Purwadhika
Startup and Coding School

HOTEL BOOKINGS CANCELLATIONS PREDICTION

Final Project Purwadhika
Iota Integer Team

Ferdi Endinanda Fuaadi
James Nathanael Nainggolan

OUTLINE

01

INTRODUCTION

About us, workflow, problem and goal

02

DATA PREPROCESSING

Data understanding, Data cleaning, and EDA

03

MACHINE LEARNING

Feature selection, Feature engineering, Modelling

04

SOLUTION

The results

05

RECOMMENDATION & CONCLUSION

Insights and what can we provide

01

INTRODUCTION

OUR TEAM

Ferdi Endinanda

Project Geological Engineer

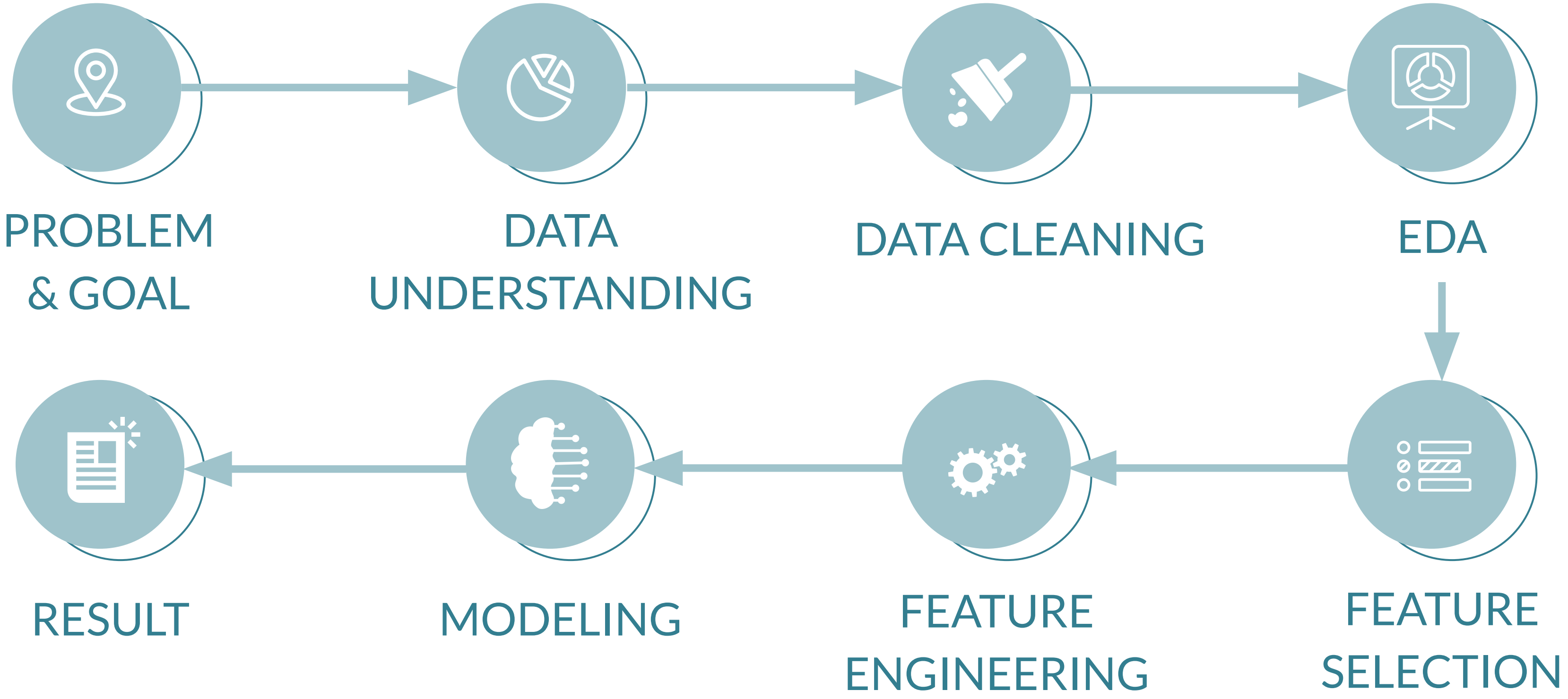
ferdiendinanda@gmail.com
https://github.com/ferdiendinanda

James Nainggolan

Aircraft Maintenance Engineer

jmsxngl@gmail.com
https://github.com/jmsxngl

WORKFLOW



PROBLEM & GOAL

DATA UNDERSTANDING

DATA CLEANING

EDA

FEATURE SELECTION

FEATURE ENGINEERING

MODELING

RESULT

THE PROBLEM

In this project we position ourselves as Data Scientist consulting for a client which is a hotel management that is facing a problem:

Experiencing an increase in booking cancellations especially in high peak seasons and they would like to be able to prioritize among those who the hotel thinks are likely to be cancelling.

THE GOAL

We are assigned to **predict booking cancellations based on customer behavior** to distinct them between the ones that are not canceling.

02

DATA PREPROCESSING

DATA UNDERSTANDING

The data is taken from kaggle https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

The data contains hotel bookings due to arrive **between 2015 and 2017** and most of this data is taken in **Portugal**

Booking data shape is **119390** rows and **32** features

Each record of data represents information related to ordering transactions that occur. **Unbalanced dataset** (is_canceled)

Useful Features :

- **hotel :** H1 (Resort Hotel) or H2 (City Hotel)
- **is_canceled** : Value indicating if the booking was canceled (1) or not (0)
- **lead_time** : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
- **adults** : Number of adults
- **children** : Number of children
- **babies** : Number of babies
- **meal** : Type of meal booked. Categories are presented in standard hospitality meal package Undefined/SC – no meal
- **market_segment** : Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- **distribution_channel** : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- **booking_changes** : Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or out
- **is_repeated_guest :** Value indicating if the booking name was from a repeated guest (1) or not (0)
- **previous_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking
- **assigned_room_type :** Code for the type of room assigned to the booking.Code is presented instead of designation for anonymity reasons
- **deposit_type :** Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories No
- **customer_type** : Type of booking, assuming one of four categories Transient - Transient-Party - Contract - Group
- **adr :** Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
- **required_car_parking_spaces** : Number of car parking spaces required by the customer
- **total_of_special_requests** : Number of special requests made by the customer (e.g. twin bed or high floor)

DATA CLEANING

01

Check NaN values → The 'company' feature has 94% NaN values and considered to be dropped, null in children filled with 0.

