Hotel Booking Prediction Report

Wirach Leelakiatiwong
August 2020

1. Introduction

In the present day, booking for a nice, beautiful accommodation are just a click away. We can browse through online travel booking platforms, search and made a booking within a minute. This benefit not only for traveler like us, but also for hotel operator around the globe. While hotel operator are enjoying benefit from online travel booking platforms to boost their occupancy rate. There are also some unavoidable drawbacks, some of them may include higher cancellation rate when online travel booking platforms offer book now pay later promotion.

Therefore, in order to maximize revenue per room and minimize loss due to cancellation, hotel operator should be able to predict whether a given reservation are likely to canceled or not? With this information, hotel operator might be able to make an adjustment on overbooking policy during high season to compensate for canceled booking. Moreover, they can utilize this information to manage their workforce more effectively.

1.1. Problem statement

This could be viewed as supervised binary classification problem. With given booking information : number of guests, number of special requirement, agencies and etc., we will try to predict whether a reservation will be confirmed or canceled.

With different baseline classification algorithms (such as Logistic Regression, Random Forest, Gradient-Boosting, Ada-Boost and Support Vector Machine), we'll evaluate them with default scikit-learn configuration and continue to develop most suitable algorithm.

1.2. Metrics

We select classification accuracy as performance evaluation metrics. The model should be able to predict whether a given booking will be canceled with enough accuracy to outperform simple overbooking policy of 20%. As shown in below table, in order to make model feasible its accuracy must be at least 70%

Model Accuracy	Average cancellation rate (%of Total booking)[1]	Expected overbooking policy	Typical overbooking policy[2]	%Improvement
65%	30%	20%	20%	-2.5%
70%	30%	21%	20%	5.0%

[1] Assume figure

[2] Typical overbooking policy reference from https://www.stayntouch.com/blog/overbooking-your-hotel/

Model Accuracy	Average cancellation rate (%of Total booking)[1]	Expected overbooking policy	Typical overbooking policy[2]	%Improvement
75%	30%	23%	20%	12.5%
80%	30%	24%	20%	20.0%
85%	30%	26%	20%	27.5%
90%	30%	27%	20%	35.0%
95%	30%	29%	20%	42.5%
100%	30%	30%	20%	50.0%

Table 1 : Model Evaluation Matrix

2. Analysis

2.1. Datasets and Inputs

This data set were extracted from hotels' Property Management System (PMS) SQL databases[3] which contains booking information for a city hotel located in the city of Lisbon and a resort hotel at the resort region of Algarve. Both hotels are located in Portugal, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, etc.[4]

2.2. Data Exploration

2.2.1. Data cleaning and imputing

The given dataset were properly cleaned. There were only minimal missing values from some features such as company (ID of the company/entity that made the booking. for NULL values, It can be interpret as hotel being booked personally. So in this case, I'll impute "NULL" with 0).

For others incomplete features, their missing value had been imputed with strategy as shown in Table 2

[3] Data from Kaggle : https://www.kaggle.com/jessemostipak/hotel-booking-demand , Originally from https://www.sciencedirect.com/science/article/pii/S2352340918315191

[4] All personally identifying information has been removed from the data

Features	No. of missing value	Imputing strategy
company	108103	Constant = 0
agent	15622	Constant = 0
country	464	Constant = 'UNK'[5]
children	4	Constant = 0

Table 2 : Imputing Strategy

2.3. Exploratory Data Analysis

2.3.1. Hotel guest profile

Top 5 visitors are from nearby European countries. With the largest portion from Portugal, following by UK, France, Spain and Germany[6] respectively.

