Springboard Data Science Capstone Project 2

Hotels Booking

Cancellation Prediction

Yang Fei
Mentor: Kenneth Gil-Pasquel
Data Science Capstone Project 2, July 2020

1. Introduction

In the tourism and travel industry, there are frequent changes in plans and cancellations. Such incidents usually bring some trouble to both tourists and hotel related workers.

Along with it, hotels accumulated a magnanimity of datasets which covered numerous booking information, including the booking time, tourist ages, parking demands ect. These datasets could be quite helpful for researchers in revenue management, machine learning, or data mining, as well as in other fields.

These datasets could be an quite beneficial part for data scientists' development of prediction models to classify a hotel booking's likelihood to be canceled. By combining with the data analysis, people may find out more potential and interesting trends and relationships in this industry, which could go further than the cancellation prediction problem itself.

WHO MIGHT CARE:

Hotel Management

It is well known that booking cancelation usually results in unavoidable financial loss to hotel management. With the help of cancelation prediction, hotel management could propose some overall arrangements and backup measures in the booking system, which could reduce the influence caused by the financial loss.

Tourists

In the peak seasons, tourists often encounter booking troubles when they make travelling plans. In fact this demand could complement hotel's cancelation problems efficiently. By observing the trend of cancelation of a hotel, tourists could know the probability of booking a hotel on their desired dates.

Hotel Industry Researchers

Through the research of the data of booking information and cancellation prediction, researchers could summarize the developing trends and regulations of this industry. Efficient and targeted strategies could be made to deal with the problems as well as to cater for the market requirement.

DATA:

The data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.

This data set contains booking information for a city hotel and a resort hotel, 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, among other things.

All personally identifying information has been removed from the data. From the article we could find further informations.https://www.sciencedirect.com/science/article/pii/S2352340918315191

APPROACH:

1. Clean data. Deal with the missing and unrealistic data points
2. EDA which could be focused on these factors:
    a. When is the peak season?
    b. What is the price trend in a year?
    c. Where are the visitors mainly from?
    d. Which kind of hotel is more popular?
    e. What kind of visitors usually make a booking?
    f. What is the lead time distribution?
    g. How long do the visitors usually stay?
    h. What is the proportion of stay week nights or weekend nights?

    The following factors are related to the target variable(cancellation) closely.

    i. When does cancellation usually happen in a year?
    j. How many bookings were canceled per year?
    k. Where are the visitors who are more likely to cancel the bookings from?
    l. What kind of hotels usually meet cancellation?
    m. What kind of visitors usually cancel the bookings?
    n. The relationship of cancellation rate and lead time
    o. The relationship of cancellation rate and planing stay time
    p. When are people more likely to cancel the booking? Weekends or week days?

3. Modeling

tools:Python's scikit learn

Target Variable: isCanceled (yes or no)

The Topic be considered as a binary classification problem of supervised machine learning.1 for cancelled and 0 for non-cancelled

DELIVERABLES:

Codes

Report on the capstone project

Presentation on the capstone project

All materials will be uploaded to my Github repository.

2. Data Cleaning and Wrangling

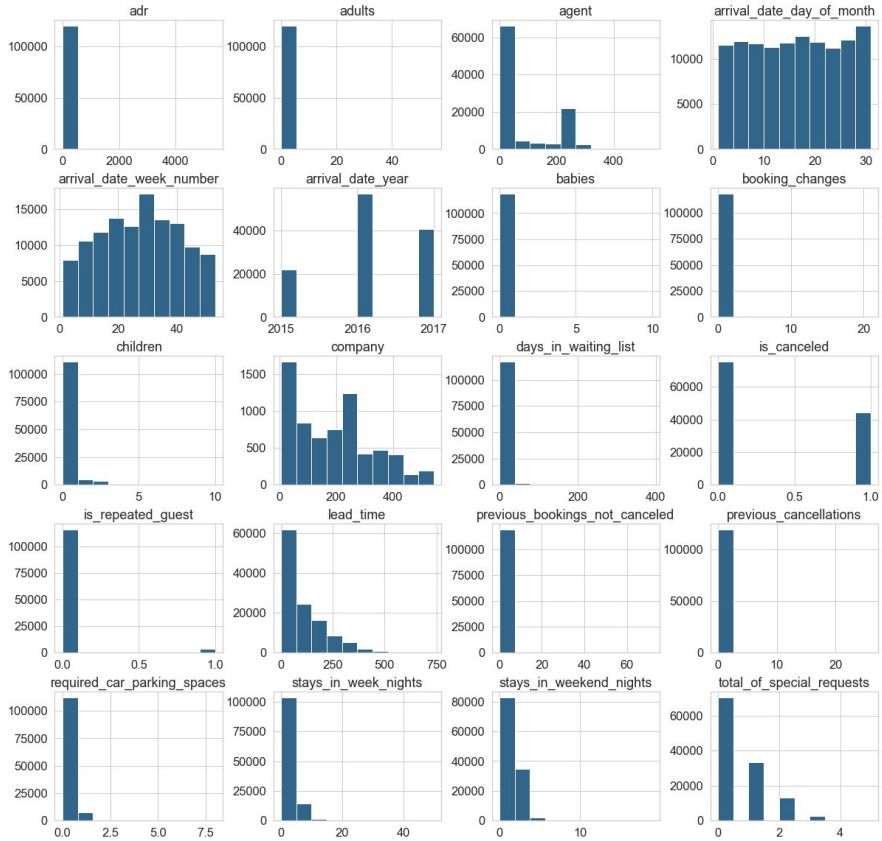
2.1 Import and check data

I got a general idea of the data to know the size of the data set. This data set contains 32 columns and 119390 rows; It is a mixed type data set. Some of the features are engineered from other variables from different database tables, which already are transferred into Dummy variables.

17 Numerical Variables: lead_time, arrival_date_year, arrival_date_week_number, arrival_date_day_of_month, stays_in_weekend_nights, stays_in_week_nights, adults, children, babies, previous_cancellation, previous_booking_number, days_in_waiting_list, adr, required_car_parking_spaces, total_of_special_requests, reservation_status, booking_changes

13 Categorical Variables: hotel, arrival_date_month, meal, country, market_segment, distribution_channel, deposit_type, customer_type, reservation_status, reserved_room_type, agent, company, assigned_room_type,

2 Dummy variables: is_concealed, is_repeated_guest,



2.2 Deal with missing data

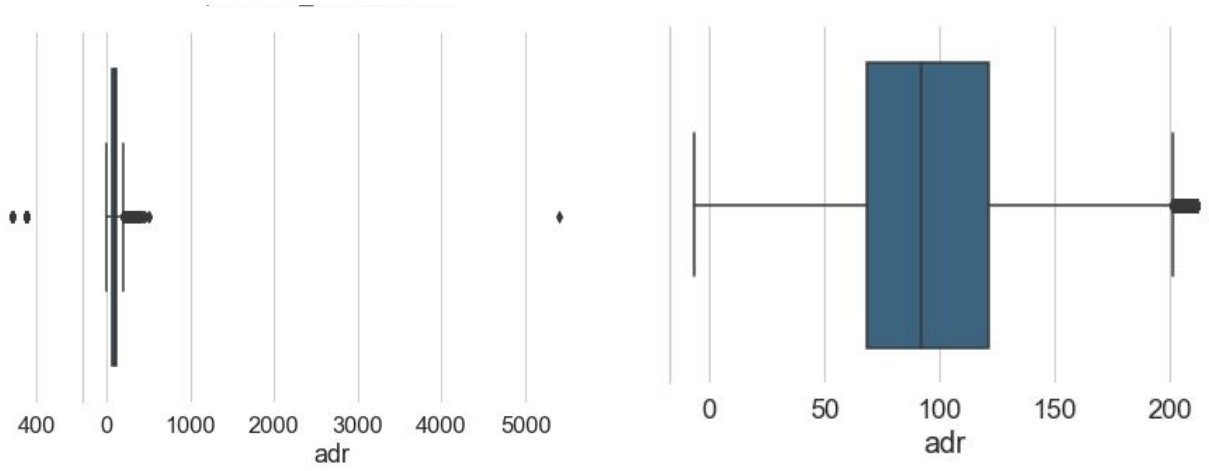
Some columns had a large proportion of missing data.
'Reservation_status','reservation_status_date' the two columns are the result of the target variable, and could not affect the target variable calculation. So I dropped them with the unnecessary columns together.

Other missing data were filled up by the mode of the features.

As the complementary information from the article, "meal" contains values "Undefined", which is equal to ‘SC’. So I replaced them by ‘SC’. "market_segment" contained 2 values "Undefined", which took little proportion of the dataset but may affect the following calculation, so dropped it.

2.3 Deal with outliers

‘Adr’ had an obvious not constant outlier which should be dropped.



3. Data Cleaning and Wrangling

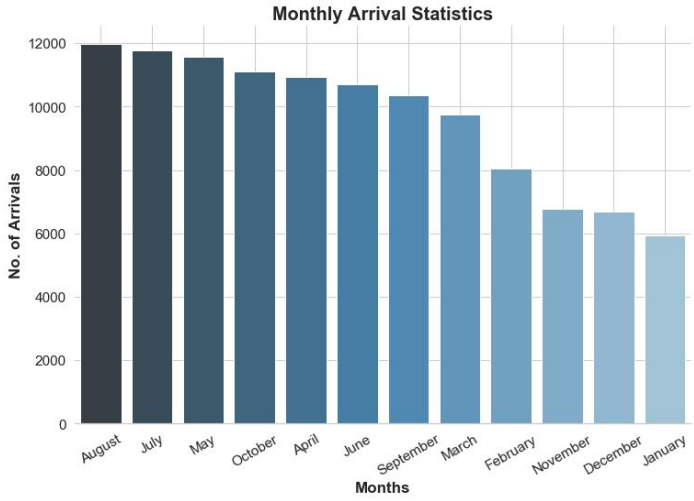
I divided the research content into two parts. Firstly some general analysis and wrangling should be taken for all the variables, aimed to find out the potential relationship between features. It also interpreted the questions most public were concerned about. Secondly the relationship between the target variable ‘is_cancelled’ and other features were to be emphasised on. A new feature ‘cancellation_rate’ was proposed to estimate the probability of booking cancellation.