02

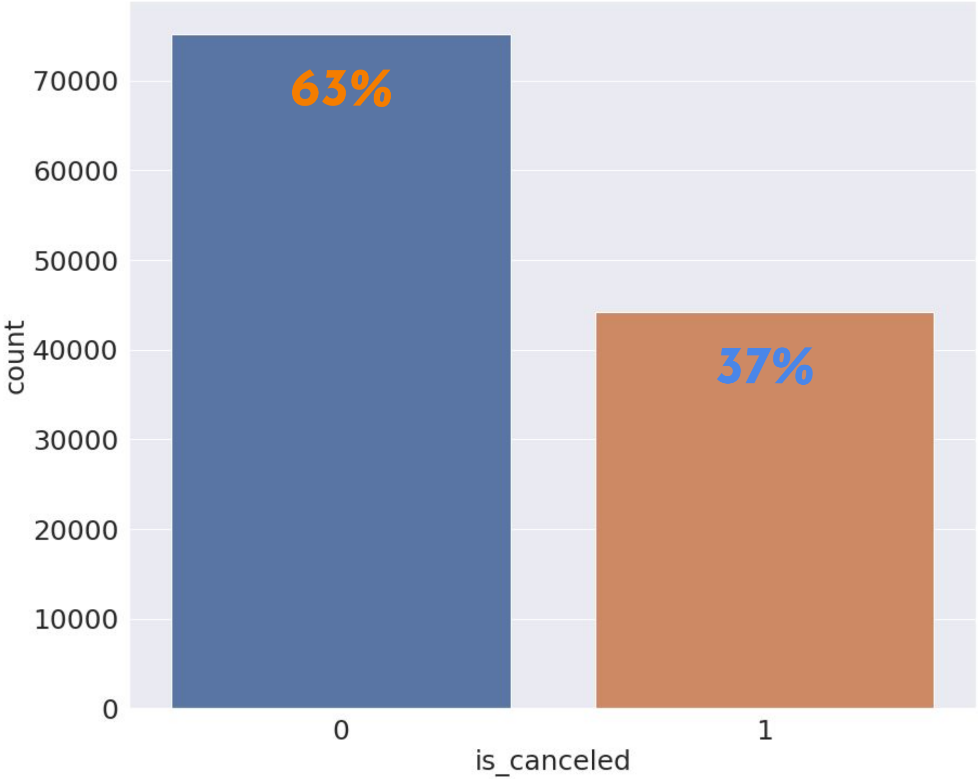
Reformat reservation status date to date and time format, also the arrival date into a new date and time feature.

03

Spotting undefined, negatives, outliers, and irrational values and changing them into mean or mode of the data. The features are meals, lead time, adr, market segment, distribution channel,

EXPLORATORY DATA ANALYSIS

Data Imbalance



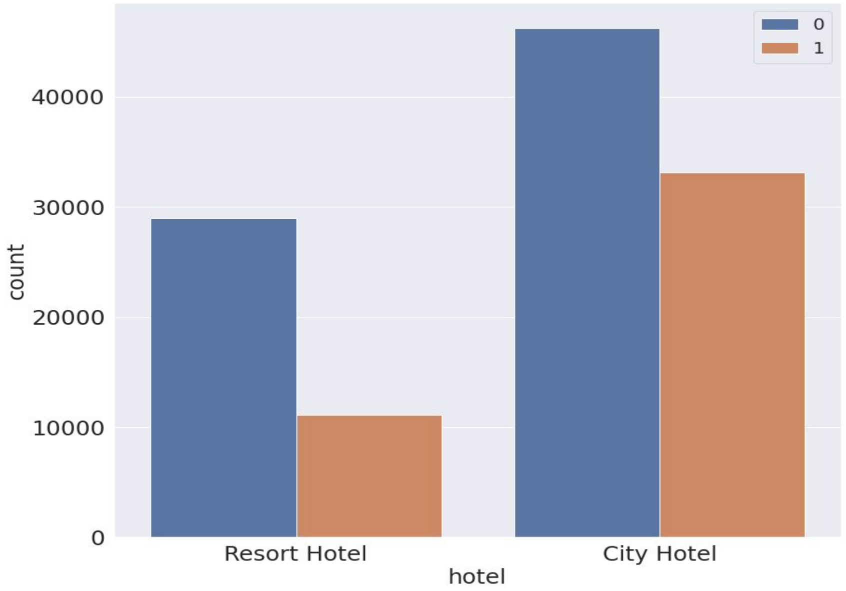
The total of not cancel (0)

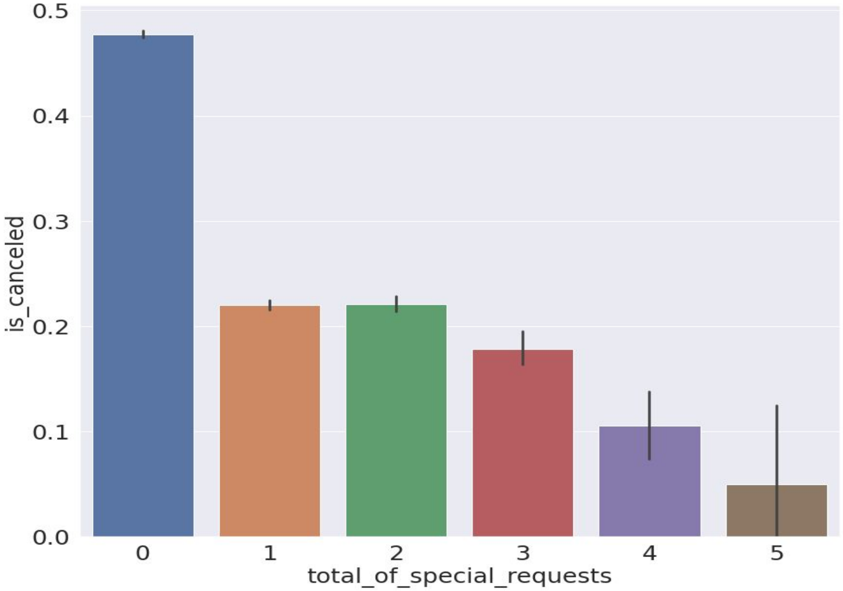
is much more than cancel (1).

This indicates data is **imbalanced**

EXPLORATORY DATA ANALYSIS

Hotel





Hotel

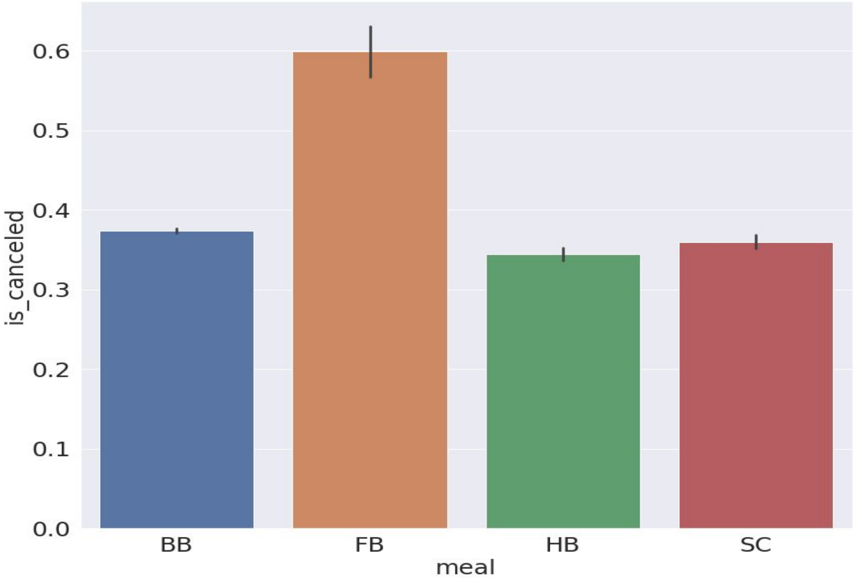
Canceled mostly happened in city hotel rather than resort hotel giving insight people tend to cancel while it is at the city that has many option of places to stay.