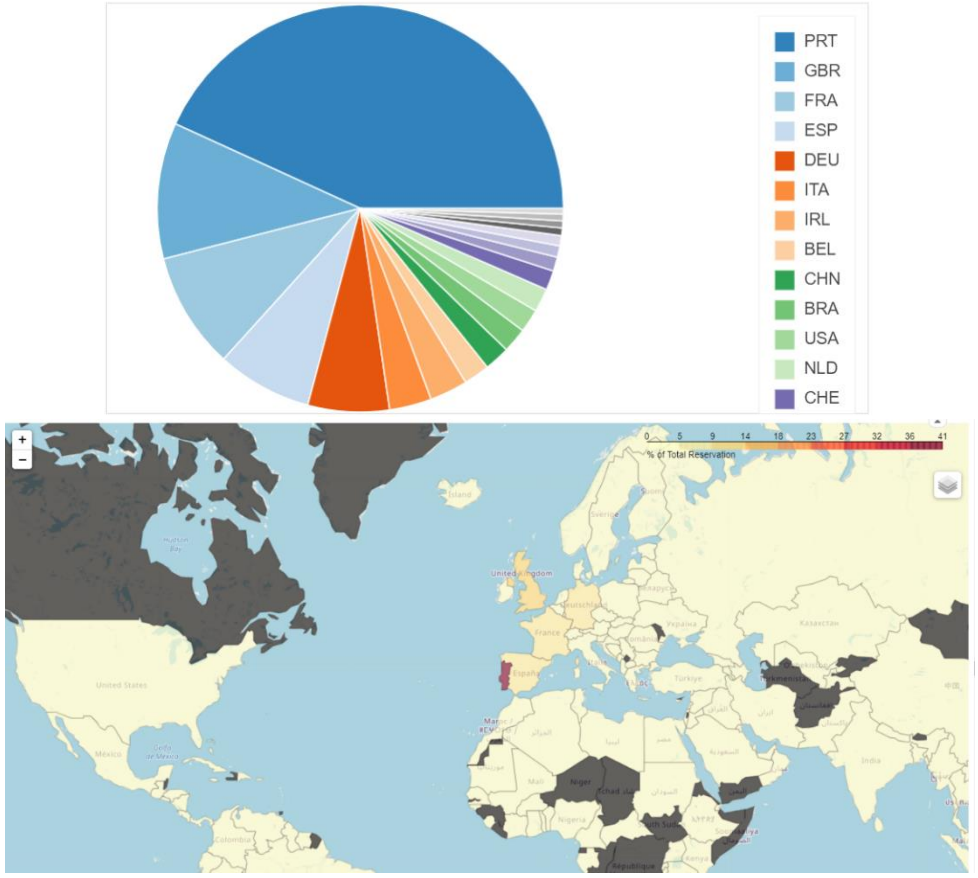
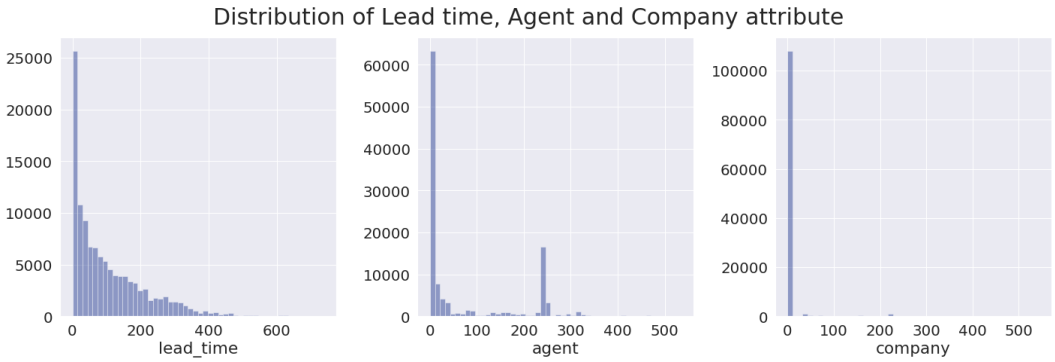
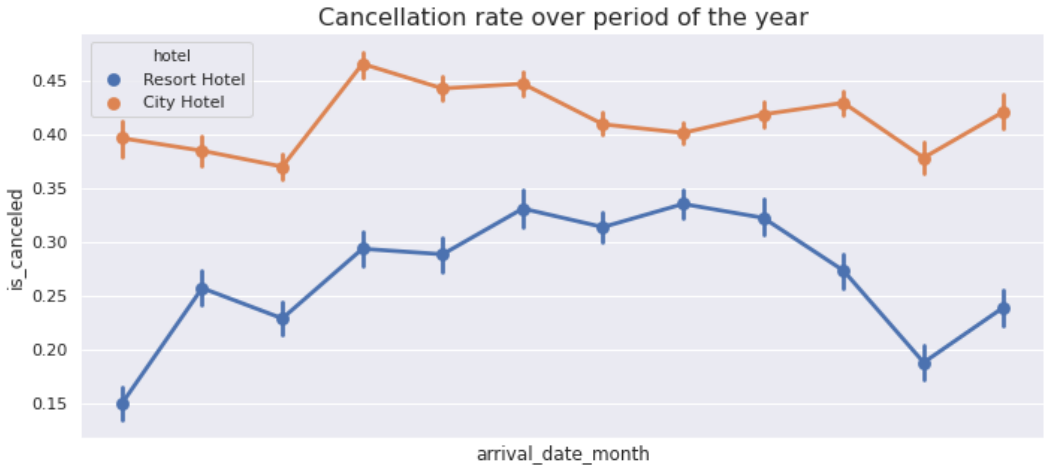
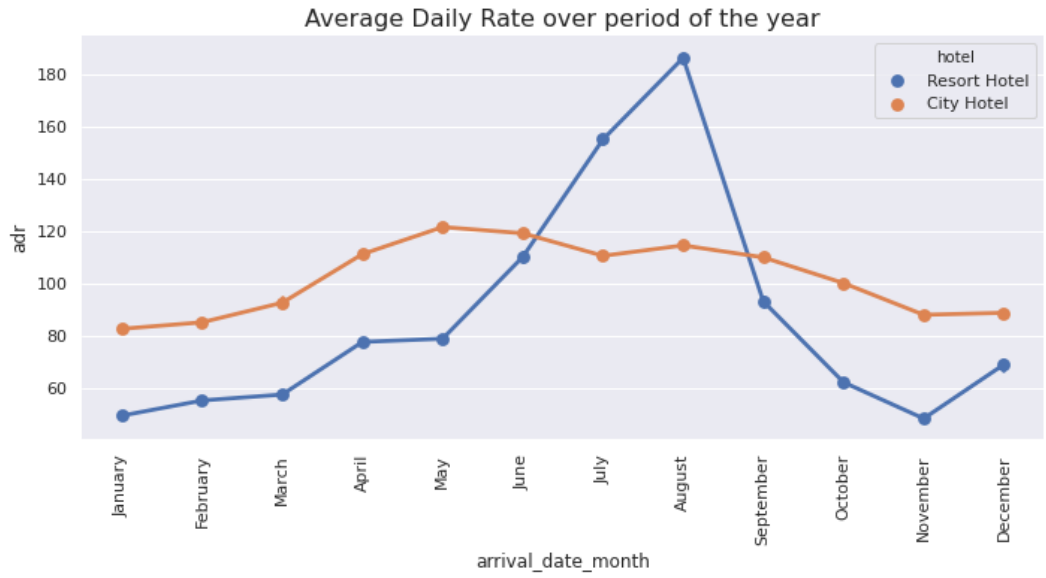


Figure 1 : Where our guest come from?

[5] UNK stand for Unknown
[6] Country of origin which are represented in the ISO 3155–3:2013 format







Observations :

- lead time (days) : Almost all reservations were made with lead times
- agent & company : Most of reservations were made personally and there are only major travel agencies which have significant amount compared to number of personally reserved

- During high season (June-August[7]) when average daily rate of resort hotel is increasing crazily. On the other hand, there are also increasing in accommodation cancellation rate which reach peak around 33%

2.3.2. Cancellation statistics

From top 20 most reservation countries, Portugal has highest rate of cancellation (nearly 60%). But we can not neglect them. Since they are largest customer segment and number of their confirmed booking alone still outnumbered other countries.

