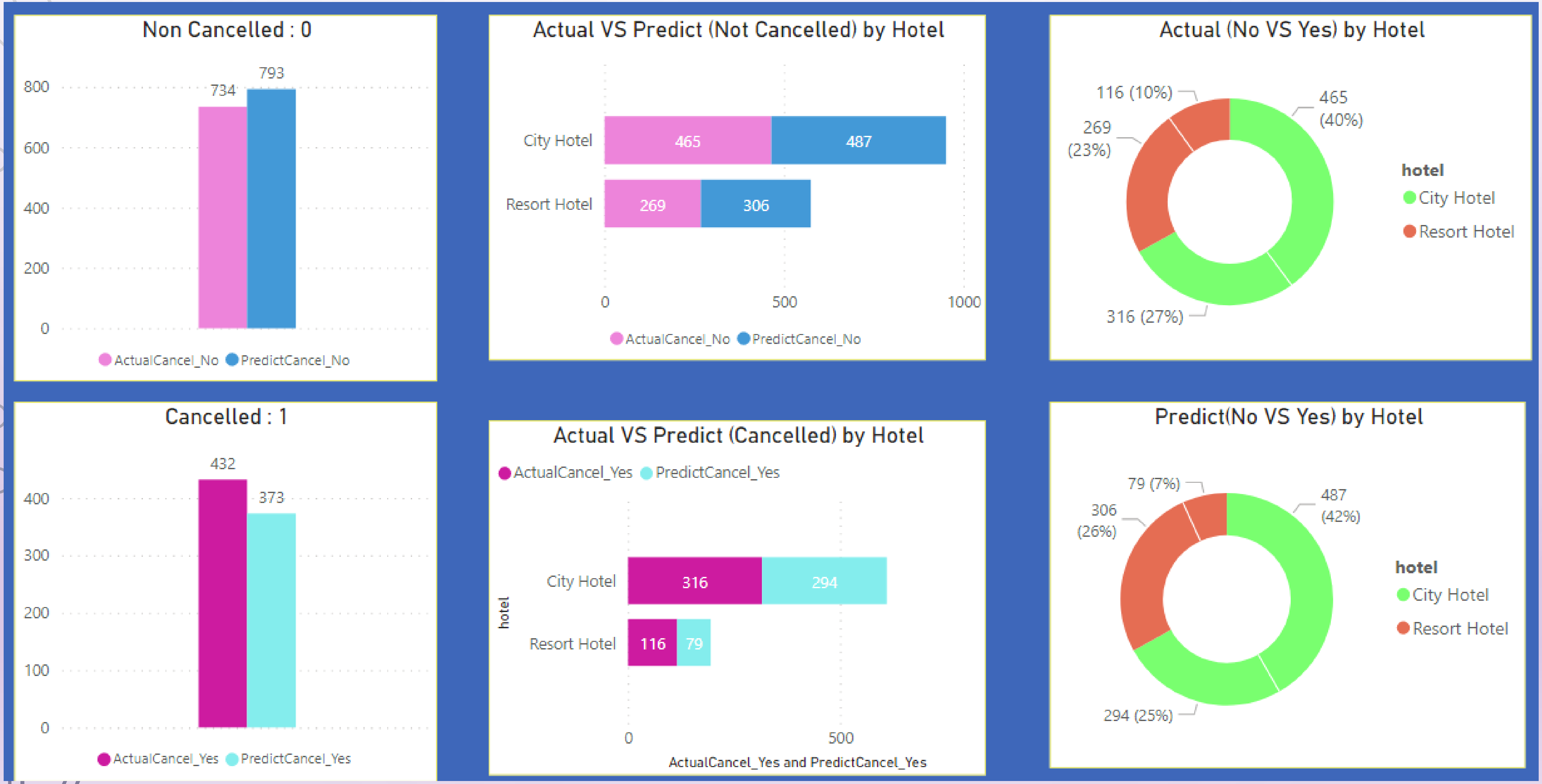
Hyperparameter Fine Tuning with CV = 5

n_estimator = 1300
max_depth = 14
gamma = 4
colsample_bytree = 1
learning_rate = 0.1