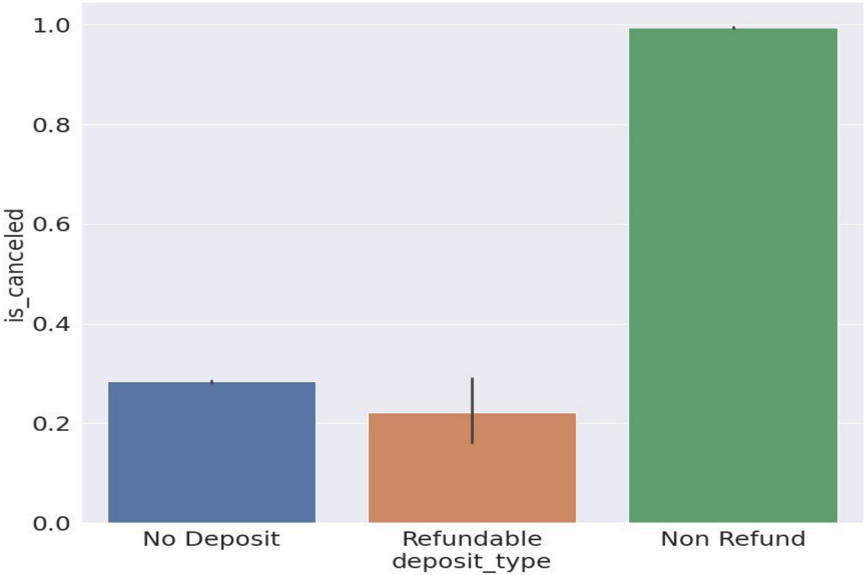
Special Request

Less special requests is much more likely to cancel than high requests. This is because the customer that has special request is more likely to have a well planned trip rather than customers without special requests.

EXPLORATORY DATA ANALYSIS

Hotel





Meal

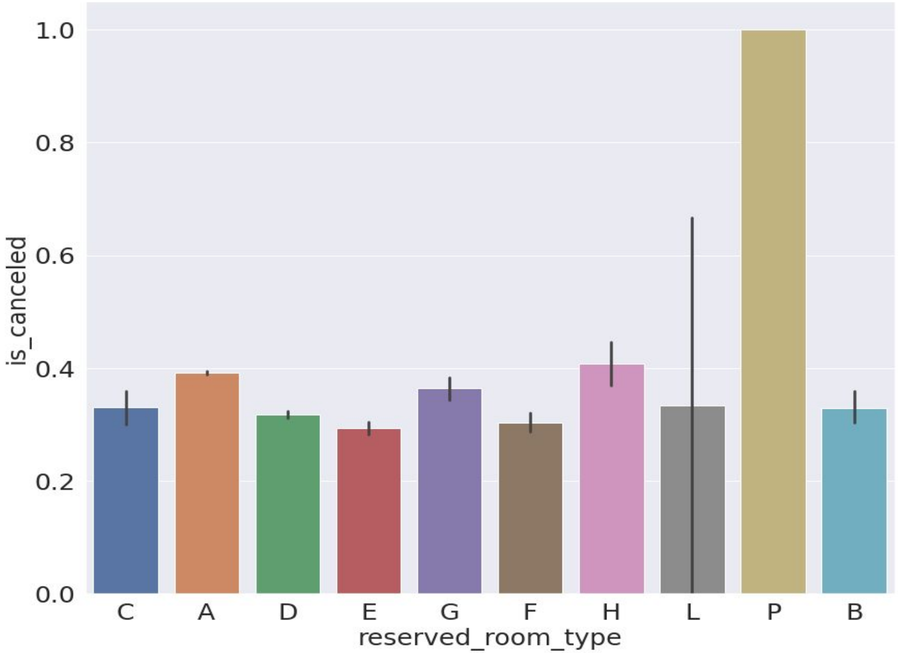
People with meal Full Breakfast tends to cancel than other types of meal order. Perhaps this is related to additional price which is more expensive than other type of meal.

Deposit Type

Non-Refund is likely to cancel all the time which is quite unusual. Usually non-refundable types of booking are the ones that are made very close to arrival date, impulsive bookings, but looking at the data the time customer made the booking and arrive to the hotel (lead_time) is quite long, perhaps there is another reason. Perhaps the terms from the hotel with non-refundable option made customers uncomfortable/confusing.

EXPLORATORY DATA ANALYSIS

Hotel

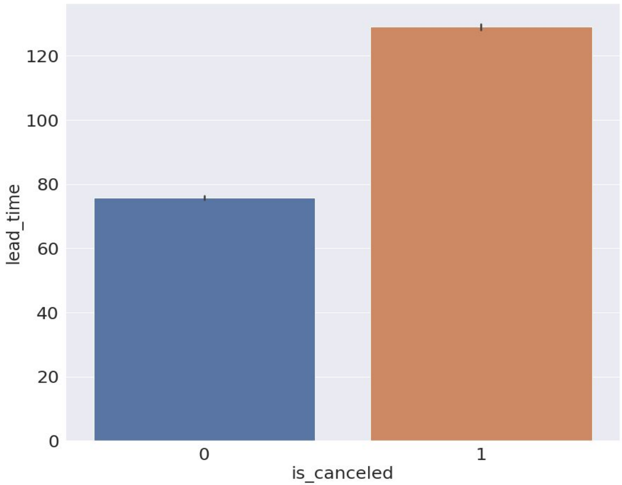


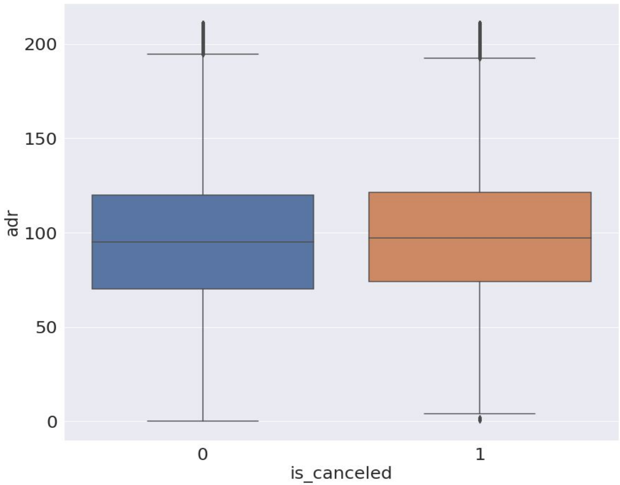
Room Type

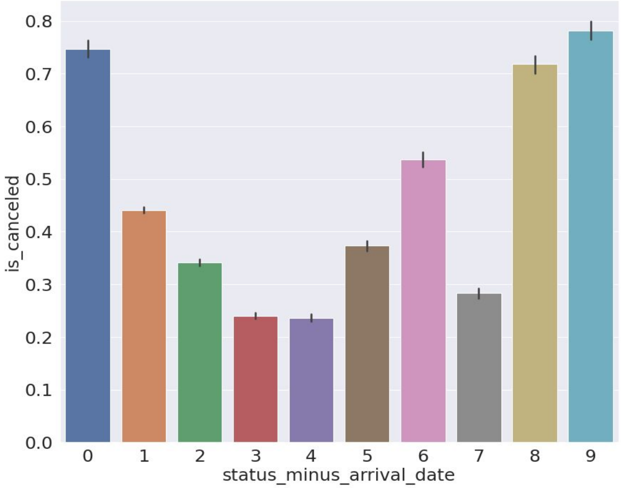
Most cancelation was made when the booking was reserved room type P. This is not very clear of why cancelation took place on these types of room, perhaps additional detail on hotel description of the room would give an answer of why these types of room are always being canceled. Probably the room type has bad review or the specs is not what customer hoped for.