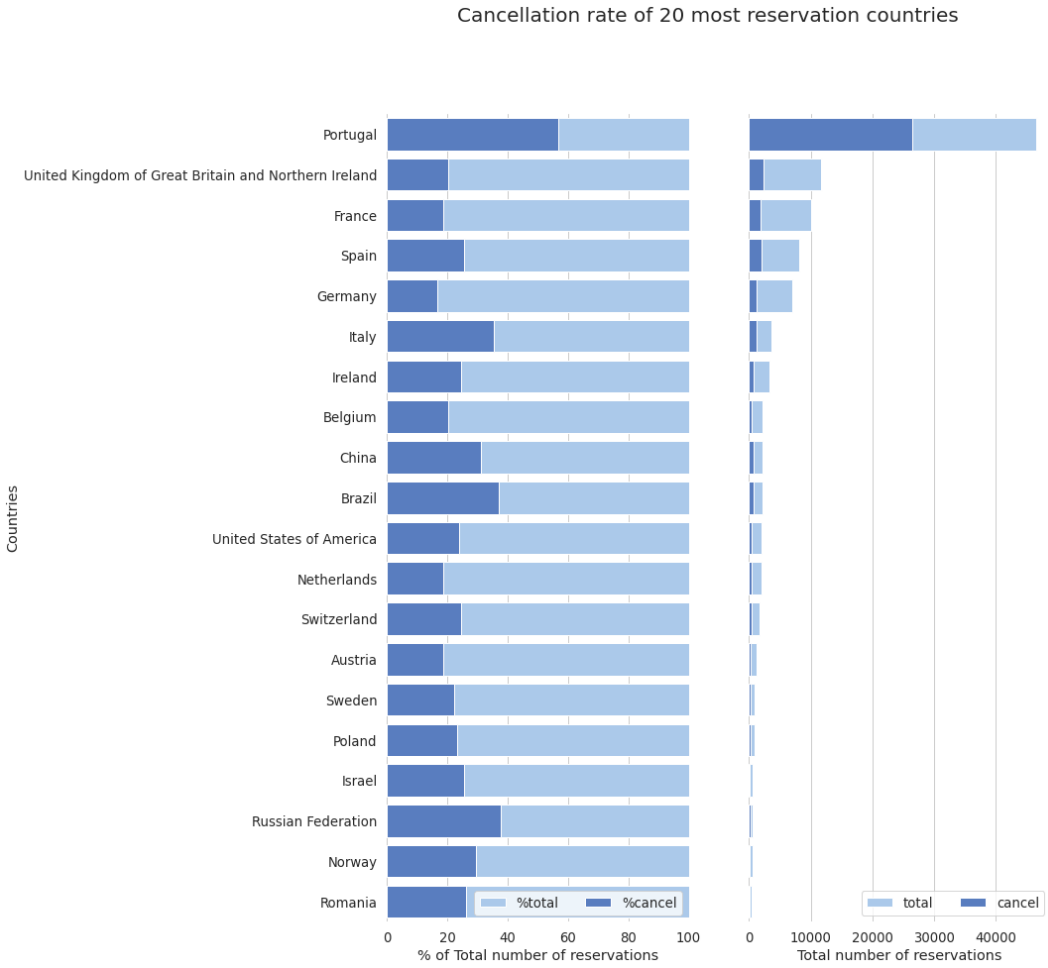


Figure 2 : Top 20 countries where our guests come from

To find out which feature has relation with cancellation rate, I decide to take a quick glance at Pearson Correlation Matrix as shown below

[7] Reference from : When to visit Lisbon? The best time of year for a holiday to Lisbon and weather (https://lisbonlisboaportugal.com/lisbon-tour/lisbon-weather-when-to-go-visit.html)