After all that, the features which had close relationships should be paid more attention.

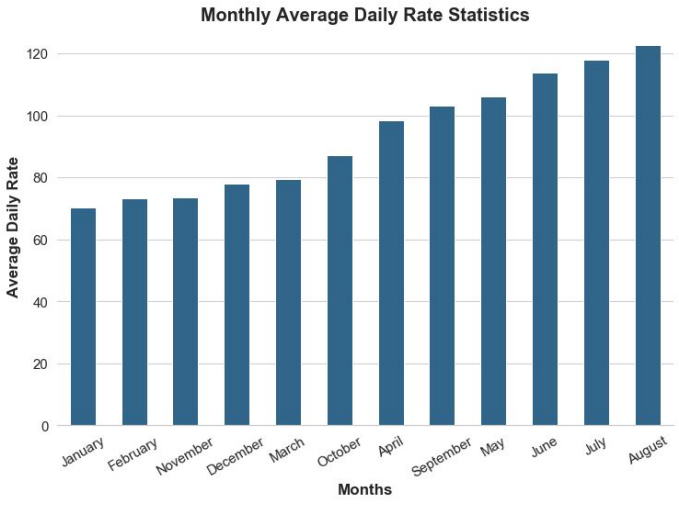
3.1 General Analysis

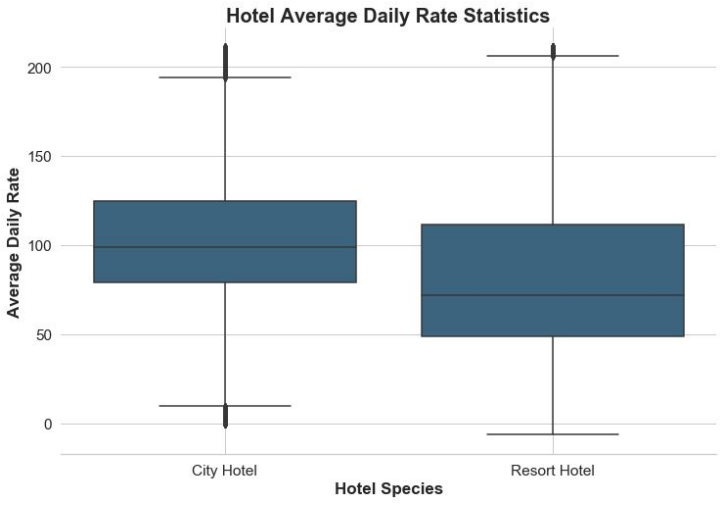
In this part I maily researched the relationship between features and the number of bookings which covered both cancelled and confirmed ones.

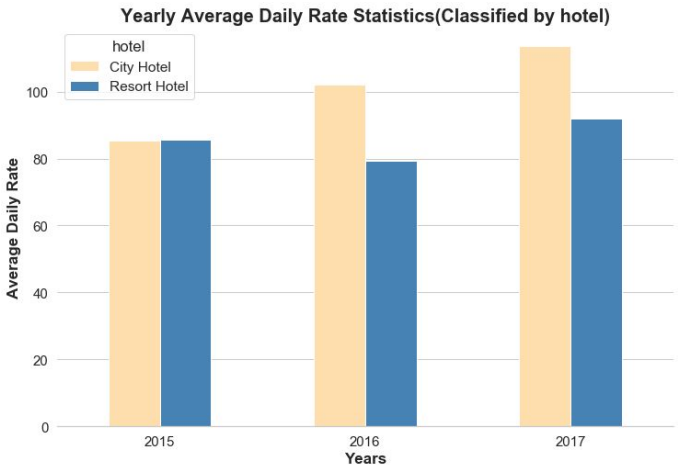
1. **Peak Season**
    - Summer (August and July) are the most popular seasons for visitors.Booking number in winter(Dec,Jan,Nov) is the lowest.
    - 2016 is the most popular year.



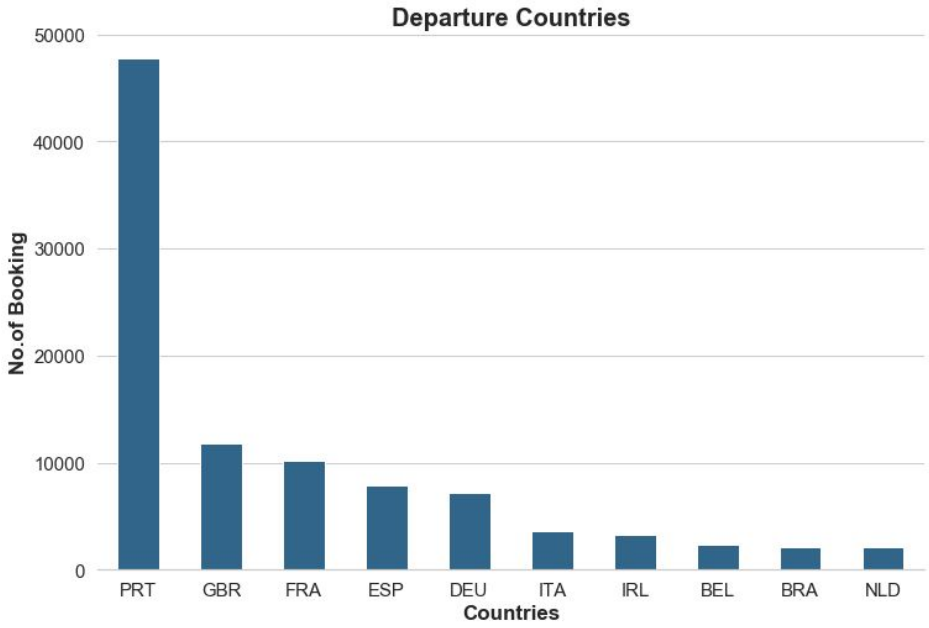
2. **Price Trend**
    - City hotels have a higher mean price than resort hotels.
    - August has the highest adr. Summer has higher adr than winter.
    - In summer resort hotels usually have a higher adr than city hotels. Resort hotels are greatly influenced by the seasons.
    - Adr(Average Daily Rate) is growing year by year.





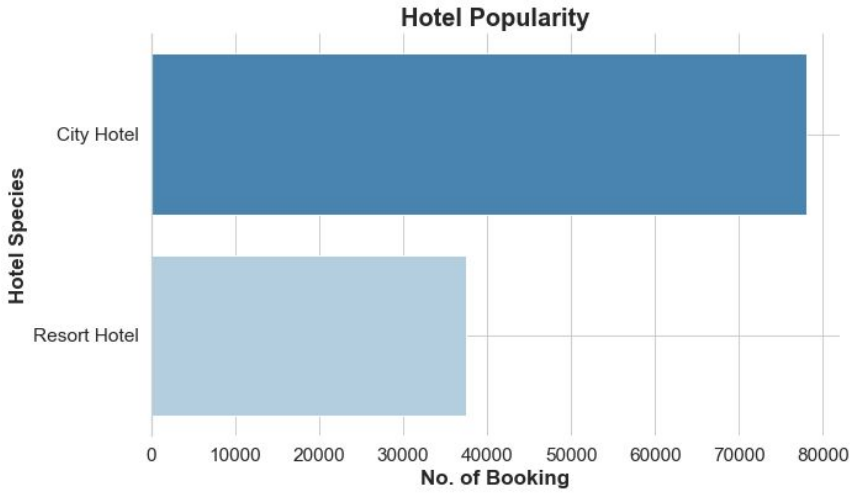


3. **Where are the visitors mainly from?**
   The dataset is created in Portugal. Except for Portugal, UK is the largest visitor's original country.