EXPLORATORY DATA ANALYSIS

Time







Lead Time

When the booking was made very far in advance, people tend to cancel rather then keeping it because the costumer plans is more likely to change the longer it gets.

ADR

Both canceled and not canceled has the same price range around 60-120 USD. But we can still see cancelation happens at slightly more expensive rates hotels.

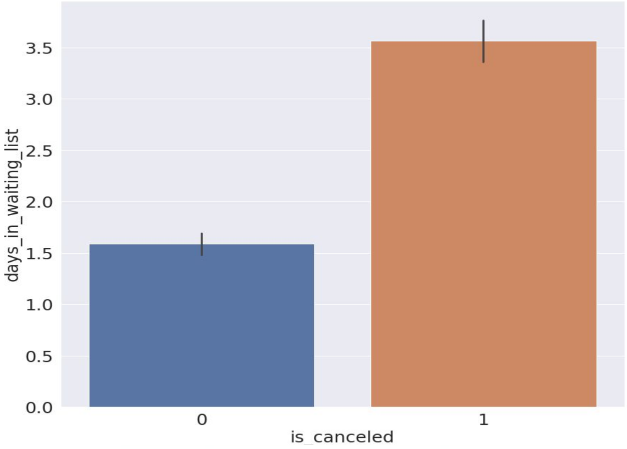
Status minus Arrival Date

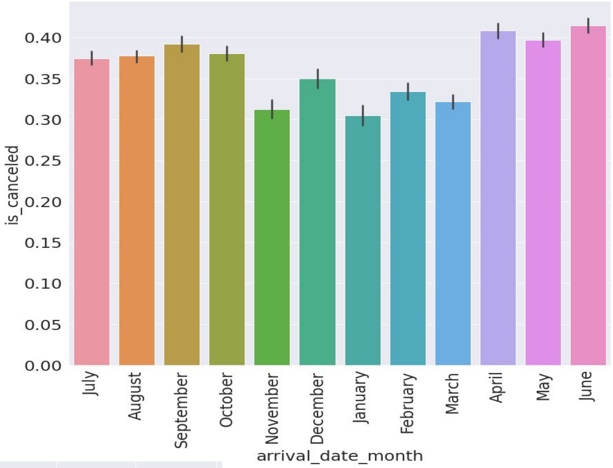
Customers tend to cancel close to the arrival date (0 days) due to sudden events/no show or a couple days before (9 days prior) due to change of plans.

Iota Integer Team

EXPLORATORY DATA ANALYSIS

Time



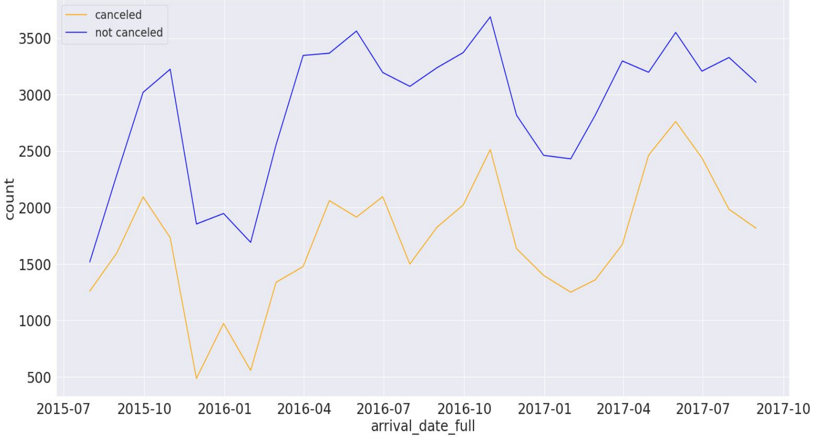


Days in Waiting List

This makes sense as customers do not like to be in waiting list for their booking even just for a day, they tend to cancel with longer wait times because unsecured
bookings is very risky for their trip.

Month Arrival Date

We can see in the graph in months that is holiday season is more likely to be canceled, this is probably due to the much more variety of customers that books the hotel such as families or groups of people that are on a holiday which can have change of plans, while in non vacation season bookings are probably made by employees for business trips that is a much more fixed booking.

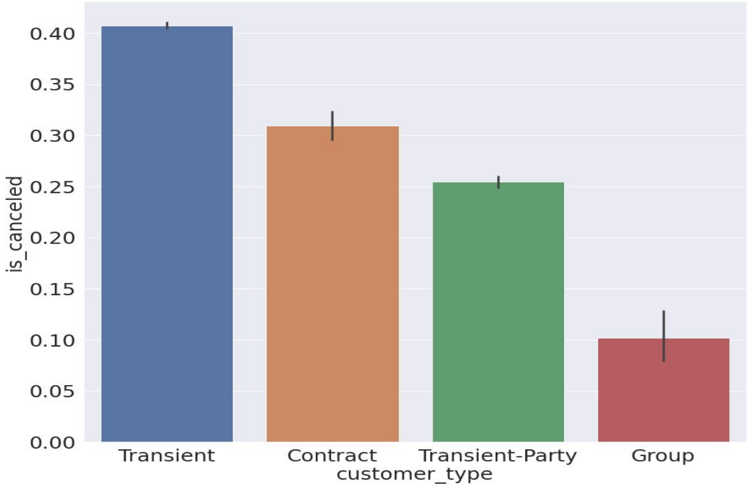


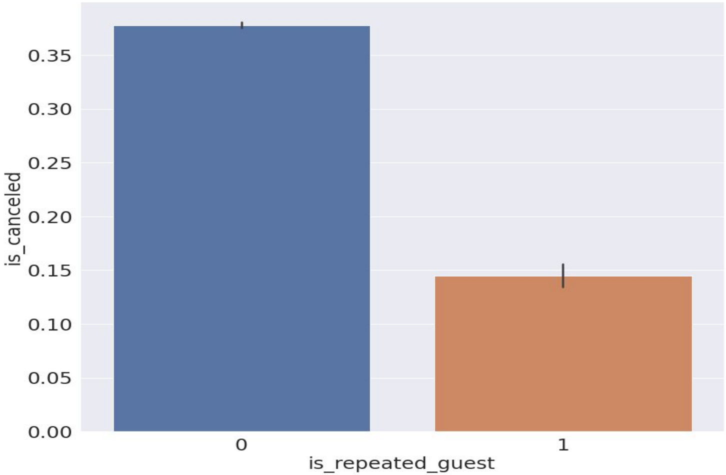
Arrival Date

From the complete set 2015-2017 data of booking arrival dates, it shows that bookings are much more made for seasonal months which are in the middle and at the end of the year. These peak seasons are the important times to predict cancelation especially in this busy moments when we don't want to miss predict giving customers a bad experience, keeping in mind cancelation are also likely to be made in these time of the year from the previous graph.