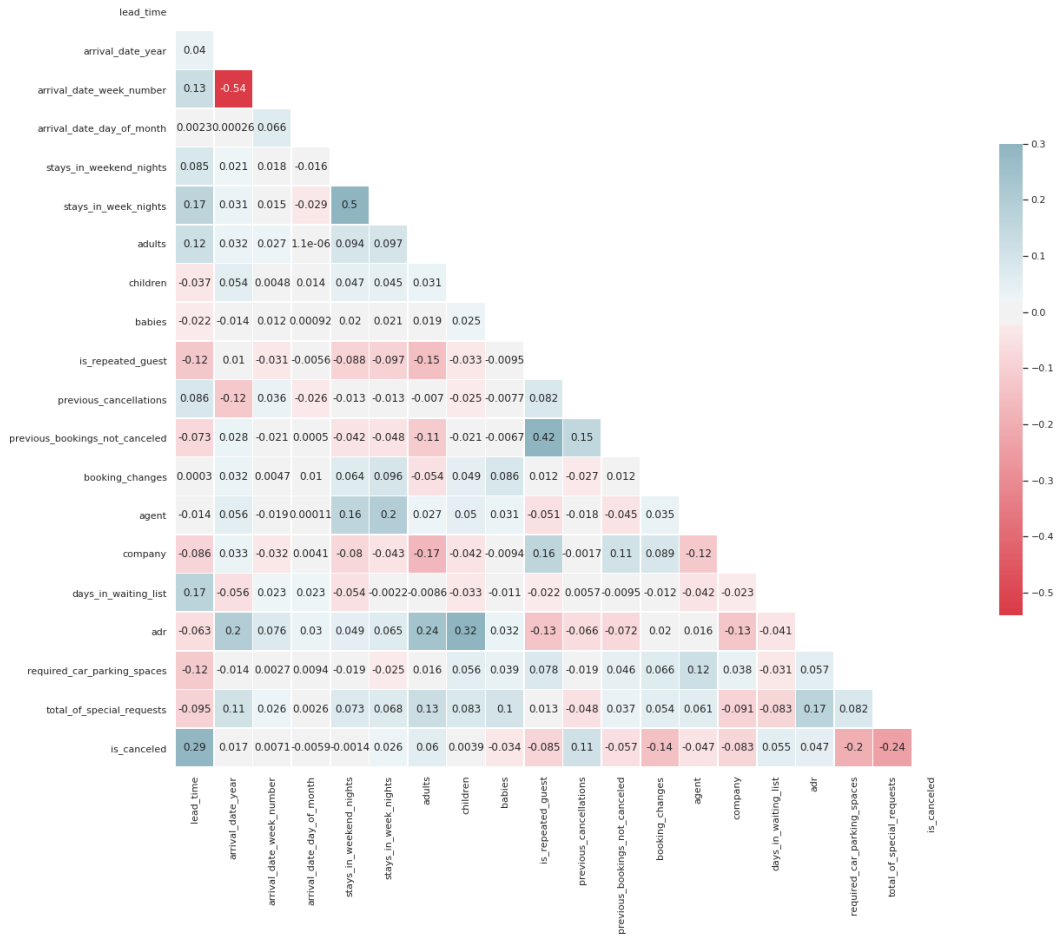


Figure 3 : Correlation matrix of numerical attributes

Finding from correlation matrix

Factors which may lead to cancellation

- **Lead time** : reservation with more lead time (booking in advance) is more likely to be cancelled

Factors which contribute for successful reservation

- **#of booking changes** : guests with booking change has tendency to visit the hotel (at the first time, guests may not available so they make booking change to the date they're available)
- **parking space requirement** : guests who required parking spaces has lower possibility to cancel the reservation
- **total number of special requests** : guests with special requests is likely to have strong intention to visit

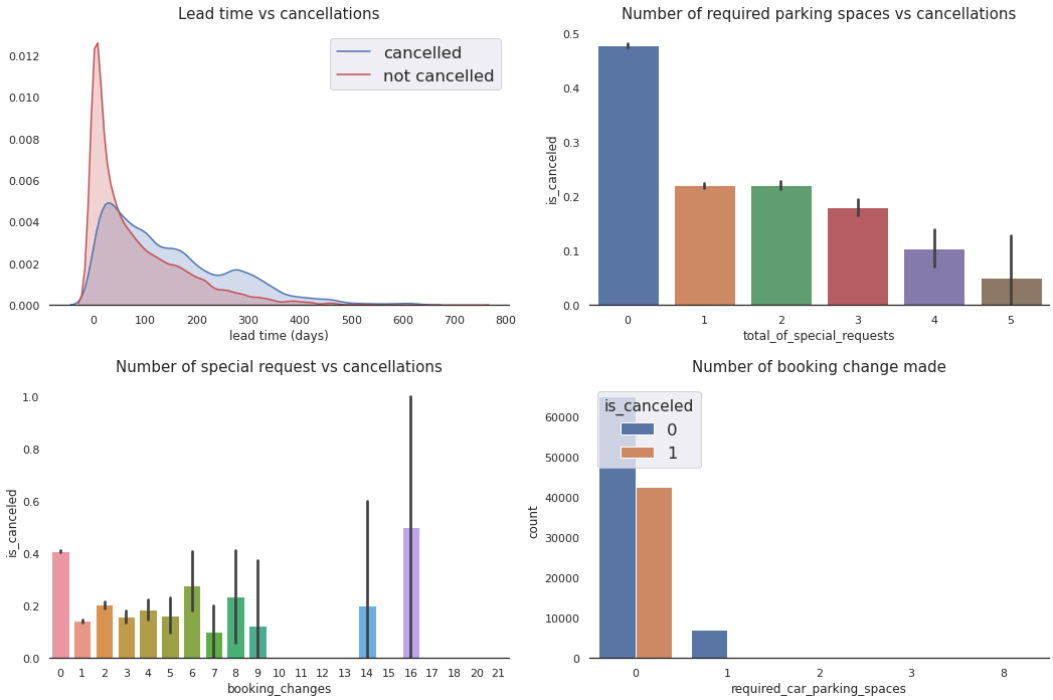


Figure 4 : Factors related to cancellation rate

To support mentioned observation, I've decided to plot above figures. Let's look what it can implies,

- Guest who not canceled the reservation often have short lead time.
- Guest with high amount of special request seem to have interest in hotel and have less cancellations than guest without special request.
- Guest with some booking change (1-5) has lower possibility of cancellations than one who doesn't make any change.
- Guest who demands single parking space never cancel!