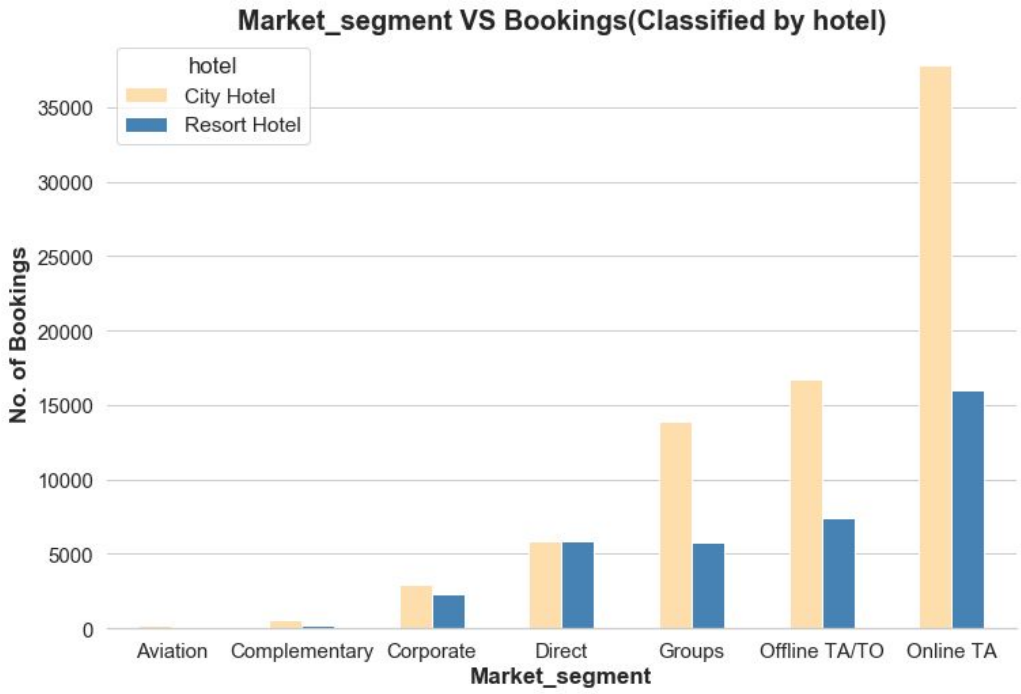


4. **Which kind of hotel is more popular?**

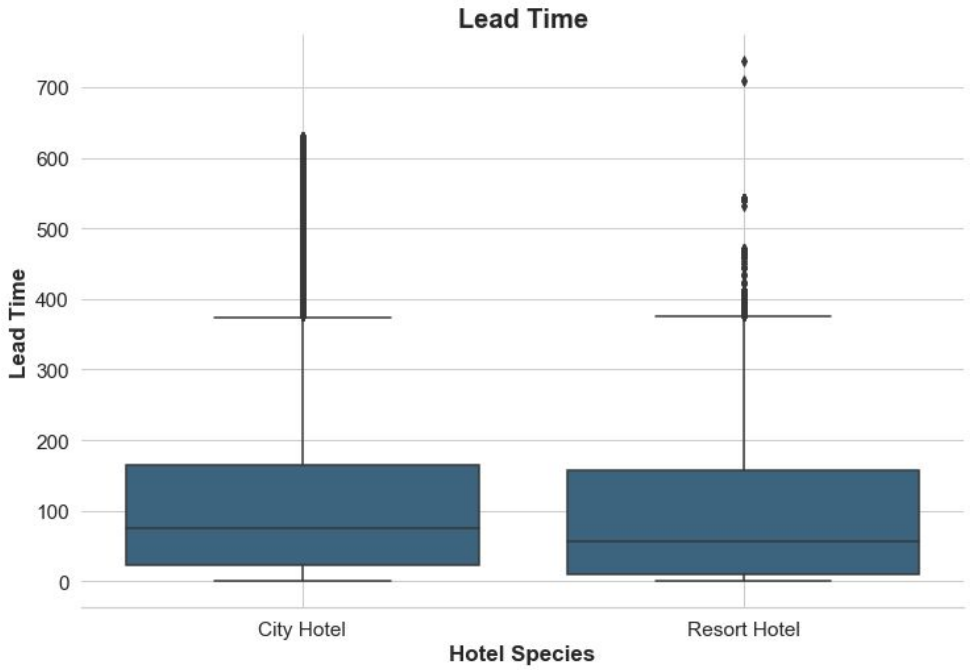
Booking number of city hotels is almost twice than resort hotels.



5. **What kind of methods by which visitors usually make a booking?**
   Most people prefer to choose the Online Travel Agents.

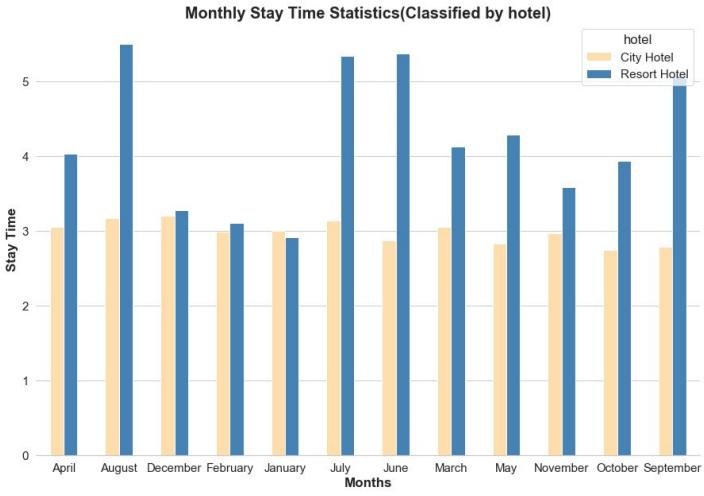


6. **What is the lead time distribution?**
   Most booking lead time concentratedly distributes below 100 days. Resort hotels lead time is a little less than city hotels



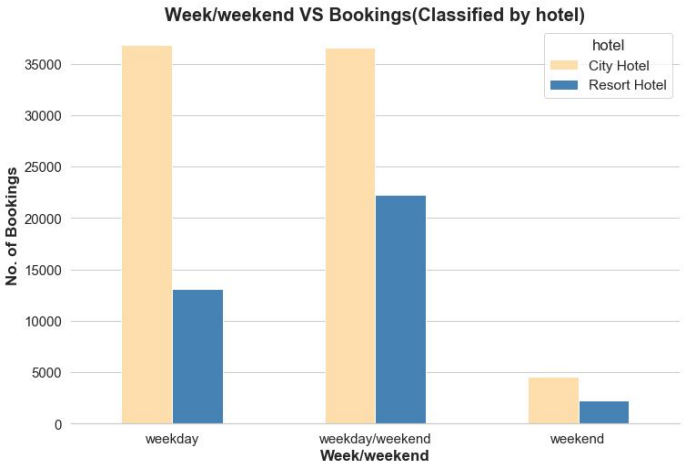
7. **How long do the visitors usually stay?**

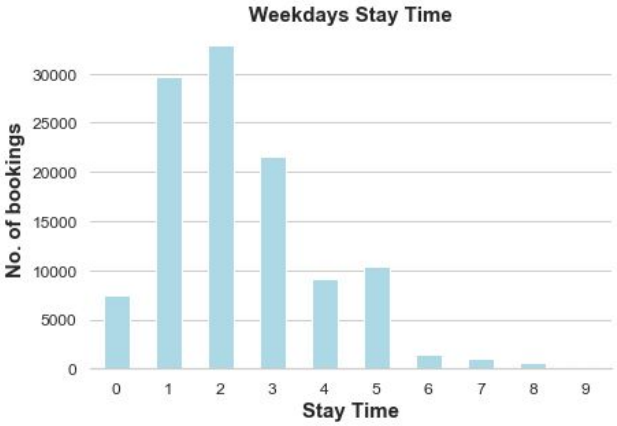
- Most bookings have a 1-3 days stay time.
- Resort hotels stay time is a little longer than city hotels.
- Resort hotels usually have a longer stay time in summer.

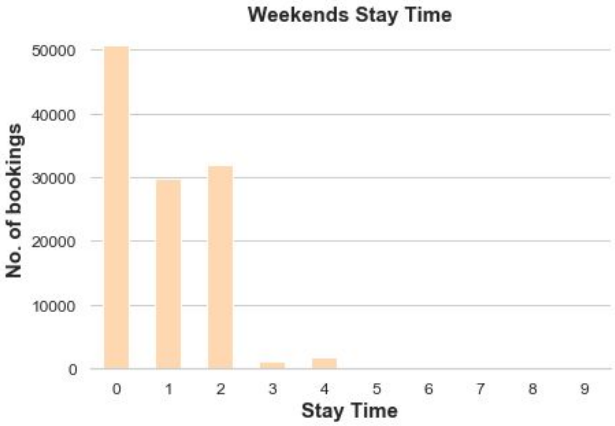


8. **What is the proportion of week nights or weekend nights stay?**

- Booking on weekdays is far more than weekends days.
- Stay time on weekdays is longer than weekends.



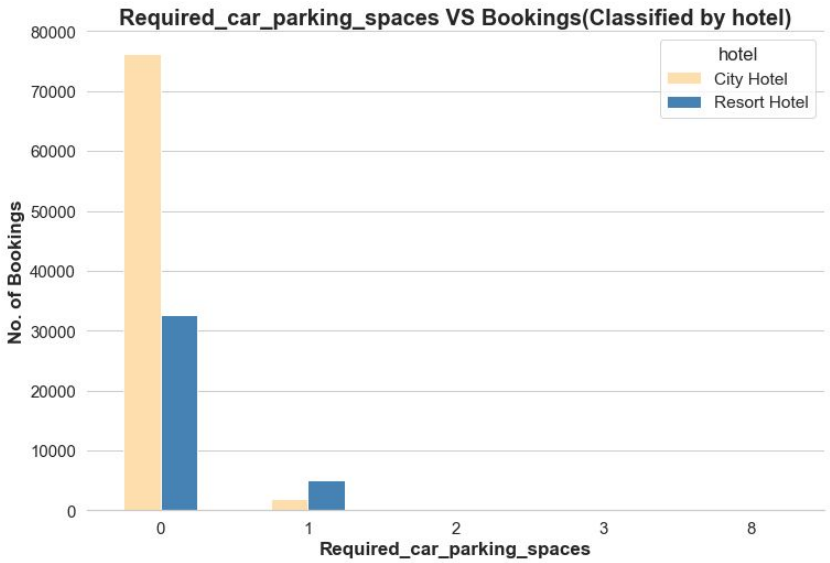




9. **What is the relationship of bookings and total number of special requests?**
   Most bookings had no special request.

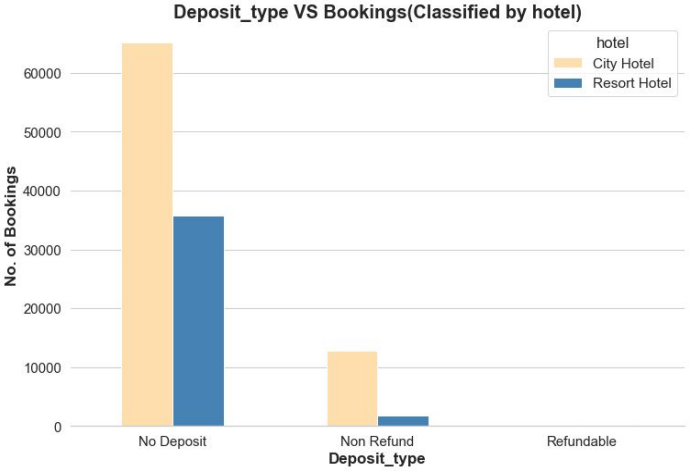
10. Will the car parking requirement affect bookings?

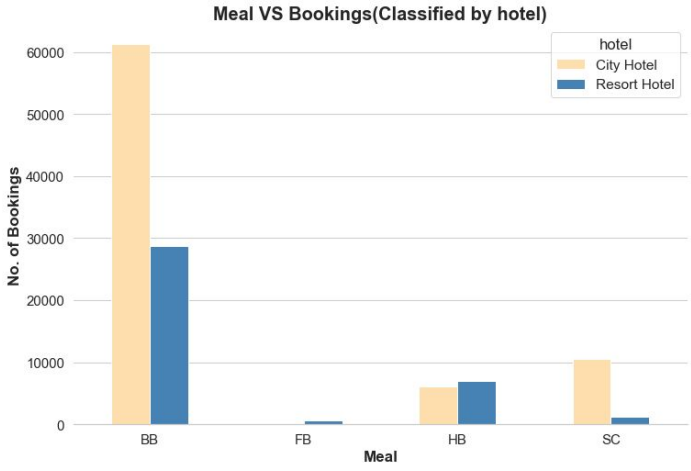
Most bookings had no request for bookings.