EXPLORATORY DATA ANALYSIS

Customer





Customer Type

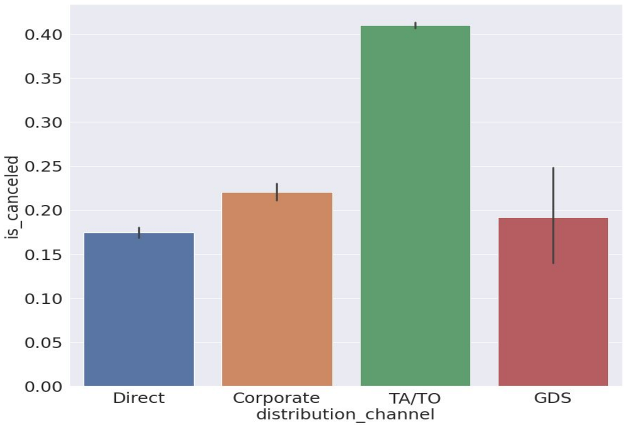
Transient type customer (people that book only for short time) are the most likely to cancel, 10 percent higher than the second highest which is contract. Transient customers might have other options for their short stay like staying at a friend, while transient group cannot easily do so. The group in this feature is different than the group in market segment. Here group is interpreted as booking lots of rooms like around 10 rooms for company meetings for example, meaning it is unlikely to be canceled.

Repeated Guest

Not repeated guests tends to cancel their booking which makes sense as it is their first stay and they are not a regular in the particular hotel, so nothing to lose if canceling. While regulars might have a special relationship with the hotel such as points system that they already acquire that can be traded for freebies.

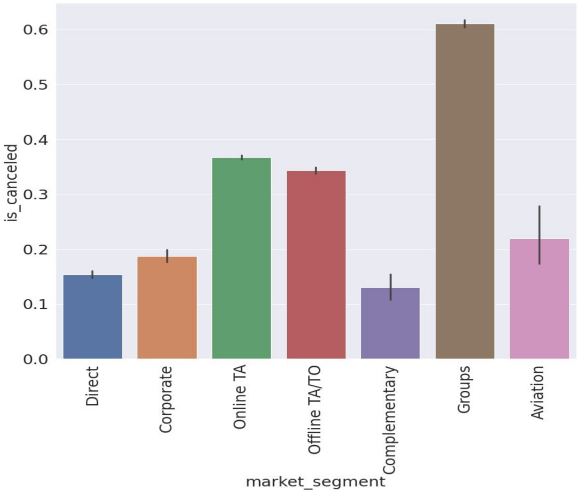
EXPLORATORY DATA ANALYSIS

Customer



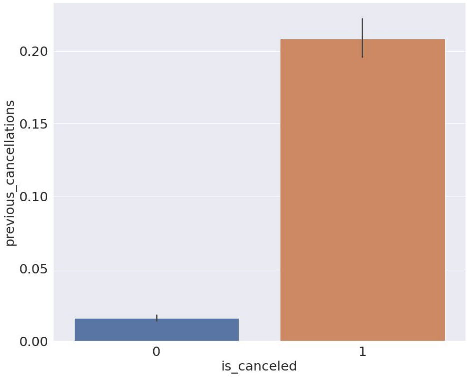
Distribution Channel

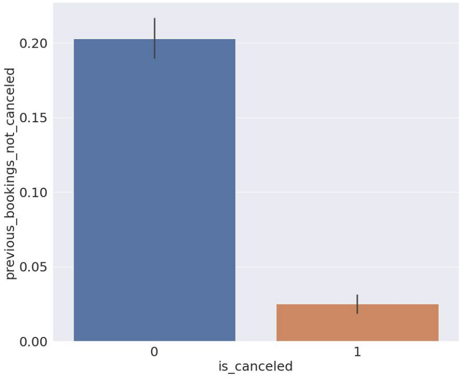
Customer bookings that was advertised or distributed through travel agent/tour operators tend to be canceled, perhaps because there were changes of plan or it is not in accordance with the customers expectation of visit.



Market segment

Group market segment tends to cancel. The group here is interpreted as customers that just comes together. They tend to cancel probably because in a group if one or two person cannot go, the whole group cancels the booking.



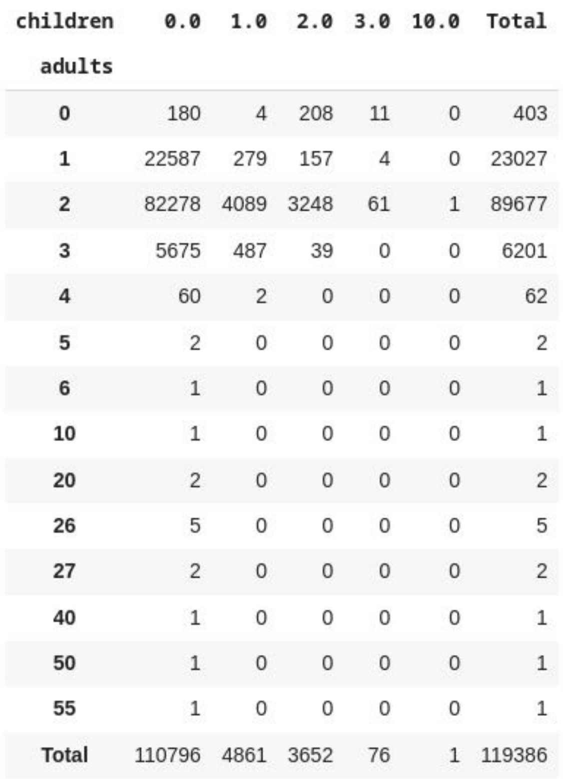


Previous cancelation and not canceled

Customers that already cancel will tend to cancel again. The same tendency for customers that not cancel their previous bookings, it unlikely they will cancel their booking.

EXPLORATORY DATA ANALYSIS

Table (Customer)



children	0.0	1.0	2.0	3.0	10.0	Total
adults						
0	180	4	208	11	0	403
1	22587	279	157	4	0	23027
2	82278	4089	3248	61	1	89677
3	5675	487	39	0	0	6201
4	60	2	0	0	0	62
5	2	0	0	0	0	2
6	1	0	0	0	0	1
10	1	0	0	0	0	1
20	2	0	0	0	0	2
26	5	0	0	0	0	5
27	2	0	0	0	0	2
40	1	0	0	0	0	1
50	1	0	0	0	0	1
55	1	0	0	0	0	1
Total	110796	4861	3652	76	1	119386

Guests

Here we can see the most bookings was made by 2 adults, followed by 1 adult, group of 3 three adults, and family with 1 to 2 child. We can conclude cancelation occurred mostly from the bookings that were made by adults only without children.