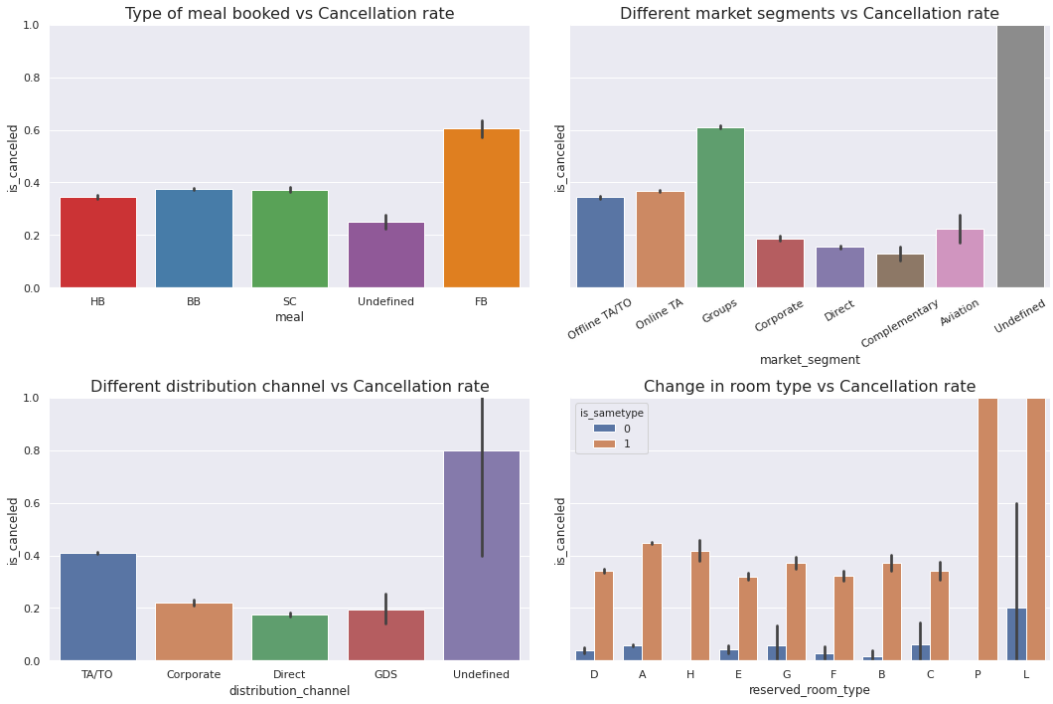


Figure 5 : Factors related to cancellation rate 2

Observations :

- Booking with FB – Full board (breakfast, lunch and dinner) have highest rate of cancellation
- Exclude undefined out of market segment, Groups segment have highest rate of cancellation than other segment. Reservation which was given as a complementary has lowest cancellation rate
- It's very surprising that reservation with unexpected change in room type having distinctively lower rate of cancellation than reservation with unchanged in room type! However, this information may not be very useful when we're trying to predict whether reservations will be canceled or not. Since it's considered as information leakage[8]

[8] "if any other feature whose value would not actually be available in practice at the time you'd want to use the model to make a prediction, is a feature that can introduce leakage to your model" : Data Skeptic

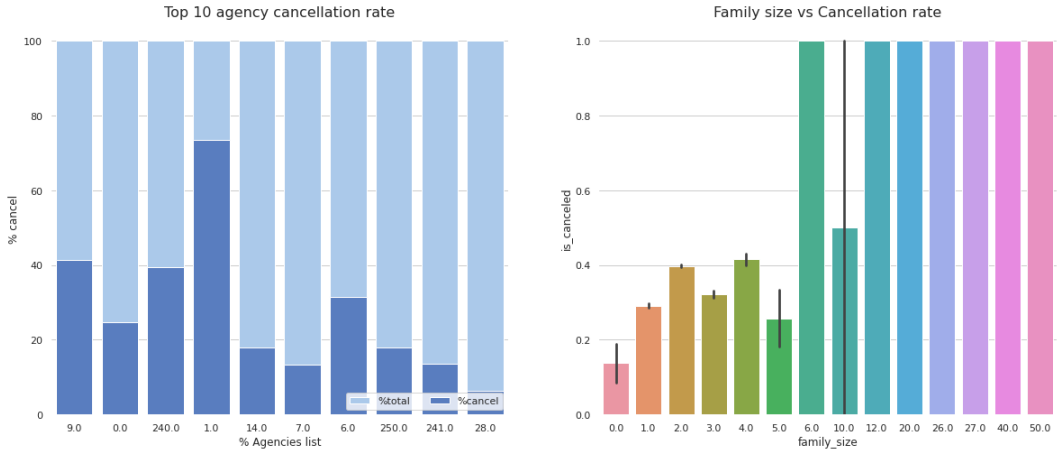


Figure 6 : Agencies and family size on cancellation rate

Observations :

- Reservation made by Agency ID=1.0 have higher possibility of cancellations than other agencies
- Almost all group visitors canceled their reservations

3. Methodology

3.1. Data Preprocessing and Feature Engineering

We begin with importing dataset with pandas.read_csv() method. After that I decided to set aside the test set for final model evaluation only. This is simply done by sklearn.preprocessing.train_test_split method. Then the training dataset was hand-clean before and along with exploratory data analysis.

During analysis, some interesting features such as family_size, stay_duration and is_sametype was discovered and can be described as

- family_size = sum of #adults + #children + #babies
- stay_duration = #stays_in_weekend_nights + #stays_in_week_nights
- is_sametype : is 1 if assigned_room_type = reserved_room_type, otherwise 0

3.2. Put preprocessing steps into pipeline

In order to make data preprocessing steps reproducible, pipeline was utilized to combine each preprocessing step into single and easily callable pipeline object.

However, our features consist of different data type which need different preprocessing steps. So, I decided to split data into 4 groups for different handling technique.

Note that some feature haven't been selected into model because issue of data leakage (such as is_sametype, days_in_waiting_list,) since when the model deploy in the future we might not have these information in hands when we want to predict cancellation.

Categorical	Country	Numerical	Combine
'arrival_date_month'	'country'	['arrival_date_day_of_month'	['stays_in_weekend_nights'
'meal'		'previous_cancellations'	'stays_in_week_nights'
'market_segment',		'previous_bookings_not_canceled'	'adults'
'distribution_channel'		'booking_changes'	'children'
'reserved_room_type'		'required_car_parking_spaces'	'babies'
'deposit_type'		'total_of_special_requests'	
'agent'		'lead_time'	
'company'		'adr'	
'customer_type'			
'is_repeated_guest'			