11. some additional questions

- Most bookings had no deposit.
- People prefered to order breakfast with bookings.





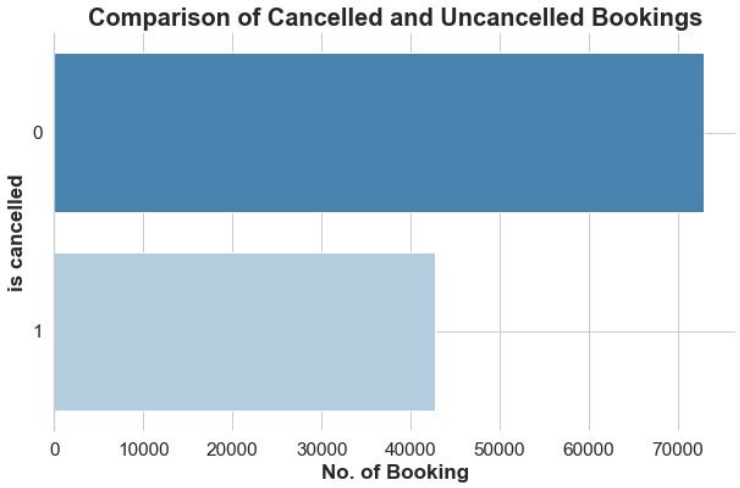
3.2 Target Related Analysis

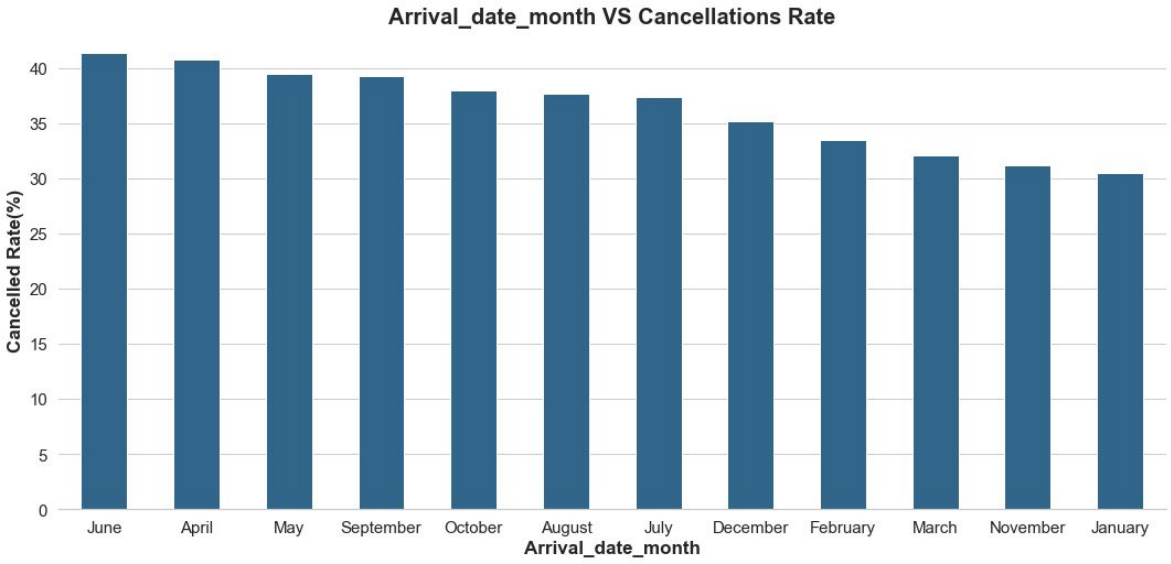
In this part I would take two variables, ‘number of cancellation’ and ‘cancellation rate’ as measurements to compare the relationship with other variables.

12. When is the peak season for cancellation?

- Booking cancellation number is far less than uncancelled ones.
- The cancellation number is higher in April,May, June and July, which could be considered as the time before holiday.

- Bookings in April, May, June are more likely to be canceled. In June the cancellation rate is more than 40%.



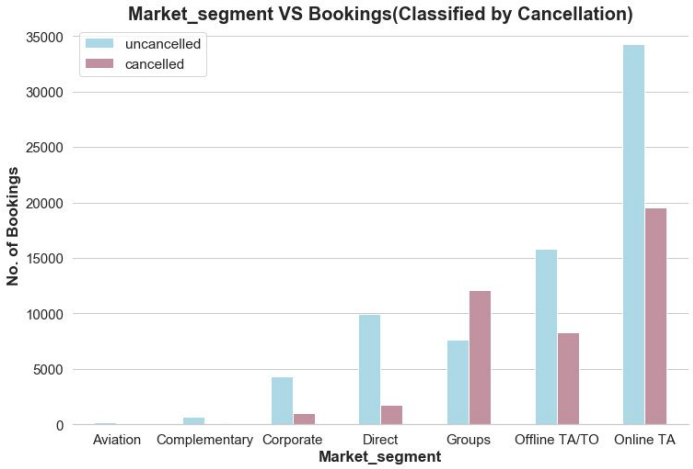


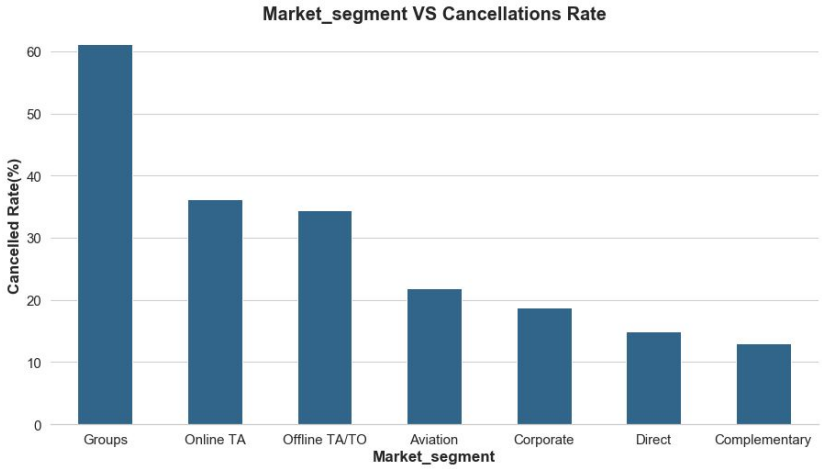
13. Which kind of hotel is more likely to be canceled?

Combined with no canceled booking number, city hotels bookings were more likely to be canceled, due to the great proportion of business trips.

14. What kind of booking methods are more likely to be cancelled?

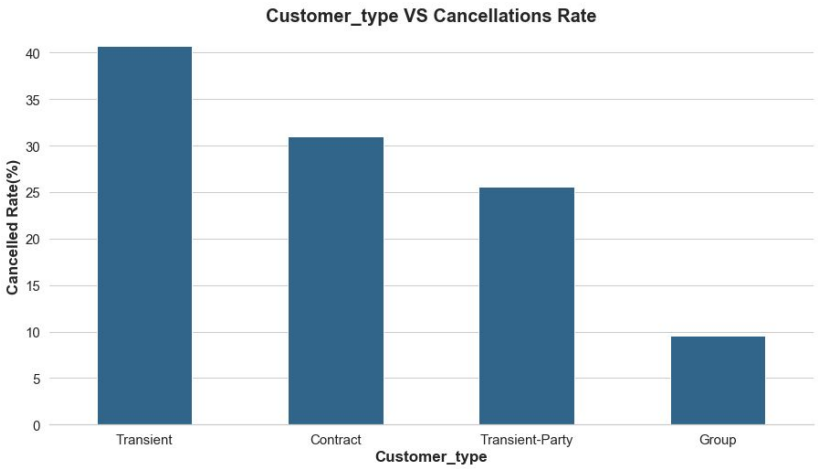
- Group bookings are more likely to suffer cancellation.
- But Online TA has the largest cancellation number.

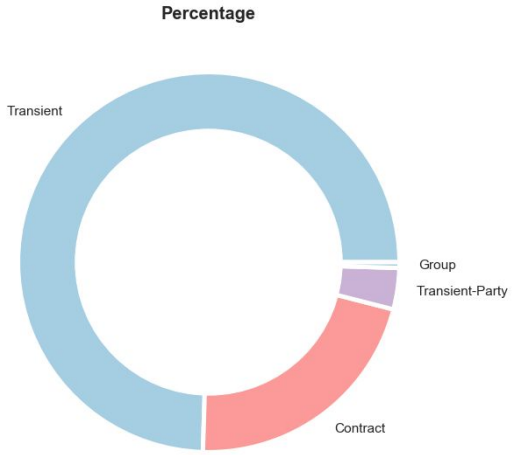




15.What type of customers are more likely to cancel the bookings?

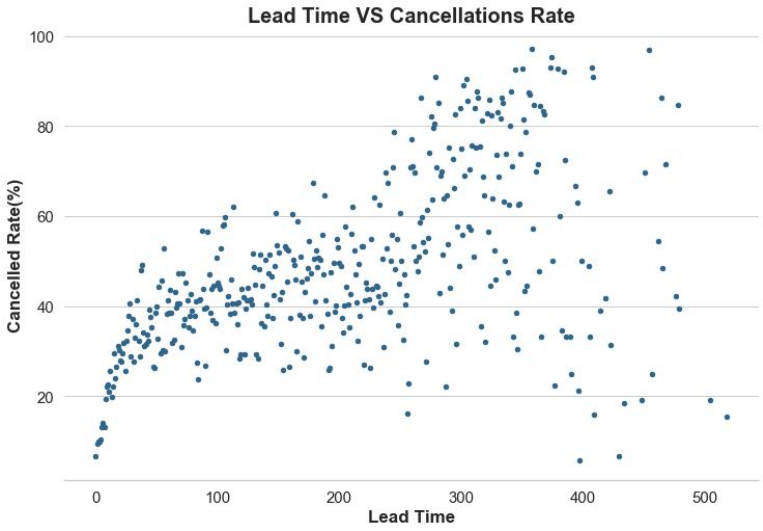
- Transient customers who have the largest cancellation number are more likely to cancel the booking.