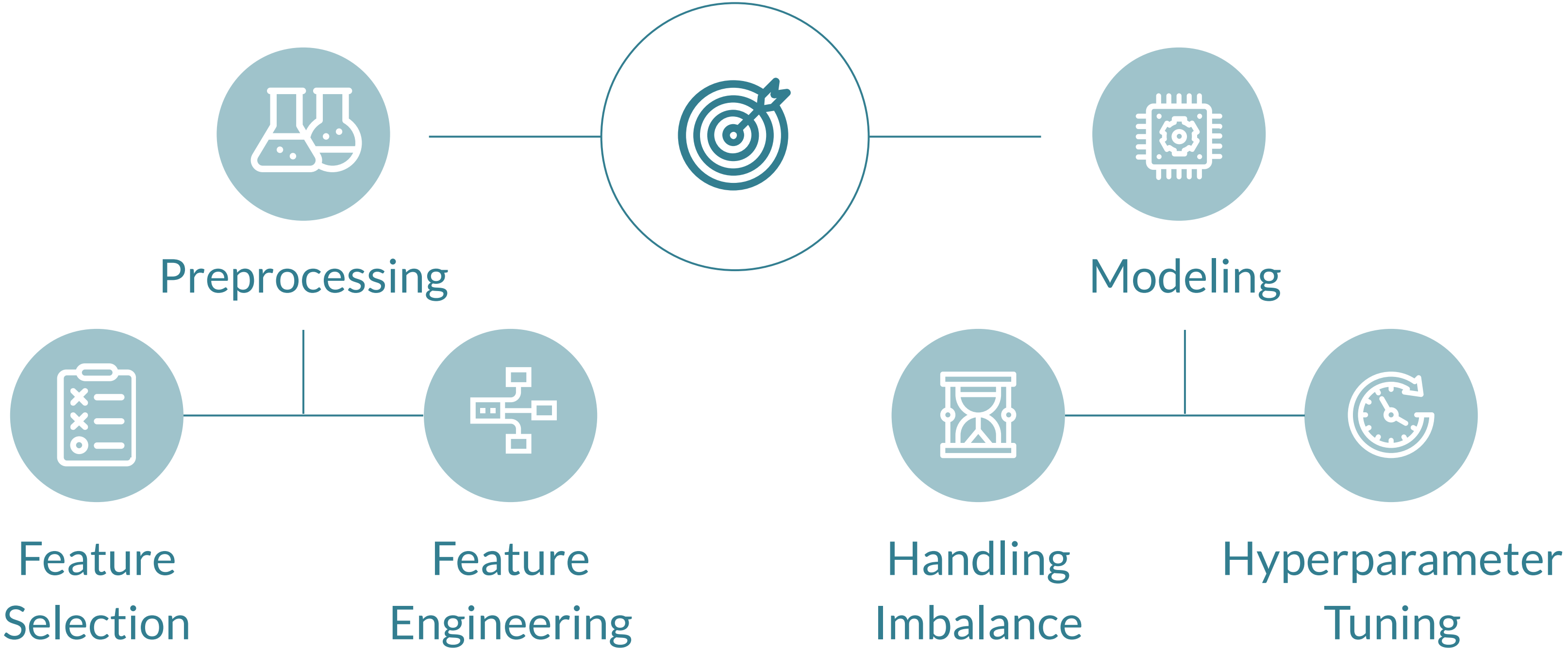
03

MACHINE LEARNING

STRATEGY

GOALS

Find the highest precision score and stable model



Preprocessing

Modeling

Feature Selection

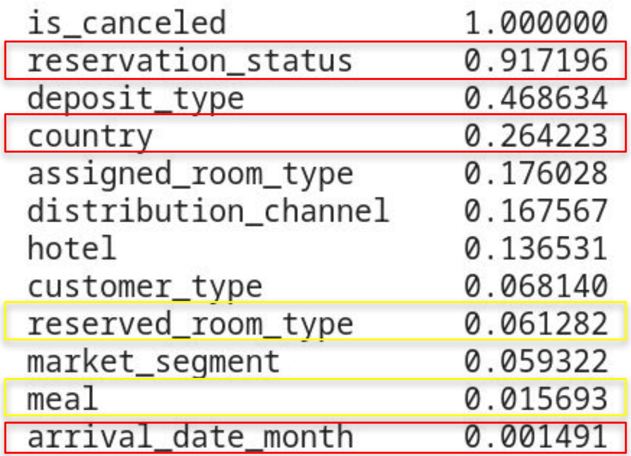
Feature Engineering

Handling Imbalance

Hyperparameter Tuning

FEATURE SELECTION

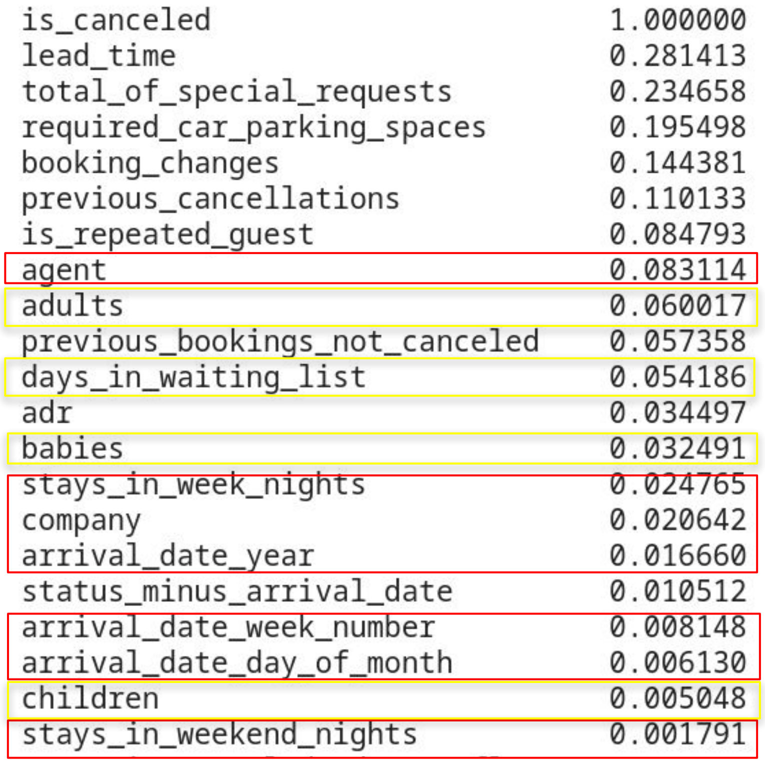
Categorical Feature



```
is_canceled                 1.000000
reservation_status          0.917196
deposit_type                0.468634
country                     0.264223
assigned_room_type          0.176028
distribution_channel        0.167567
hotel                       0.136531
customer_type               0.068140
reserved_room_type          0.061282
market_segment              0.059322
meal                        0.015693
arrival_date_month          0.001491
```

Feature selection with correlation is a quantified approach to help us decide which features are not important and best be dropped.

Numerical Feature



```
is_canceled                        1.000000
lead_time                          0.281413
total_of_special_requests          0.234658
required_car_parking_spaces        0.195498
booking_changes                    0.144381
previous_cancellations             0.110133
is_repeated_guest                  0.084793
agent                              0.083114
adults                             0.060017
previous_bookings_not_canceled     0.057358
days_in_waiting_list               0.054186
adr                                0.034497
babies                             0.032491
stays_in_week_nights               0.024765
company                            0.020642
arrival_date_year                  0.016660
status_minus_arrival_date          0.010512
arrival_date_week_number           0.008148
arrival_date_day_of_month          0.006130
children                           0.005048
stays_in_weekend_nights            0.001791
```

🟥 dropped

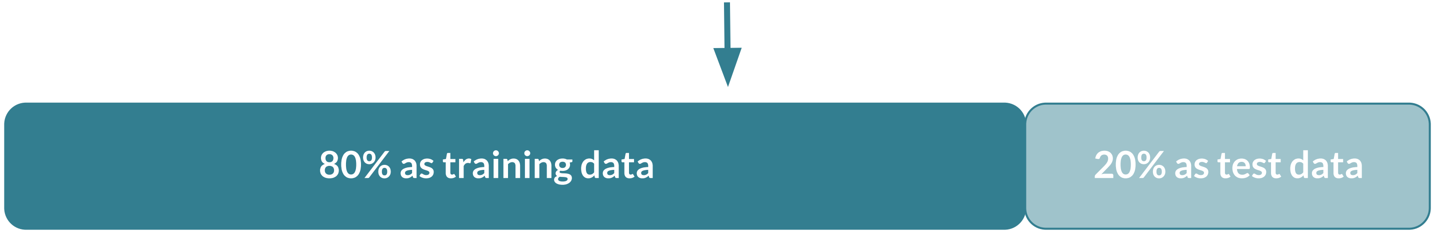
🟨 no significant effect

FEATURE ENGINEERING

Several treatments:

