Hotel booking demand:

This dataset set belongs to Hotel Booking[1], it has a summary about various booking information of two hotels: a resort hotel and a city hotel. I have used the logistic regression and random classifier algorithm to predict when a customer is going to cancel their booking or not.

This dataset contains total of 32 columns and 119390 rows.

Columns	Description	Data Types
babies	Number of babies	float 64
Days_in_waiting_list	No of days in waiting	Int 64
Is_repeated_guest	Repeated guest (1) or not (0)	Int 64
agent	Travel agency ID	Float 64
children	Number of children	Float 64
Distribution_channel	Booking distribution channel	object
meal	Type of meal	object
adults	Number of adults	Int 64
Booking_changes	Number of changes to booking	Int 64
Required_car_parking_space	No of car spaces req	Int 64
Customer_type	Type of customer	object
Stays_in-weekend_nights	Number of weekend nights	Int 64
Previous_cancellations	No of previous cancellations	Int
Total_of_special_requests	No of special request	Int
Market_segment	Market segment designation	Object
Deposit_type	Type of deposit	Object
Country	Country of origin	Object
Adr	Average Daily Rate	Float
Lead_time	No of days elapsed before entering date	Int
Is_canceled	Booking canceled or not	Int

Fig: 2

Setup of target variable:

Here our target variable is whether a customer is going to cancel their booking or not, So our target variable is 'is_canceled'.

Data Preprocessing and Cleaning:

Firstly I have checked is there any null data in our dataset.

```
hotel                                0
is_canceled                          0
lead_time                            0
arrival_date_year                    0
arrival_date_month                   0
arrival_date_week_number             0
arrival_date_day_of_month            0
stays_in_weekend_nights              0
stays_in_week_nights                 0
adults                               0
children                             4
babies                               0
meal                                 0
country                            488
market_segment                       0
distribution_channel                 0
is_repeated_guest                    0
previous_cancellations               0
previous_bookings_not_canceled       0
reserved_room_type                   0
assigned_room_type                   0
booking_changes                      0
deposit_type                         0
agent                            16340
company                         112593
days_in_waiting_list                 0
customer_type                        0
adr                                  0
required_car_parking_spaces          0
total_of_special_requests            0
reservation_status                   0
reservation_status_date              0
dtype: int64
```

Fig: 6

We can observe there are about 16340 null values in agent column and 112593 null values in company variable. So I have filled all the null values in agent column with 'no agent' and similarly replaced all null values in company with 'no company'. I have removed some unnecessary variables like arrival date, arrival month, arrival date week, arrival date of month because these variables doesn't contribute in prediction as they are arriving

[1] [1] "Hotel booking demand | Kaggle." https://www.kaggle.com/jessemostipak/hotel-booking-demand (accessed May 04, 2020).

Insights from the data:

I'm comparing all my input variables with my target variable to check how they are influential or not, as follows:

is_canceled	hotel	0
0	City Hotel	46228
	Resort Hotel	28938
1	City Hotel	33102
	Resort Hotel	11122

Fig: 16

The hotel booking column has equal contribution on the target variable 'is canceled', so this variable is biased during prediction its not useful for prediction.

assigned_room_type	is_canceled	
A	0	41105
	1	32948
B	0	1651
	1	512
C	0	1929
	1	446
D	0	18960
	1	6362
E	0	5838
	1	1968
F	0	2824
	1	927
G	0	1773
	1	780
H	0	461
	1	251
I	0	358
	1	5
K	0	267
	1	12
L	1	1
P	1	12

reserved_room_type	is_canceled	
A	0	52364
	1	33630
B	0	750
	1	368
C	0	624
	1	308
D	0	13099
	1	6102
E	0	4621
	1	1914
F	0	2017
	1	880
G	0	1331
	1	763
H	0	356
	1	245
L	0	4
	1	2
P	1	12

Fig: 17

Here both in the assigned room type and reserved room type most of the cancellation has been done on the hotel type A. Hotel A it self has 80% of cancellation so it's baised column so we are removing these two columns.

total_of_special_requests	is_canceled	
0	0	36762
	1	33556
1	0	25908
	1	7318
2	0	10103
	1	2866
3	0	2051
	1	446
4	0	304
	1	36
5	0	38
	1	2

Fig: 18

Here we can see the no of special requests do have influence on the target variable which means if a customer have a special request then they are almost interested to continue with booking.

is_canceled	required_car_parking_spaces	
0	0	67750
	1	7383
	2	28
	3	3
	8	2
1	0	44224

Fig: 19

When a customer request for a car parking they are definitely going to continue with their booking. The number of people who are cancelling are too less 7383.

So are comparing all the variables, here we are dropping ‘reserved room type’, ‘assigned room type’, ‘reservation status date’ & ‘reservation status’. As all these are skew variables.

Correlation Plot:

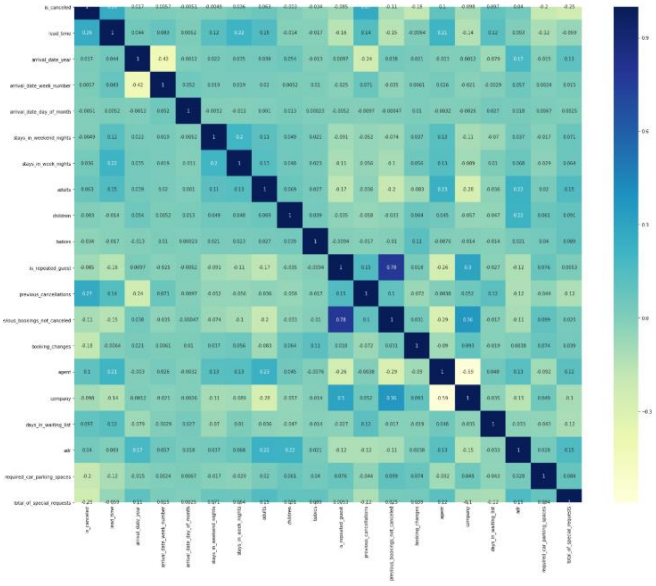

Fig: 20

After looking at correlation plot we see the ‘previous booking not canceled” and ‘company’ are two highly correlated variables. So I’m removing these variables.

After cleaning I have ended up with below dataset:

```
is_canceled                     0
lead_time                       0
stays_in_weekend_nights         0
stays_in_week_nights            0
adults                          0
children                        0
babies                          0
meal                            0
country                         0
market_segment                  0
distribution_channel            0
is_repeated_guest               0
previous_cancellations          0
booking_changes                 0
deposit_type                    0
agent                           0
days_in_waiting_list            0
customer_type                   0
adr                             0
required_car_parking_spaces     0
total_of_special_requests       0
```

Fig: 21

Below is the target variable which is not imbalanced.

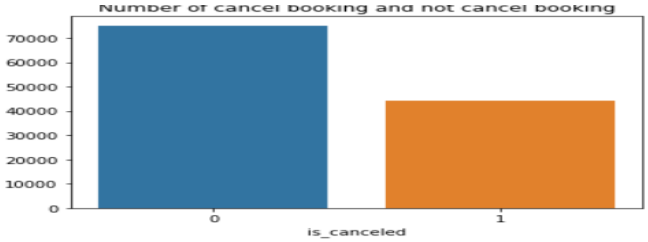

Fig: 22

Applying the data mining algorithm:

Logistic Regression:

As my target variable is categorical which has 0 & 1, so I have used logistic regression to predict which customers are going to cancel their booking or not. I have independent variables of 20 and one target variable.

Confusion Matrix:

Predict	0	1	All
Actual			
0	3293	1292	4585
1	1977	783	2760
All	5270	2075	7345

Accuracy is 0.71
Precision is 0.64
Recall is 0.50

Fig: 32

The accuracy of this model is 71% which means it is able to predict the bookings that are going to get cancelled or not. The precision tells about how precise the model is about to predict the true positive booking cancellations out of the predicted booking cancellation. So our model have a precision value of 0.64 which means its not able to classify the true positives to precisely. Similarly Recall tells about how many of the predicted positives are actually positive, we have a recall value of 0.5 which means it's a good precision value.

Random Forest Classifier:

I have used random forest classifier **f**or the better accuracy in predicting the booking cancellations and to know what are the influential variables that are acting on target variable in my dataset.

Confusion Matrix:

Actual			
0	3008	1592	4600
1	1836	935	2771
All	4844	2527	7371

Accuracy is 0.86
Precision is 0.85
Recall is 0.77

Fig: 33

After applying the random forest classification algorithm the accuracy of the model has improved to 86%, the classification of true booking cancellation out of predicted cancellation i.e. the precision indicator value has improved to 85% which is good value. The recall value is 77%, which is also a good improvement in classifying the true cancellations out of positively predicted cancellations.

Random Forest has an additional feature in ranking it all variables. Below are the influential variables in ascending order. Looking at it we can infer that lead time, adr and deposit type are the top 3 influential variables on the target variable

	0
babies	0.001181
days_in_waiting_list	0.003627
is_repeated_guest	0.004657
agent	0.006771
children	0.009906
distribution_channel	0.013760
meal	0.018372
adults	0.019500
booking_changes	0.022191
required_car_parking_spaces	0.026464
customer_type	0.028425
stays_in_weekend_nights	0.031323
previous_cancellations	0.036746
stays_in_week_nights	0.050259
total_of_special_requests	0.060666
market_segment	0.065016
country	0.133075
deposit_type	0.139547
adr	0.158374
lead_time	0.170140

Fig: 34

Conclusion:

After applying the classification algorithms on my dataset, we can infer that random forest classification has done good prediction in predicting the booking cancellation with an accuracy of 86% which is a 15% improvement from logistic regression. We got to know the top most influential variables are lead time, adr, deposit type by using random classifier algorithm. For the future work we can apply some high complex algorithms like extreme gradient boosting can bring the generalized classification accuracy and by including features like free parking availability, amenities provided, seasonality bookings can bring also help in better prediction.