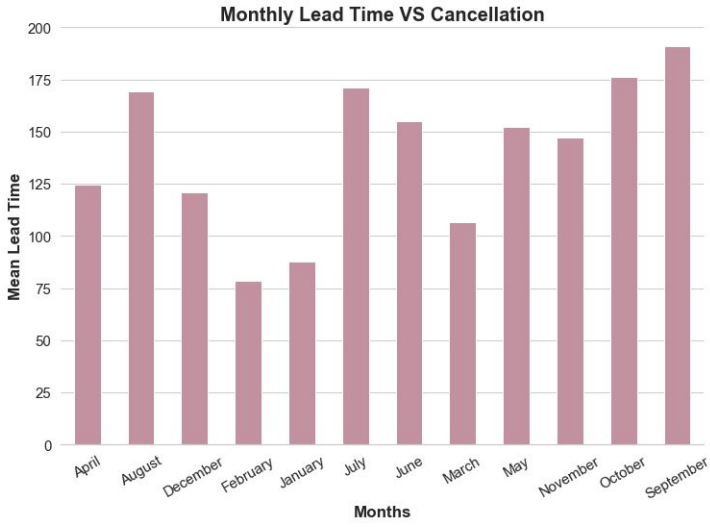




16.The relationship of cancellation rate and lead time

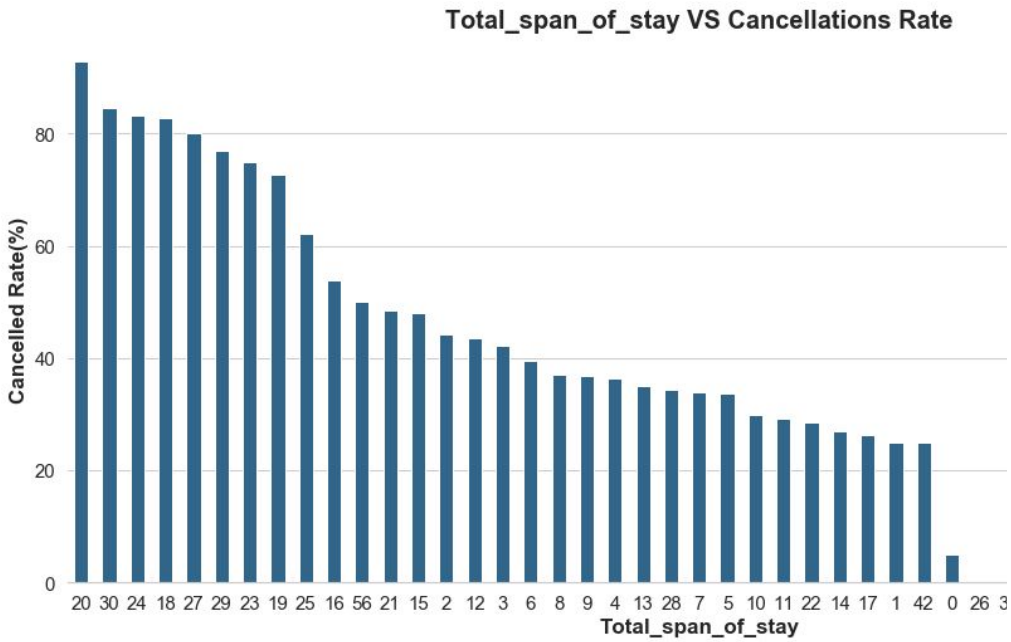
- Lead time of cancellations is usually located below 100 days.
- As the lead time increases, the probability of cancellation increases. But it is not a linear relationship.

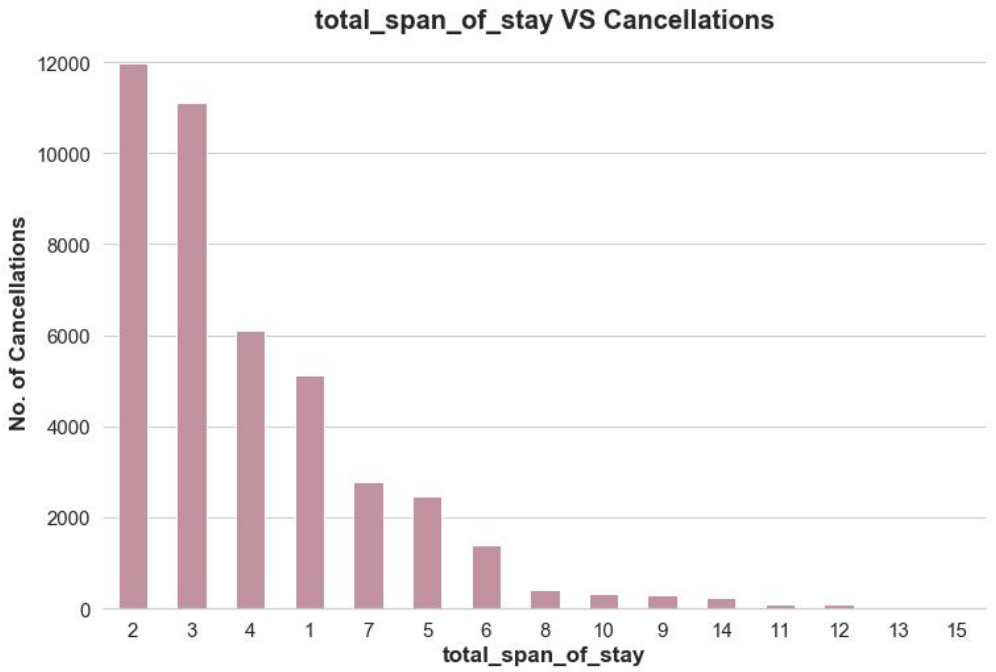




17.The relationship of cancellation rate and planing stay time

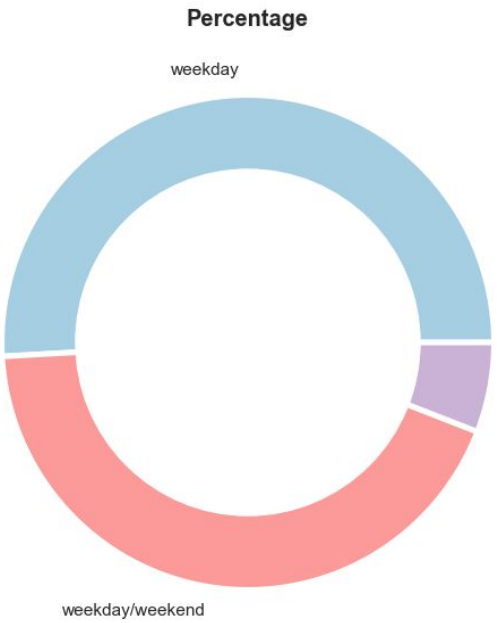
- 20-30 days stay time has the largest cancellation rate.
- But 2-4 days stay time has the largest cancellation number.

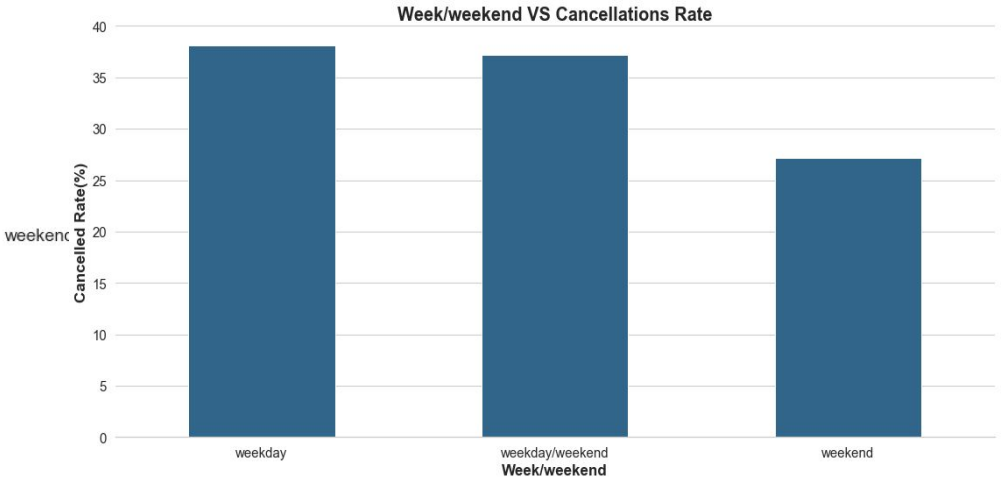




18.When are people more likely to cancel the booking? Weekends or week days?

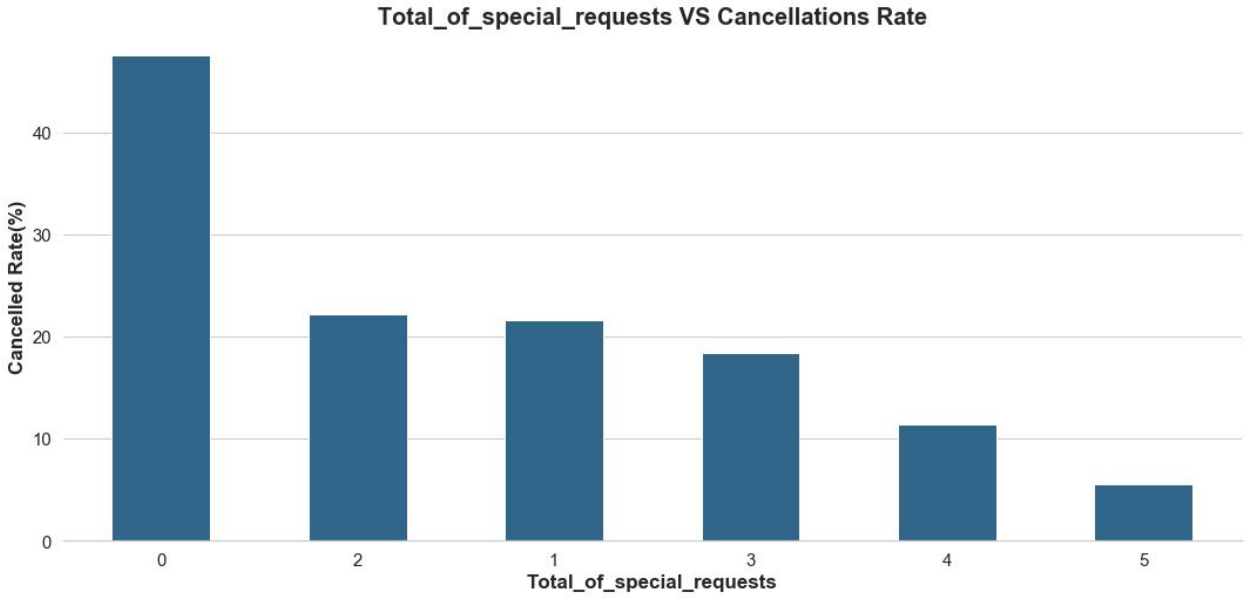
- Weekend has the smallest cancellations.
- People are more likely to cancel booking on weekdays.