Table 3 : Different Features Categories

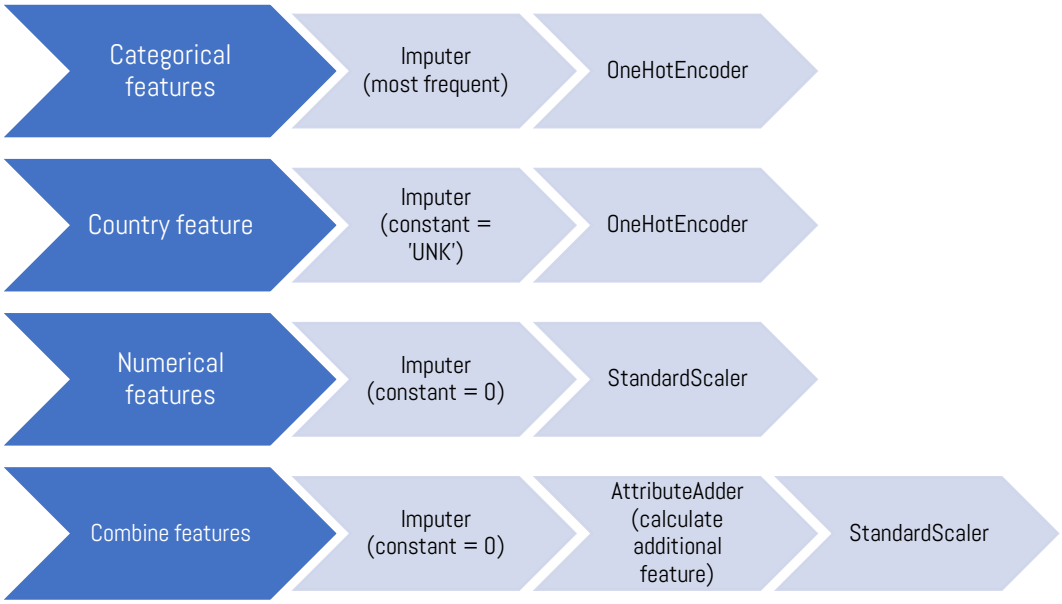


Figure 7 : 4 Pipeline constructed for handling different data type

3.3. Modeling

After we got our data preprocessed, we're now ready to feed it into our model. Since our task is to predict whether given reservation will be canceled or not? (This is a supervised classification problem). Therefore, I decide to compare following algorithms

- Logistic Regression
- Support Vector Machine with Polynomial Kernel
- Support Vector Machine with RBF kernel
- Random Forest Classifier
- Extreme Gradient Boosting (XGBoost) Classifier
- Adaptive Gradient Boosting (AdaBoost) Classifier

All algorithm will be compared based on their default configuration (untuned) by scikit-learn. Comparison will be conducted with 4-fold cross validation method on the training set and will be compared by mean accuracy. The accuracy results as shown below

Model	Mean accuracy (%)	STD (%)
Logistic Regression	82.6	0.2
SVM (Polynomial kernel)	85.8	0.2
SVM (RBF kernel)	85.9	0.2
Random Forest	88.5	0.2
XGBoost	87.2	0.3
AdaBoost	82.8	0.2

Table 4 : Cross validation results of different model

Evaluation result show that RandomForestClassifer are the best among others. However it also have extremely long training time compare to XGBoost and AdaBoost model. Therefore, we'll try to optimize parameter for RandomForest, XGBoost and AdaBoost model