Train Score : 90.9 %
Test Score : 86.71

→ No Overfitting

Actual vs Predict (Cancellation Status) Power BI



Conclusions

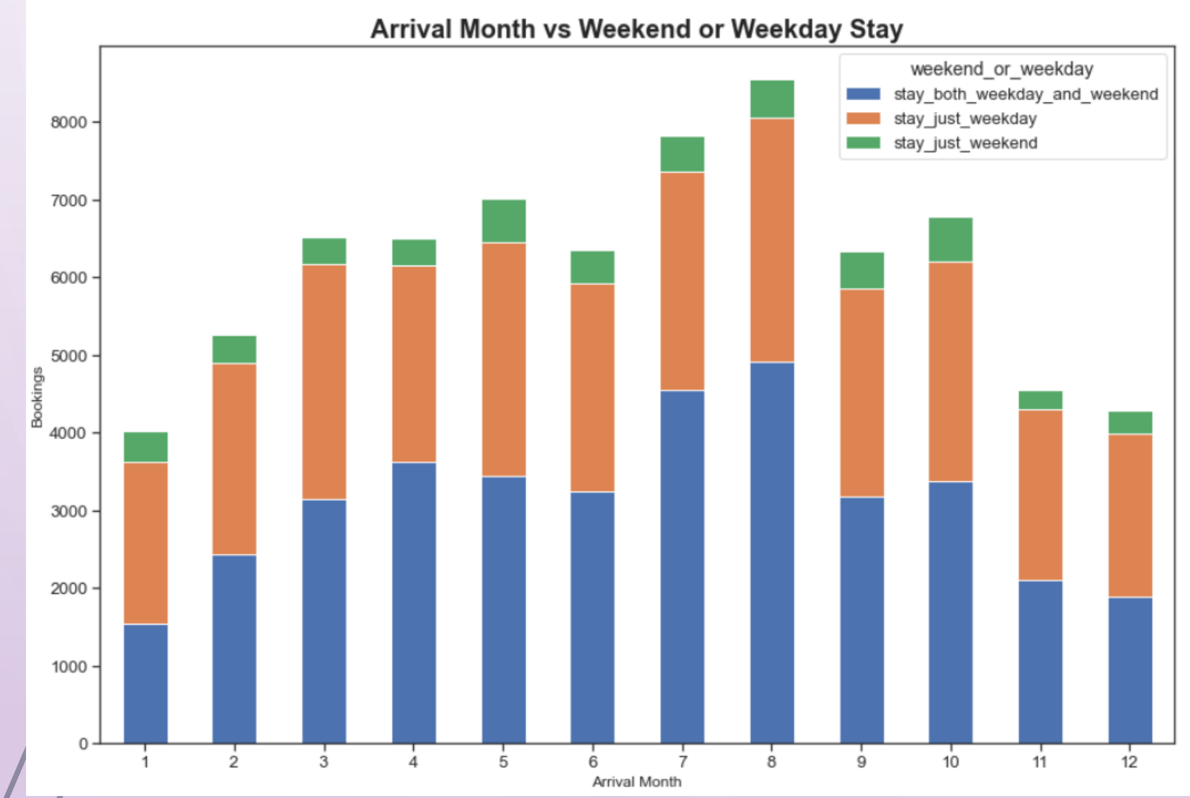
- eXtreme Gradient Boost is the best model to make prediction in this project

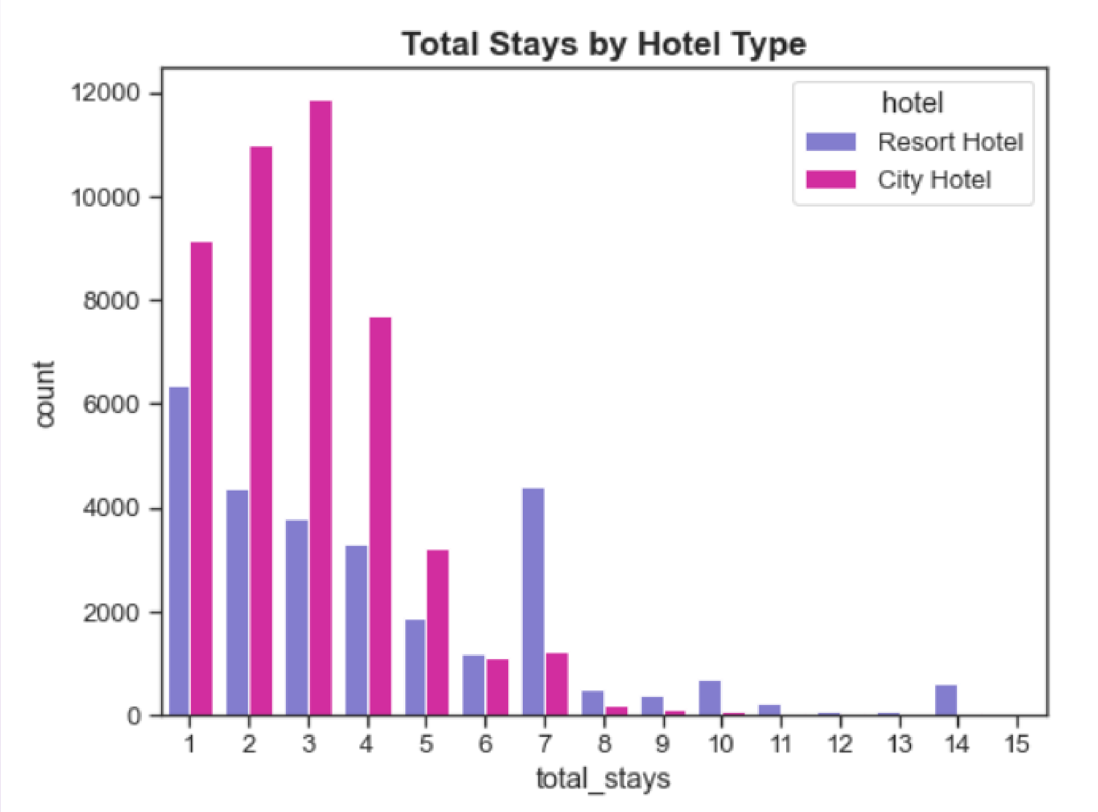
- Almost 37% of bookings were cancelled.

- It can help the Hotel staff to contact clients if the model predict "will cancel" with early notification, so the hotel can have more time to resell the room.

- Or perhaps approach the client in a way to make them feel special and keep their reservation

Interesting Insight





Future Opportunities

- Applying Deep Learning Model for prediction
- Perform more through EDA