1. Remove unnecessary features in modeling to prevent leaking
2. Encoding categorical data as additional features for modeling
3. Splitting data between train and test

Target	**: is_canceled**
OneHotEncoder	**: hotel, meal, distribution_channel, deposit_type, customer_type**
BinaryEncoder	**: country, market_segment, reserved_room_type**
Standard scaler	**: adults**
Passthrough	**: babies, booking_changes, total_of_special_requests**



MODELING

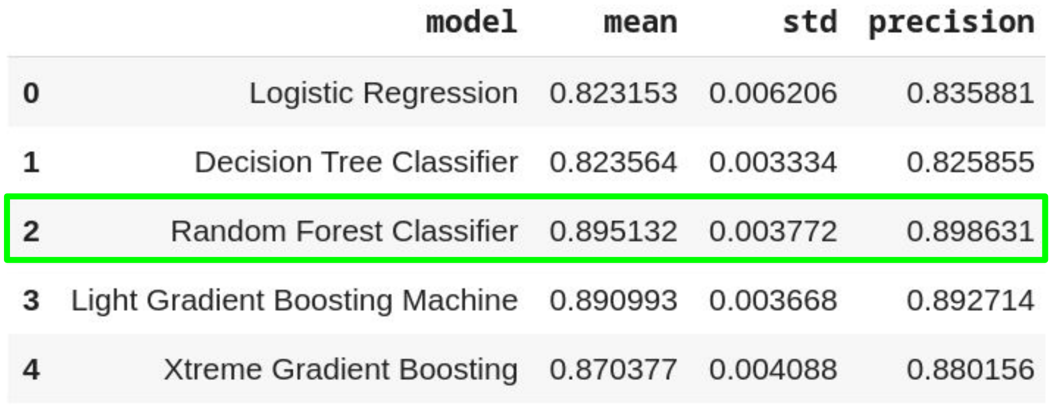
Machine Learning Algorithms

Base Model

- Logistic Regression
- Decision Tree Classifier

Ensemble Model

- Random Forest Classifier
- Light Gradient Boosting Machine
- eXtreme Gradient Boosting



	model	mean	std	precision
0	Logistic Regression	0.823153	0.006206	0.835881
1	Decision Tree Classifier	0.823564	0.003334	0.825855
2	Random Forest Classifier	0.895132	0.003772	0.898631
3	Light Gradient Boosting Machine	0.890993	0.003668	0.892714
4	Xtreme Gradient Boosting	0.870377	0.004088	0.880156

HANDLING IMBALANCE

The methods for handling imbalance:

- Random Under Sampling
- Random Over Sampling

Random Forest Classifier Precision Score			
model	mean	std	precision
Base	0.895132	0.003772	0.898631
Random Under Sampling	0.902095	0.003384	0.847828
Random Over Sampling	0.929119	0.002500	0.883394

highest precision score
and the most stable

HYPERPARAMETER TUNING

Tuning **Random Forest Model**

1. Expected characteristics with a high precision score
2. Stable value on the base model and during the balancing process

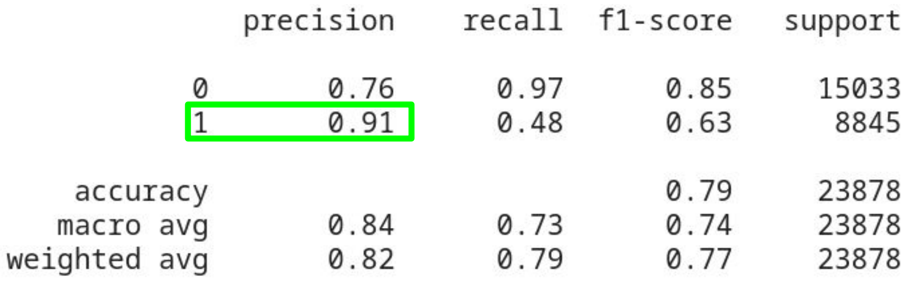
Random Forest Classifier Precision Score	
Random Forest Before Tuning	0.883394
Random Forest After Tuning	**0.912669**

Score significant increasing. Tuned model will be used for prediction

04

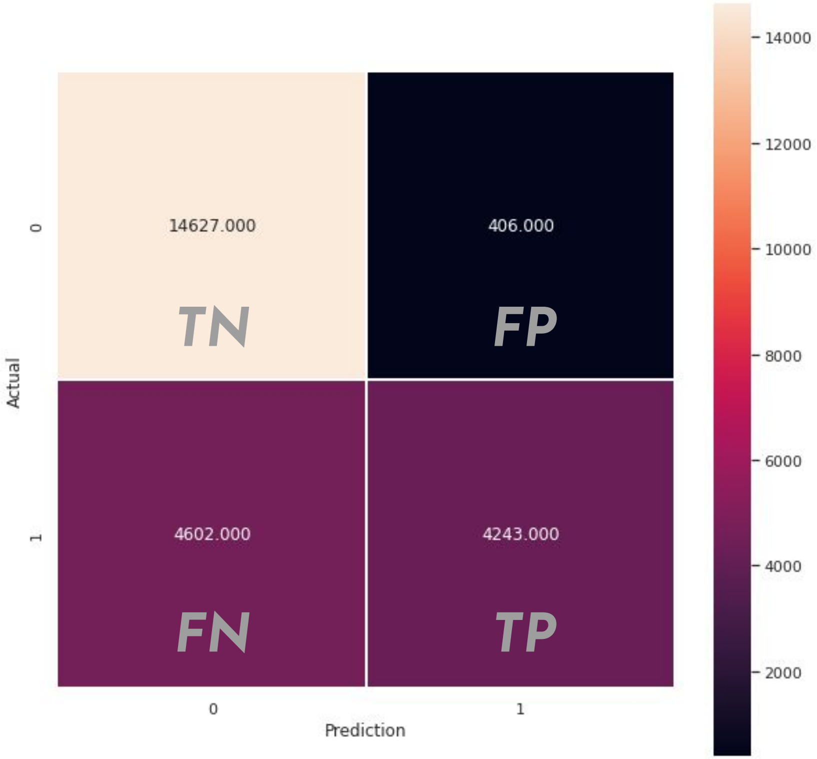
SOLUTION

RESULT

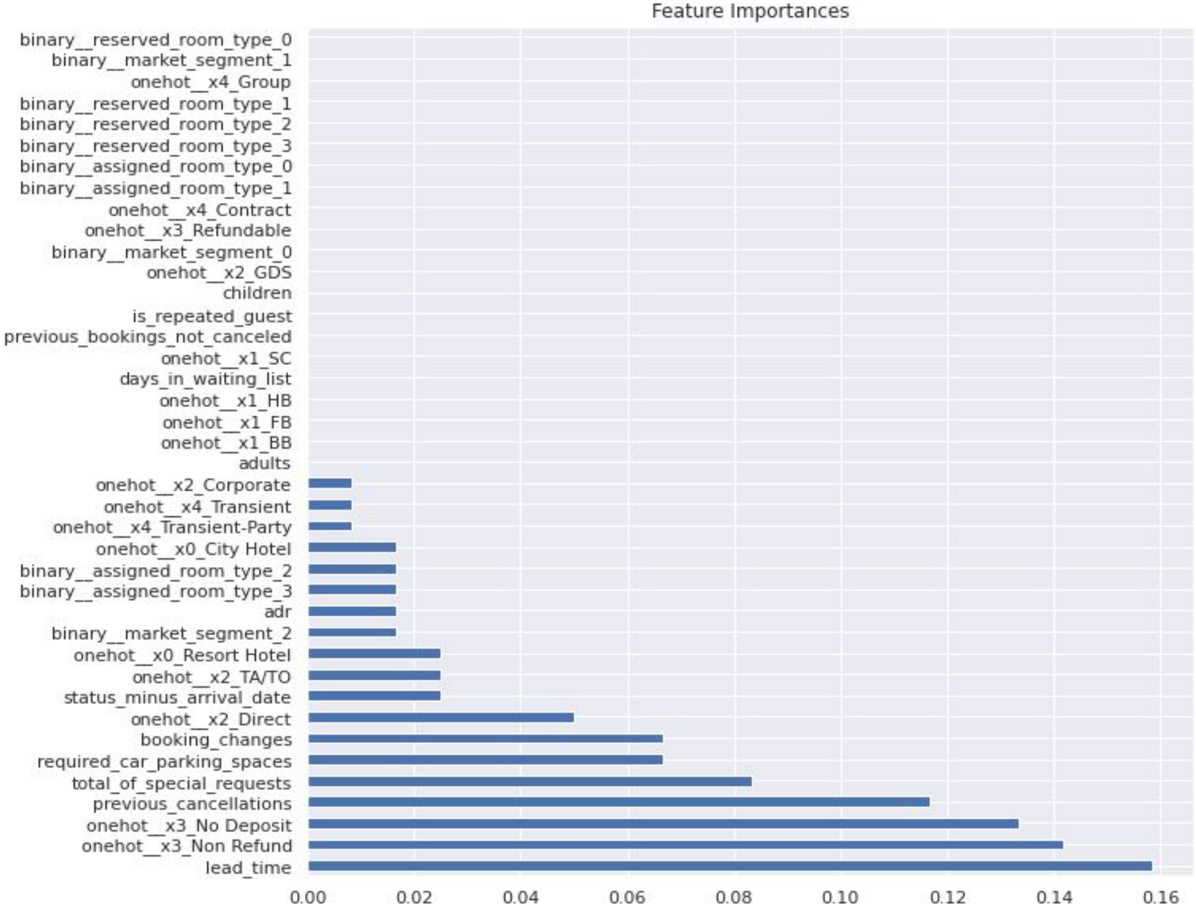


```
              precision    recall  f1-score   support

           0       0.76      0.97      0.85     15033
           1       0.91      0.48      0.63      8845

    accuracy                           0.79     23878
   macro avg       0.84      0.73      0.74     23878
weighted avg       0.82      0.79      0.77     23878
```

Random Forest Precision Score	
Base Model	0.898631
Balanced Model	0.883394
Balanced & Tuned Model	0.912669



FEATURE IMPORTANCE



Not important:

1. Meal type
2. Days in waiting list
3. Guest
4. Reserved room type
5. Assigned room type

05

RECOMMENDATION & CONCLUSION

CONCLUSION

our model could make hotels **avoid a loss of those customers** due to the various reasons of cancelling by knowing who to prioritize **37%**

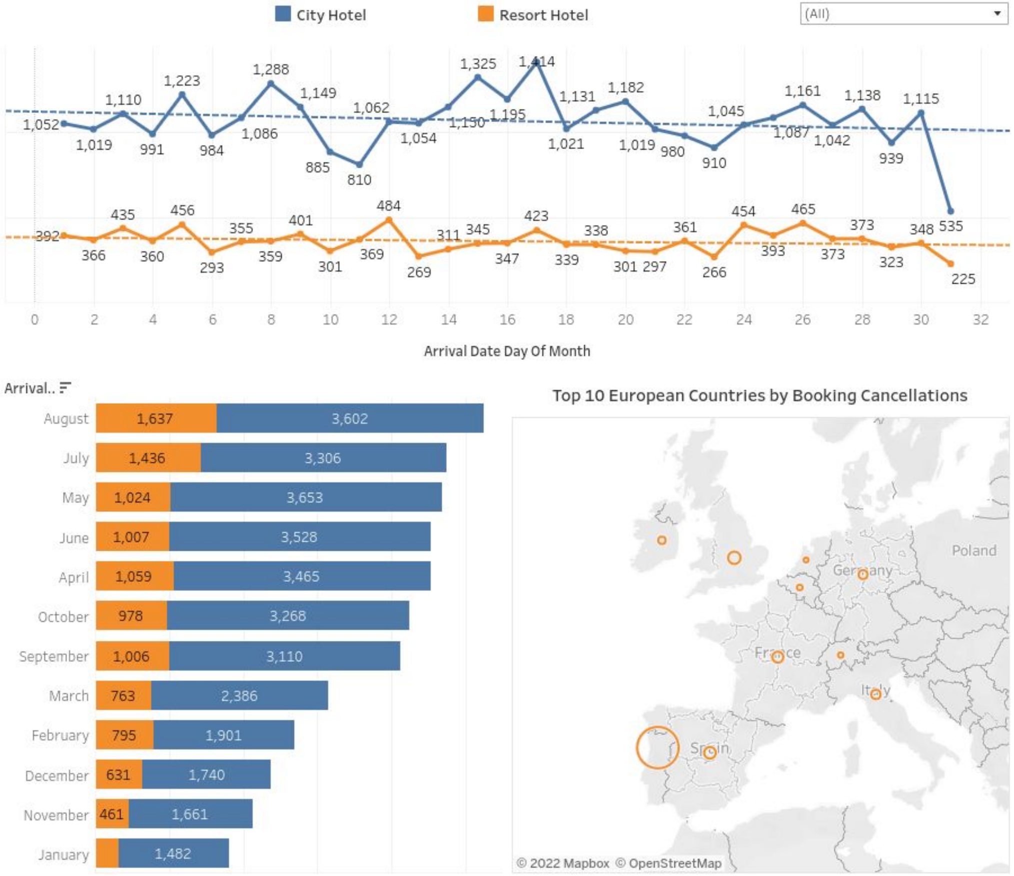
91% prediction success, the hotel can gain up to **4m USD** for the next 2 years

Help the **inventory allocation** and **pricing decision**

RECOMMENDATION

- Hotel regulations and policies for deposit type should be revised
- Promoting saving packages especially in high seasons
- A sharing room (bunk beds) as an alternative option especially for transient customers
- Early detection system has to be properly managed especially in high season
- Offering a form with multiple itinerary choice they can easily pick
- Promotion should be direct to give a more personal approach to customers

TABLEAU DASHBOARD



SCAN ME

https://public.tableau.com/views/hotel_bookings_16580755476150/Dashboard?:language=en-US&:display_count=n&:origin=viz_share_link

Thank You