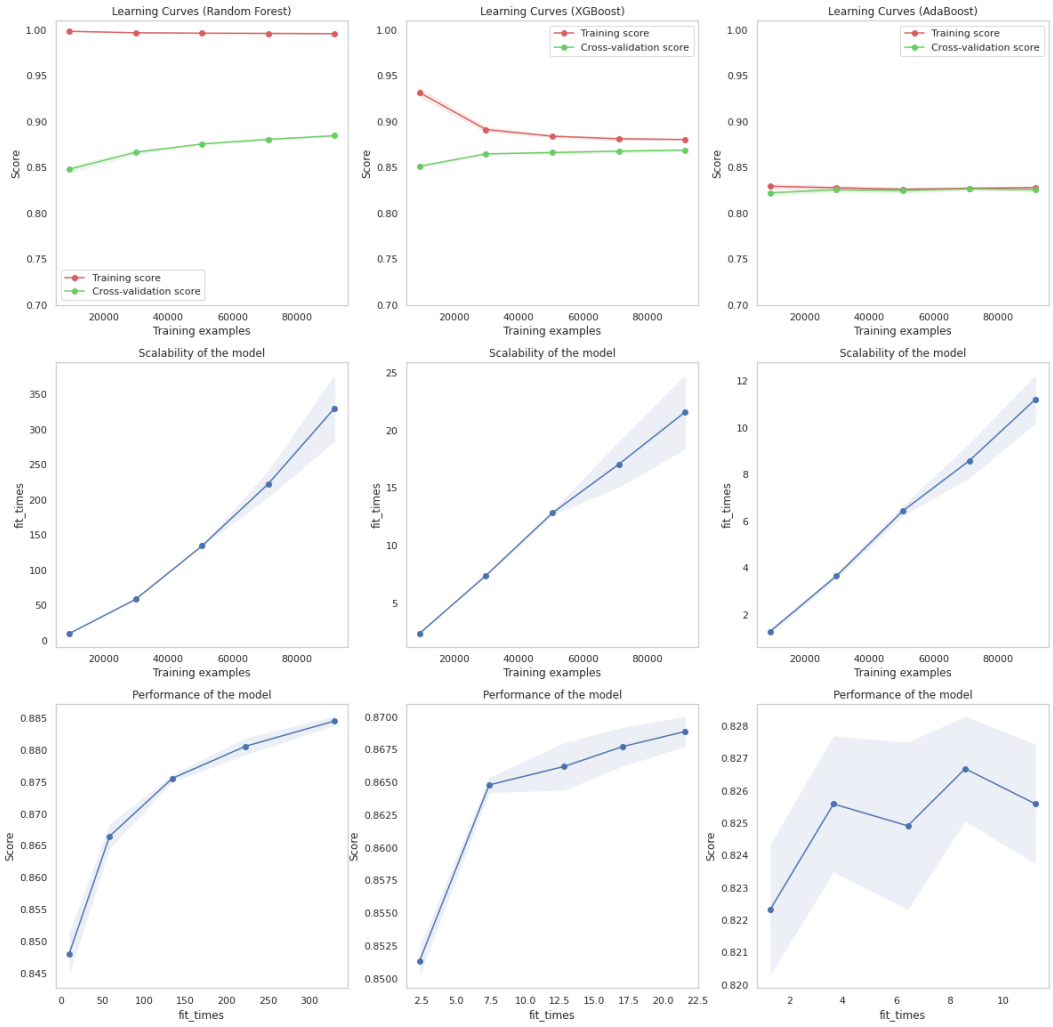


Figure 8 : Learning curve of Random forest , XGBoost and AdaBoost model

3.4. Hyperparameters tuning

We use RandomizedSearchCV to fine tune our Random Forest[9] , XGBoost[10] and AdaBoost[11] and after fine tuning and trade off between training time and accuracy, XGBoost seem to be the most suitable model with set of following hyperparameters

- Learning rate = 0.1
- Max_depth = 15
- n_estimators = 270

[9] See more about RandomForestClassifier hyperparameters at : https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[10] See more about XGBoost model at https://xgboost.readthedocs.io/en/latest/python/python_api.html#module-xgboost.sklearn

[11] See more about AdaBoost model at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

- reg_lambda = 0.001

With combination of optimum hyperparameters, we were able to marginally improve XGBoost model accuracy from 87.2% to 88.8%. We could try to fine tune with more hyperparameters but it's very time consuming and the results might not be promising.

4. Results

4.1. Model performance

Our candidate model (XGBoost model with fine-tuned parameters) has accuracy of 87.2% on cross-validation set and 89.2% on test set which could be considered as much more improvement from breakeven of benchmark accuracy (about 70%)

Take scenarios of high season at resort hotel as an example, with average cancellation rate of 30%. Normally hotel operator would set overbooking policy with typical margin of 20%. However, our model able to detect possibility of booking cancellation with accuracy of 89.2% which mean that 89.2% of 30% cancellation rate can be detected. Therefore, hotel operator may consider to adjust overbooking policy to have larger overbooked ratio up to 26.8% (Of course some safety margin should be applied).

4.2. Feature importance

The feature importance of multiple tree-based classifier are shown in figure below. It's not very surprise that lead time has most predictive power among others features. In addition, deposit requirement also have strong predictive power, since all non-refund type were almost canceled. Others importance features such as adr, countries, number of special request, agencies etc. has shown significant predictive power as we have discussed in detailed in EDA section. Surprisingly, 'arrival_date_day_of_month' come up in top rank with no clue during hand-on analysis. This could be improved if we could modify and create new features and eliminate features with low predictive power to further improve model performance.