19. Will the number of special requests affect cancellation rate?

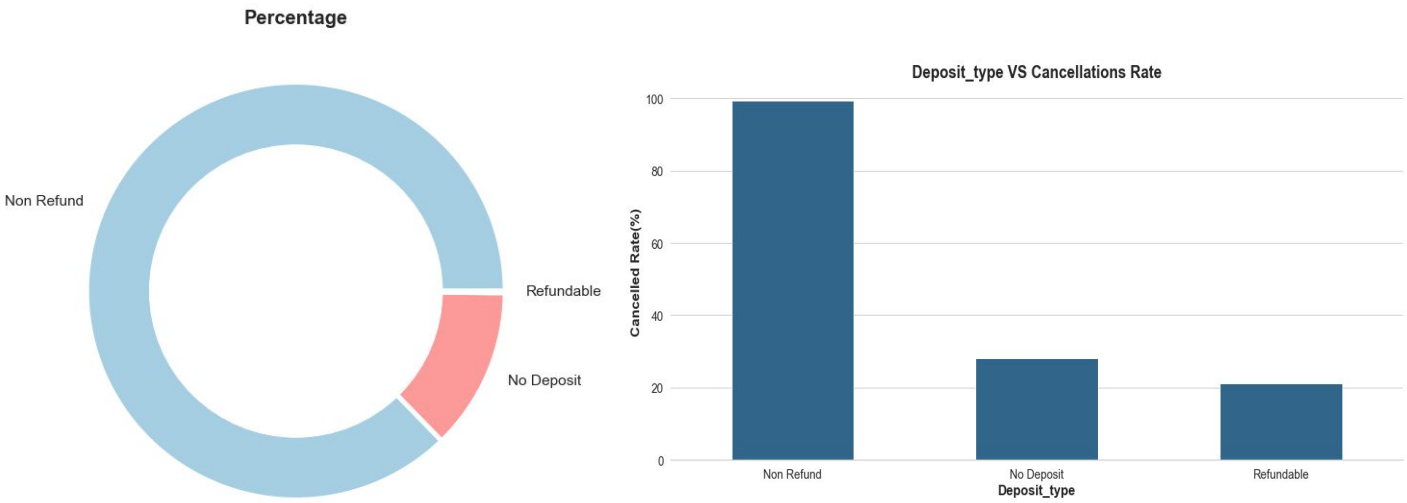
- It is obvious that people without any request had the largest cancellation rate. As the number of special requests increases, the rate comes down.

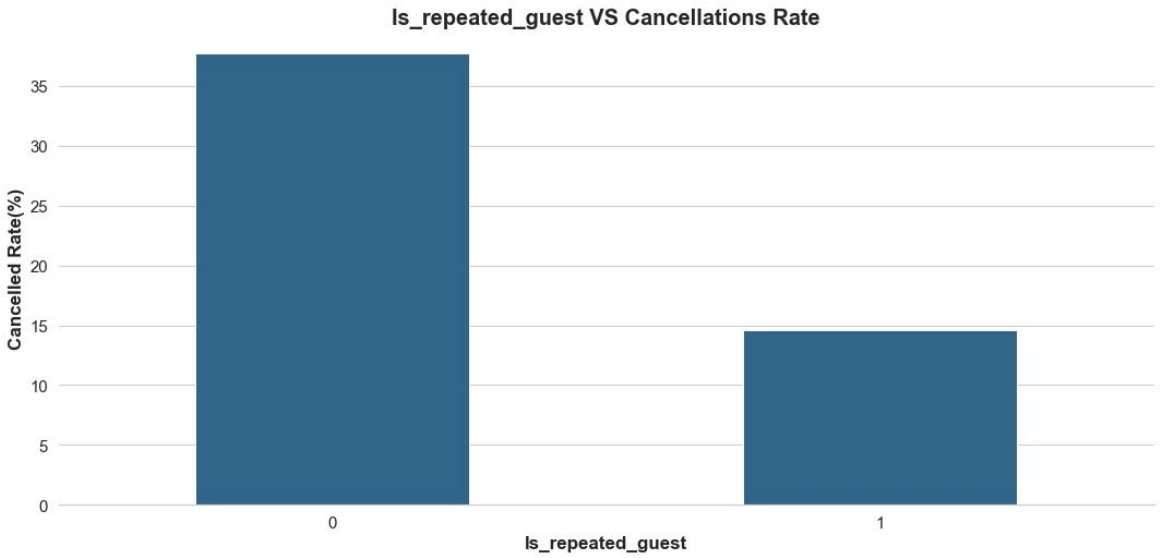


20.some additional questions

- Compared with refundable and no deposit customers , non refund customers are more likely to cancel the bookings. That is really weird, different from common sense. In fact the customers who had business trips took a large proportion of the cancellation. The contracts between the companies and the hotels may interpret the non refund data.

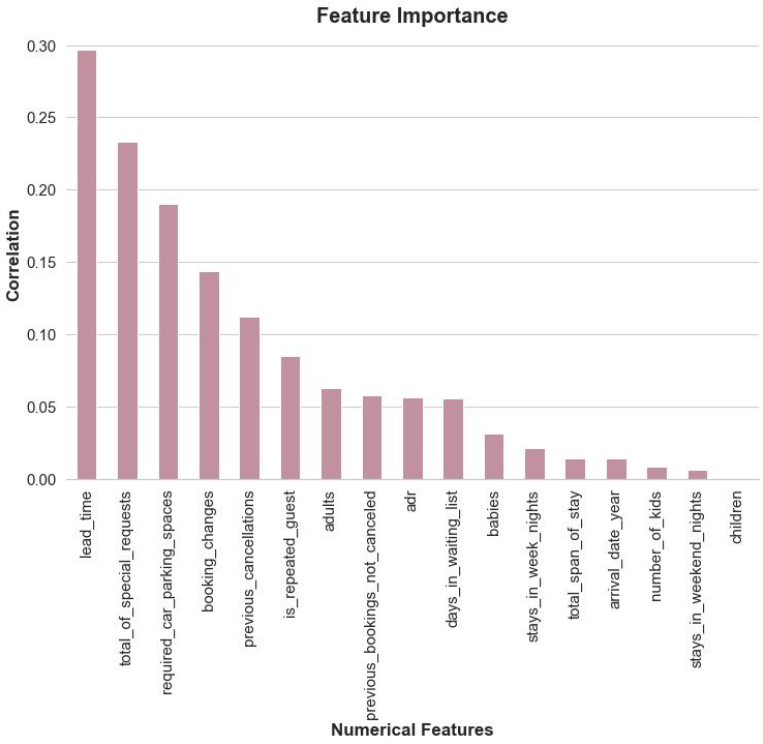
- Repeated guests had a higher cancellation rate.





4.STATISTICAL ANALYSIS

4.1 evaluate the correlation of all the features



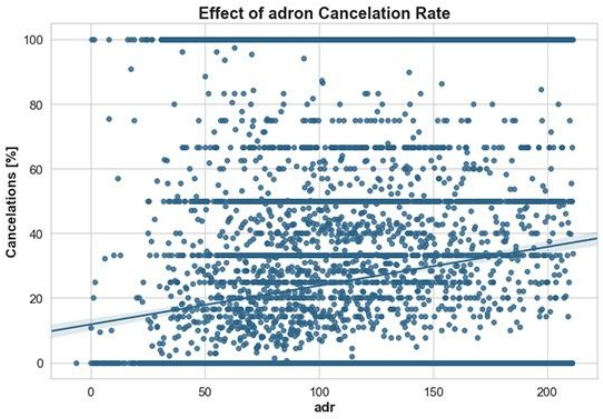

```
lead_time                       0.296587
total_of_special_requests       0.233214
required_car_parking_spaces     0.190400
booking_changes                 0.143636
previous_cancellations          0.111872
is_repeated_guest               0.084991
adults                          0.062545
previous_bookings_not_canceled  0.057877
adr                             0.056329
days_in_waiting_list            0.055710
babies                          0.030956
stays_in_week_nights            0.021369
total_span_of_stay              0.013695
arrival_date_year               0.013650
number_of_kids                  0.008493
stays_in_weekend_nights         0.005782
children                        0.000444
Name: is_canceled, dtype: float64
```

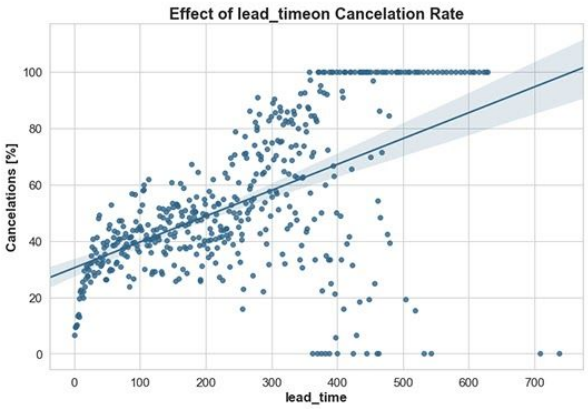
- lead_time, total_of_special_requests, required_car_parking_spaces, booking_changes and previous_cancellations are the 5 most important numerical features.
- booking_change Will be affected by target variables, so I won't include it.

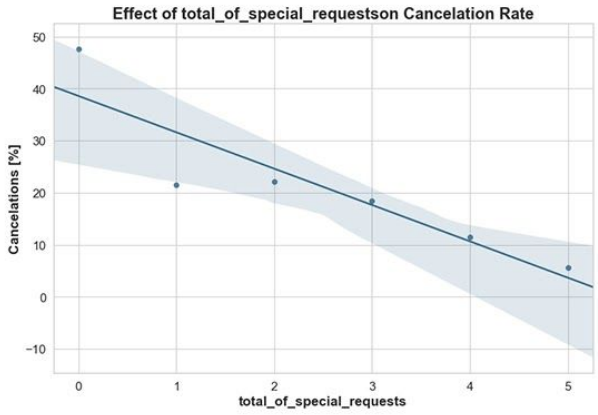
4.2 Research the most important numerical features

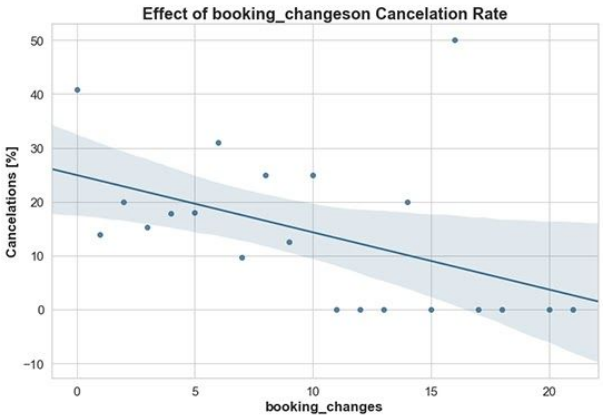
After researching the scatter diagrams, none of the features had a linear correlation with the cancellation rate.

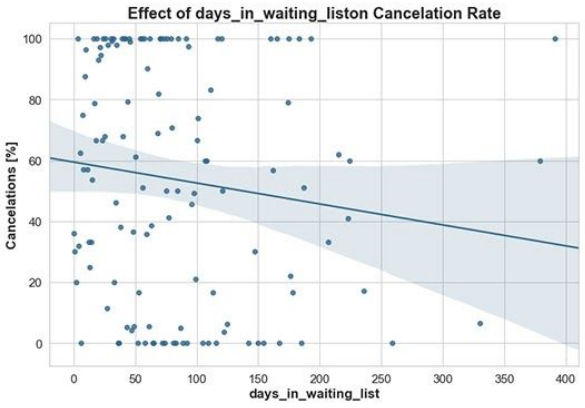
However, I could find out some law between the lead time and cancellation rate. It seems that before 50 days lead time the cancellation raised quite fast. After that it is obviously slowed down.











5.Modeling

4.1 Modeling Preparation

Firstly I need to transfer all the variables into numerical ones.Since there were more than two classes in one column. I decided to use the Label Encoding method.

Secondly I split the data into dependent and independent variables. Dependent variable was also considered as the target, ‘is canceled’ column.

4.2 Modeling Comparison

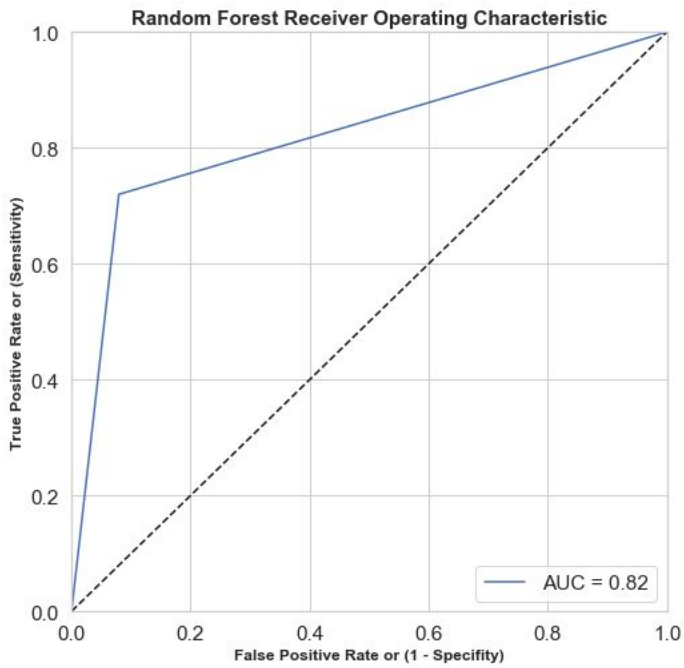
Since it is a classification supervised problem, I took ‘DummyClassifier’, ’RandomForestClassifier‘,’LogisticRegression’, ’KNeighborsClassifier’, ’GradientBoostingClassifier’,’DecisionTreeClassifier’ and ‘XGBRegressor’ as modeling methods.

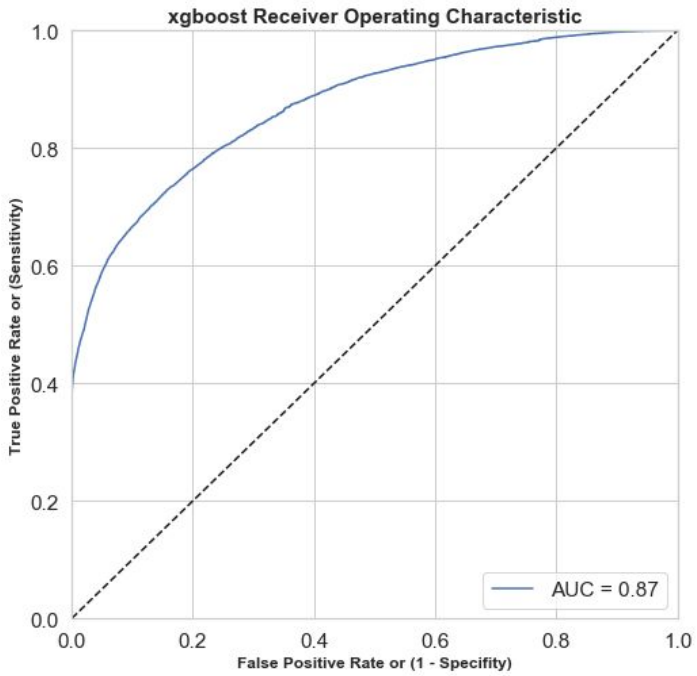
I took 3 metrics to compare the modeling methods: auc, Accuracy score, and K-fold cross validation.

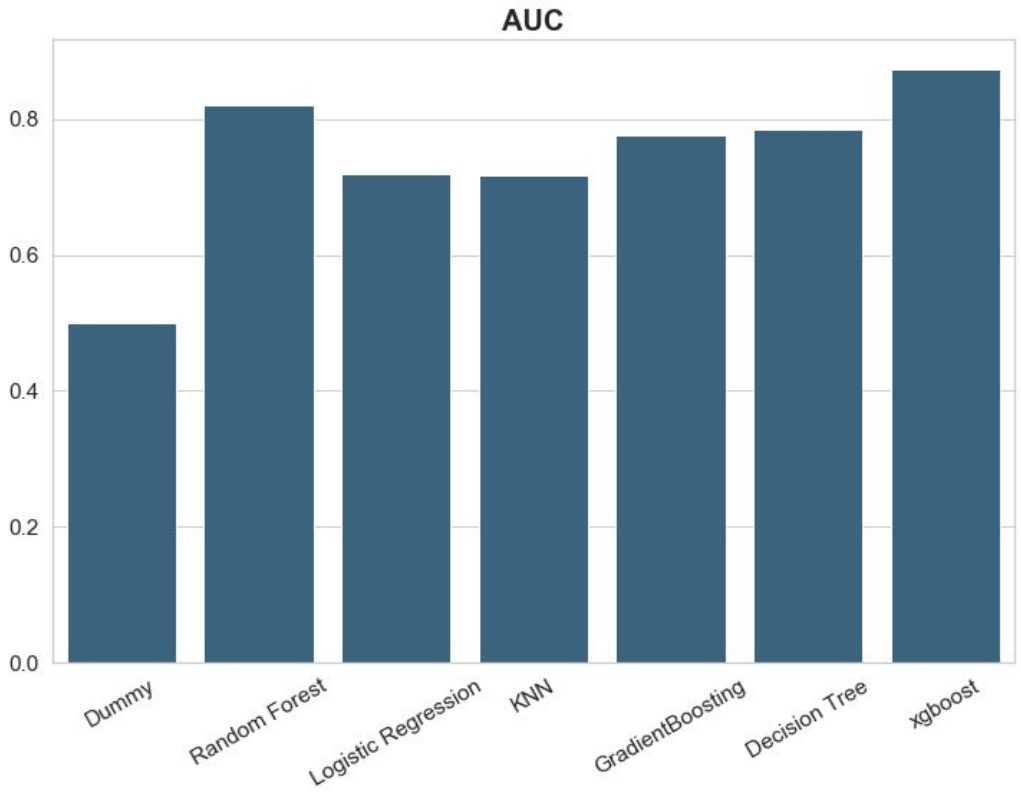
4.2.1 AUC score

After comparing the 6 modeling methods, xgboost got the highest AUC score, which is 0.87, followed by RandomForest.

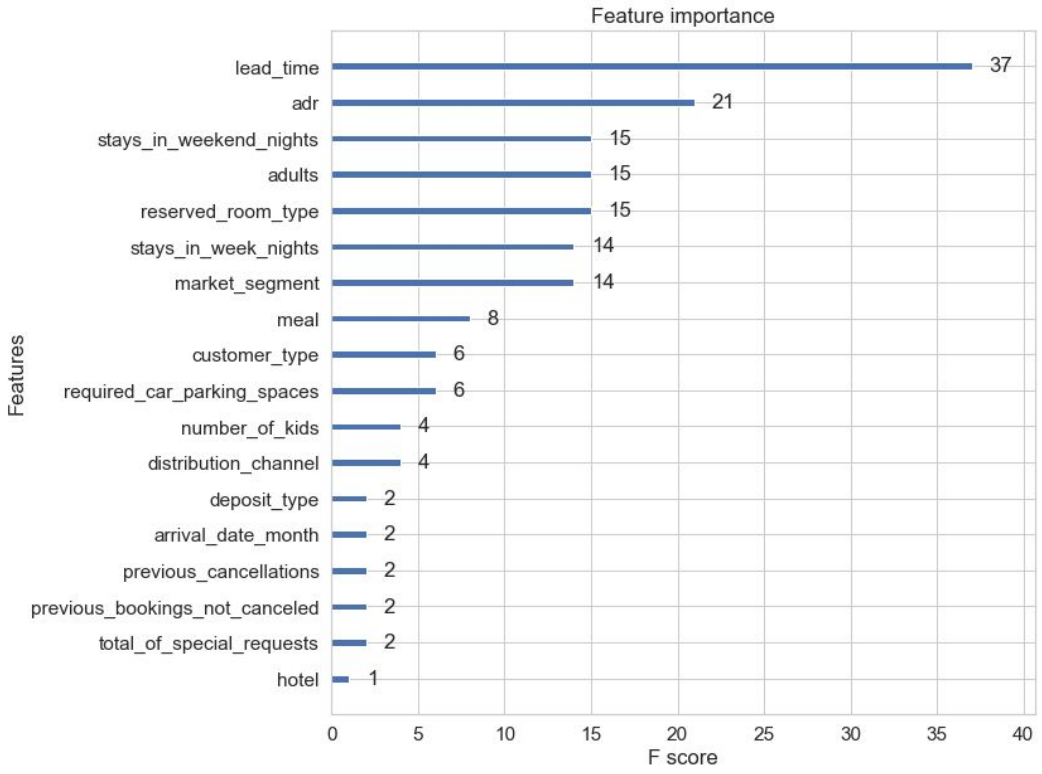
	Dummy	Random Forest	Logistic Regression	KNN	GradientBoosting	Decision Tree	xgboost
0	0.50042	0.820201	0.719212	0.716848	0.7769	0.783842	0.873781



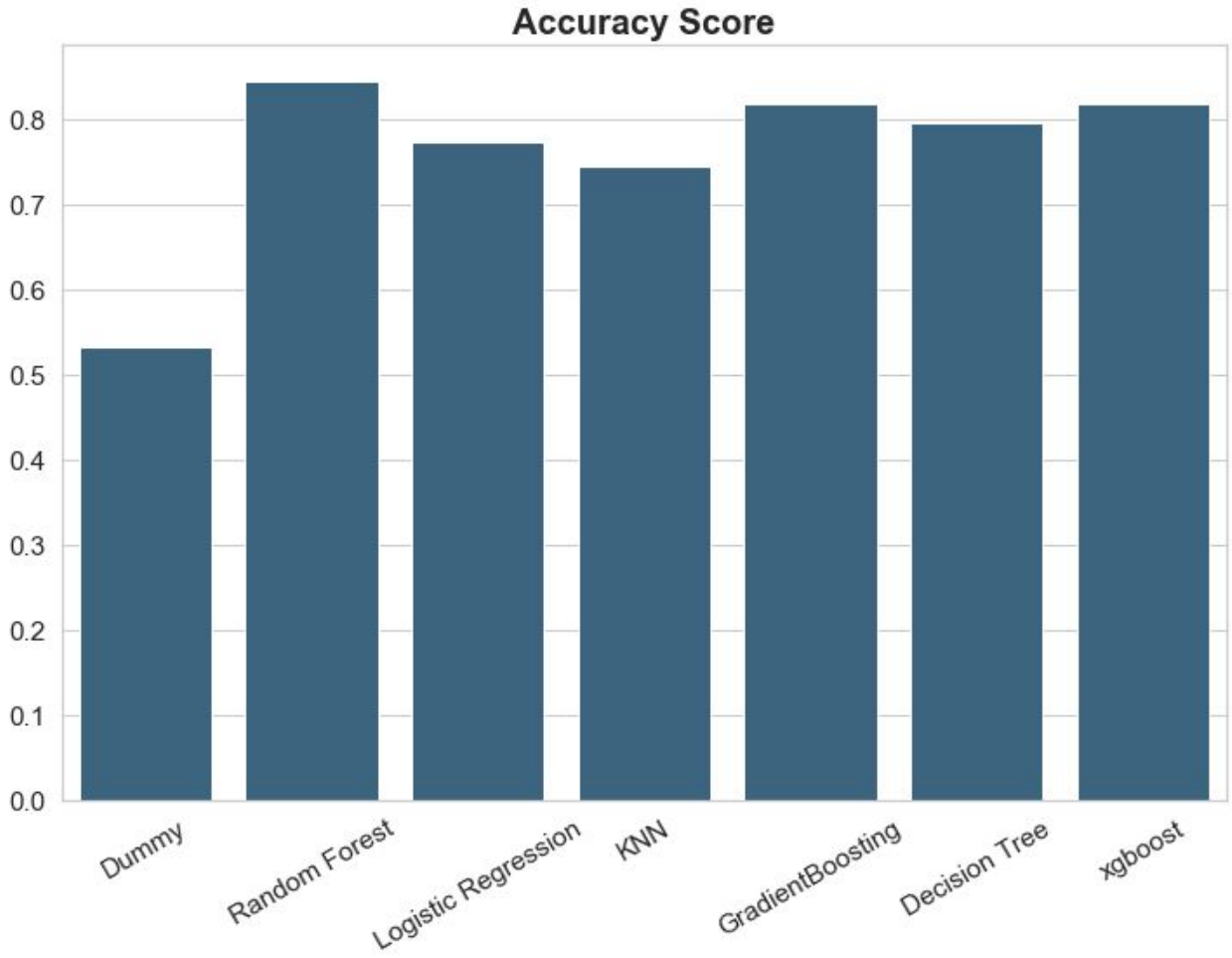




I checked the feature importance of xgboost, I got the following plot:



4.2.2 Accuracy score



Random Forest had the greatest performance in accuracy score.

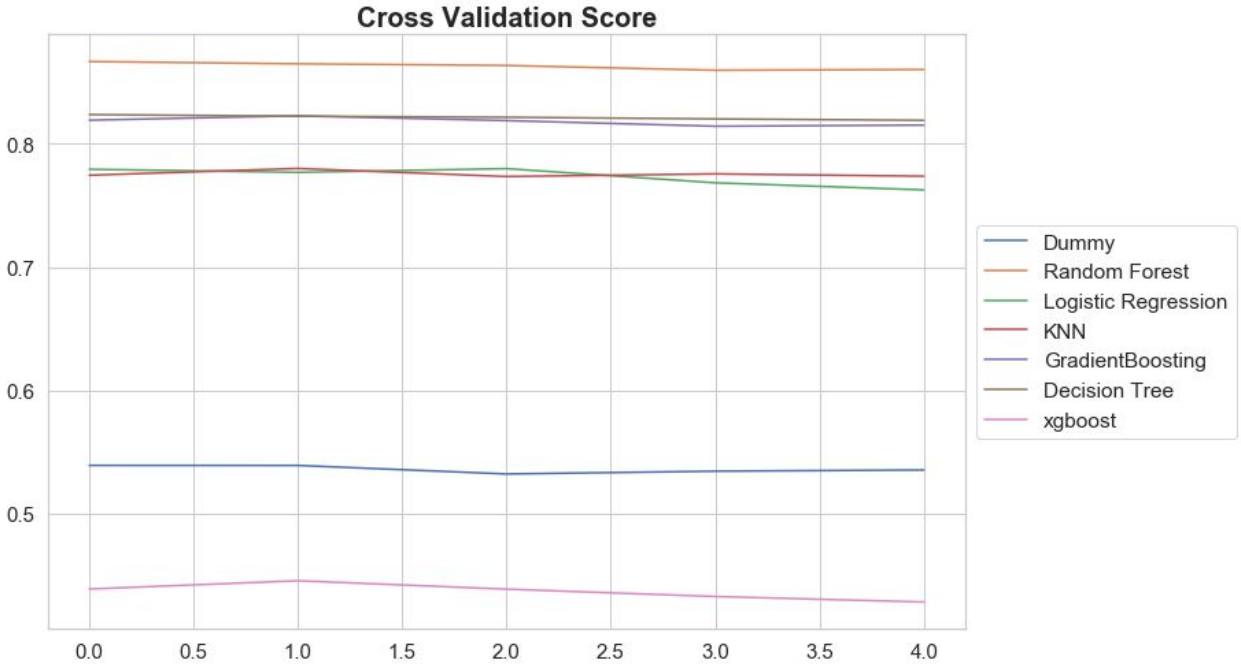
	Dummy	Random Forest	Logistic Regression	KNN	GradientBoosting	Decision Tree	xgboost
0	0.533497	0.845545	0.774436	0.745789	0.817739	0.796421	0.818048

4.2.3 Cross Validation

:

	Dummy	Random Forest	Logistic Regression	KNN	GradientBoosting	Decision Tree	xgboost
count	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000
mean	0.536252	0.863091	0.773502	0.775553	0.818106	0.821463	0.437089
std	0.003036	0.003025	0.007597	0.002669	0.003391	0.001823	0.006572
min	0.532376	0.859726	0.762706	0.773563	0.814352	0.819153	0.428555
25%	0.534625	0.860331	0.768459	0.773822	0.815260	0.820278	0.433015
50%	0.535620	0.863619	0.776937	0.774601	0.818894	0.821619	0.438999
75%	0.539297	0.864960	0.779445	0.775682	0.819240	0.822440	0.439092
max	0.539340	0.866819	0.779965	0.780094	0.822786	0.823825	0.445787

Random Forest had the greatest performance after K-fold Cross validation.



After hyperparameter optimization, there is little improvement of RF.

4.3 Feature Importance of Random Forest

```
 1) lead_time                      0.201259
 2) adr                            0.150833
 3) deposit_type                   0.147614
 4) arrival_date_month             0.072284
 5) total_of_special_requests      0.060666
 6) stays_in_week_nights           0.060448
 7) market_segment                 0.055489
 8) previous_cancellations         0.053882
 9) stays_in_weekend_nights        0.037192
10) customer_type                  0.031932
11) reserved_room_type             0.025205
12) required_car_parking_spaces    0.021722
13) adults                         0.020920
14) meal                           0.018058
15) hotel                          0.013272
16) distribution_channel           0.012901
17) number_of_kids                 0.009293
18) previous_bookings_not_canceled 0.004957
19) is_repeated_guest              0.002074
```

As we can see, lead_time, adr, deposit_type are the most important features, which is a little different from Xgboost.

Inclusion:

•Xgboost has the best AUC Score of all, however with the lowest Cross Validation Score.
•Random Forest has the best average performance in all the metrics.
•‘Lead_time’, ‘ADR’ are the most important features in both Xgboost and Random Forest. According to different models, the feature importances may change.
•Hyperparameter Optimization cannot improve the performance greatly all the time.