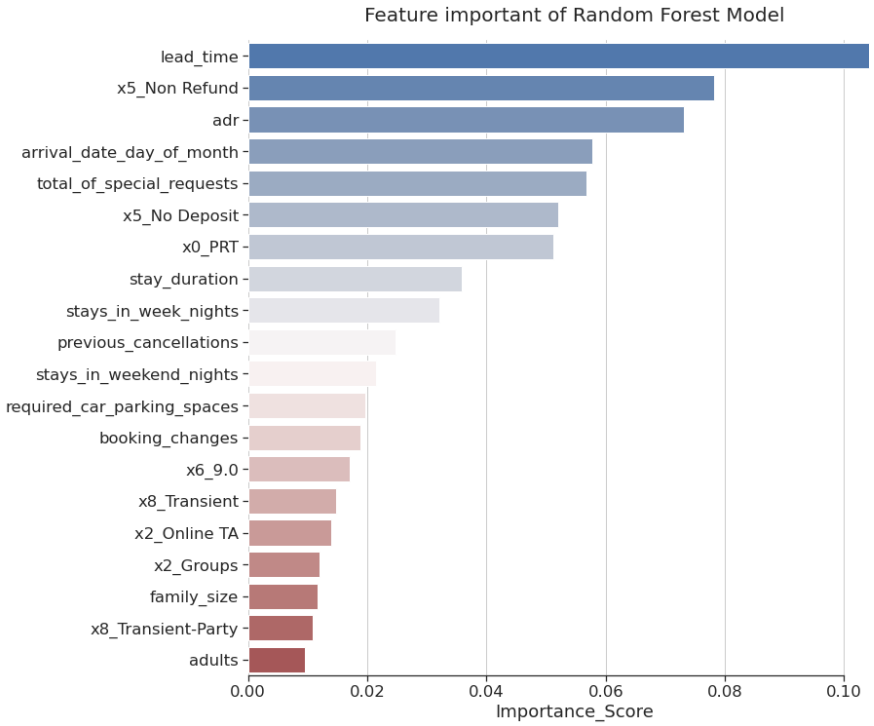


Figure 9 : Feature importance of random forest classifier

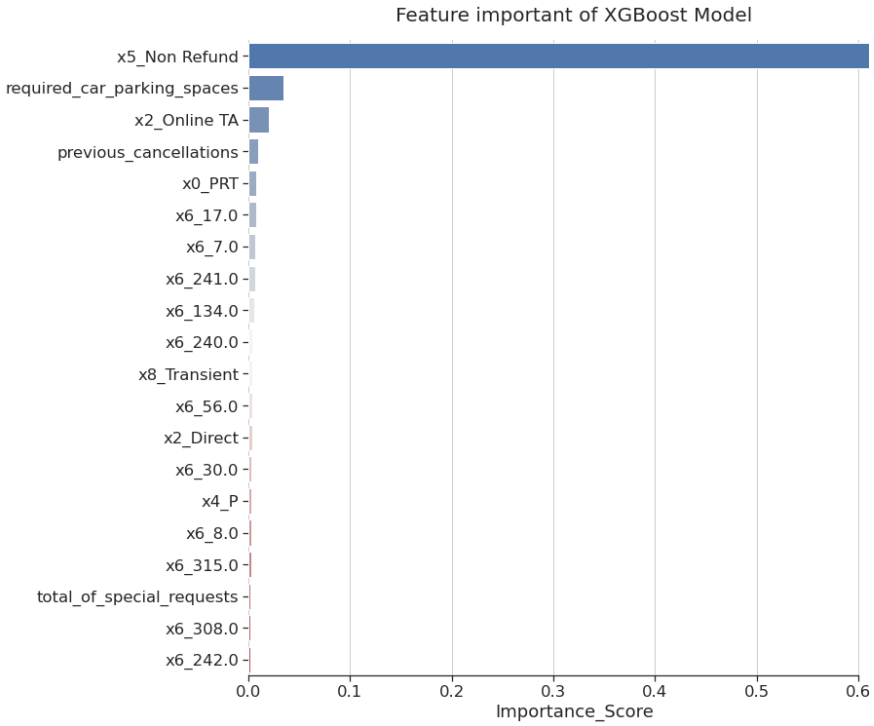


Figure 10 : Feature importance of XGBoost

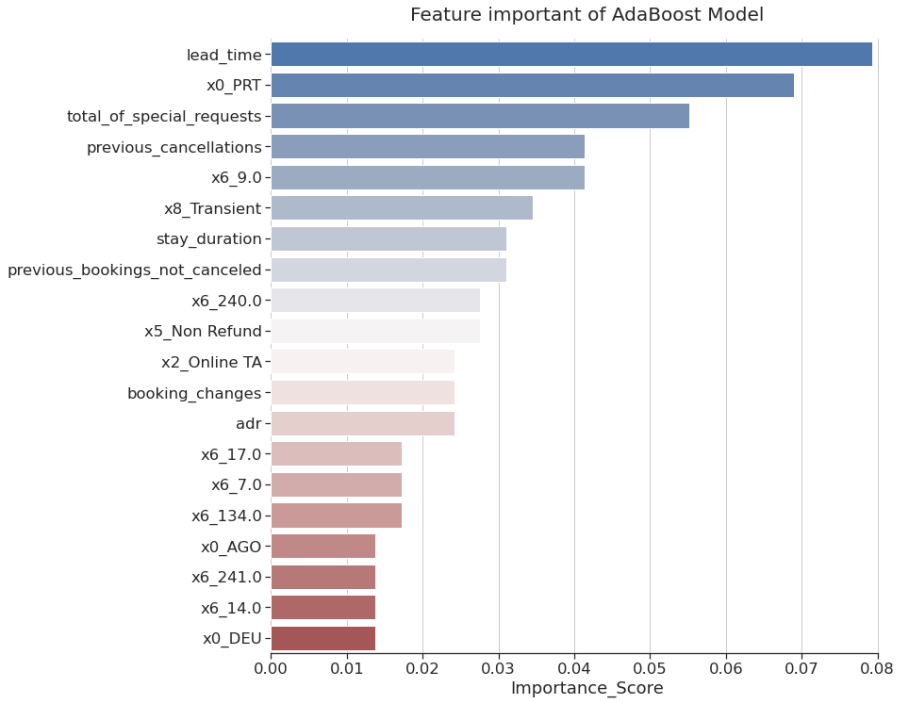


Figure 11 : Feature importance of AdaBoost

5. Conclusion

We have explored the data and able to extracted some useful insights since exploratory data analysis. We've found that the booking are likely to be canceled if they have incredibly long lead time, have little or no special requirements, booked with full broad meal ,etc. Some specific group of guests also have higher chance of cancellation than the others. For example, guests from Portugal themselves has higher rate of cancellation than some foreign travelers[12].

Moreover, we have trained classification models to predict whether a given reservations will be canceled or not. The outcome of the model quite satisfy (with an accuracy of 89.2%). Although it can be improved by training with more data and adjusting regularization parameters. But with current performance, our model has shown significant improvement from benchmark which can be utilized to adjust overbooking policy to achieve higher occupancy rate. In addition, even low season period when booking amount doesn't fulfill hotel capacity. The model can be used to manage hotel workforce according to fluctuation in demand in each period of years.

[12] detailed finding can be found in exploratory data analysis section (section 2.3)

6. Appendix

6.1. Data Dictionary

variable	class	description
hotel	character	Hotel (H1 = Resort Hotel or H2 = City Hotel)
is_canceled	double	Value indicating if the booking was canceled (1) or not (0)
lead_time	double	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
arrival_date_year	double	Year of arrival date
arrival_date_month	character	Month of arrival date
arrival_date_week_number	double	Week number of year for arrival date
arrival_date_day_of_month	double	Day of arrival date
stays_in_weekend_nights	double	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	double	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	double	Number of adults
children	double	Number of children
babies	double	Number of babies

variable	class	description
meal	character	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
country	character	Country of origin. Categories are represented in the ISO 3155–3:2013 format
market_segment	character	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
distribution_channel	character	Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
is_repeated_guest	double	Value indicating if the booking name was from a repeated guest (1) or not (0)
previous_cancellations	double	Number of previous bookings that were cancelled by the customer prior to the current booking
previous_bookings_not_canceled	double	Number of previous bookings not cancelled by the customer prior to the current booking
reserved_room_type	character	Code of room type reserved. Code is presented instead of designation for anonymity reasons
assigned_room_type	character	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons

variable	class	description
booking_changes	double	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
deposit_type	character	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
agent	character	ID of the travel agency that made the booking
company	character	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
days_in_waiting_list	double	Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type	character	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking
adr	double	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
required_car_parking_spaces	double	Number of car parking spaces required by the customer
total_of_special_requests	double	Number of special requests made by the customer (e.g. twin bed or high floor)

variable	class	description
reservation_status	character	Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why
reservation_status_date	double	Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel