Hotel booking demand

By shykah alrushud

In this project we will build a model based on our dataset “ **Hotel Booking demand** “ to predict the probability of the customer to cancel the booking or not.

This data set contains booking information for the hotel type , 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 .

This dataset contains :

119390 reservations

32 features

you can find it at Kaggle in this link:

Hotel booking demand dataset | Kaggle

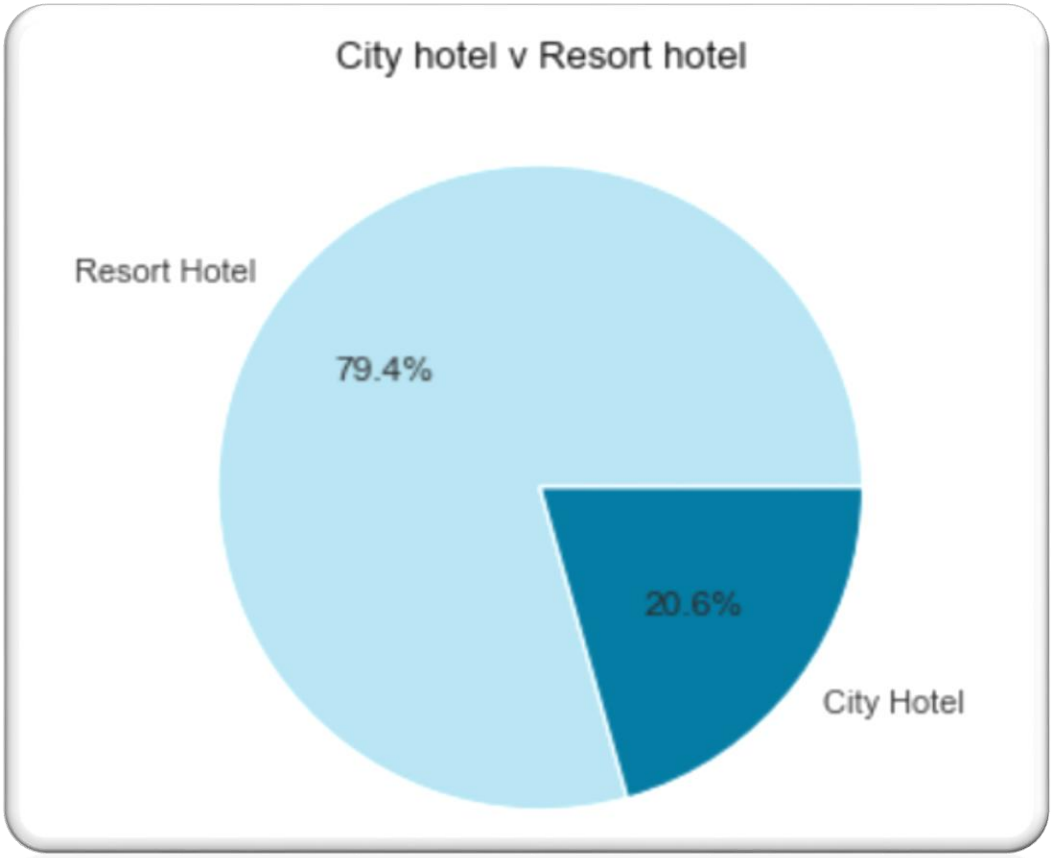
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weeke
0	Resort Hotel	0	342	2015	July	27	1	
1	Resort Hotel	0	737	2015	July	27	1	
2	Resort Hotel	0	7	2015	July	27	1	
3	Resort Hotel	0	13	2015	July	27	1	
4	Resort Hotel	0	14	2015	July	27	1	
...	...	...	...	...	...	...	...	
119385	City Hotel	0	23	2017	August	35	30	
119386	City Hotel	0	102	2017	August	35	31	
119387	City Hotel	0	34	2017	August	35	31	
`	City Hotel	0	109	2017	August	35	31	

Data Cleaning:

- Drop the duplicate rows.
- Drop the rows with column "***adult***=0 "
- Converting the datatype of ***children*** and ***agent*** from float to integer.
- Remove column **company , days_in_waiting_list, arrival_date_year, assigned_room_type, booking_changes,reservation_status, days_in_waiting_list.**
- Replace null values with 0 in ***agent*** feature.
- Replace NULL with "unknown" in column ***country***.
- Fill NULL value in ***adr*** with "mean".

- From this dataset I was curious to find answers for some questions next pages .

what is the hotel types and which one is more demand ?



Which type of hotel has the highest number of cancellations?

So, as you seen
Resort Hotel
has highest no. of cancellations

What is the percentage of repeat customers?

- Percentage of repeated guests = 7.936507936507937 %

Its just a little..

what is the busy month?

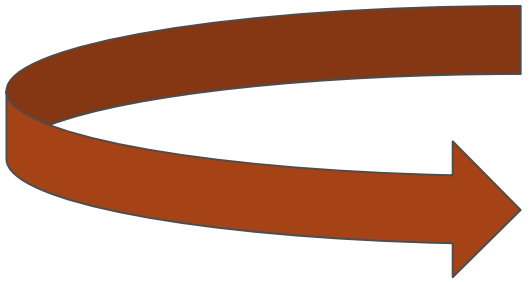
	month	no of guests
0	November	111
1	May	18
2	August	13
3	October	10
4	January	10
5	December	5
6	March	5
7	July	5
8	September	5
9	June	4
10	February	2
11	April	1

How many guest from each country ?

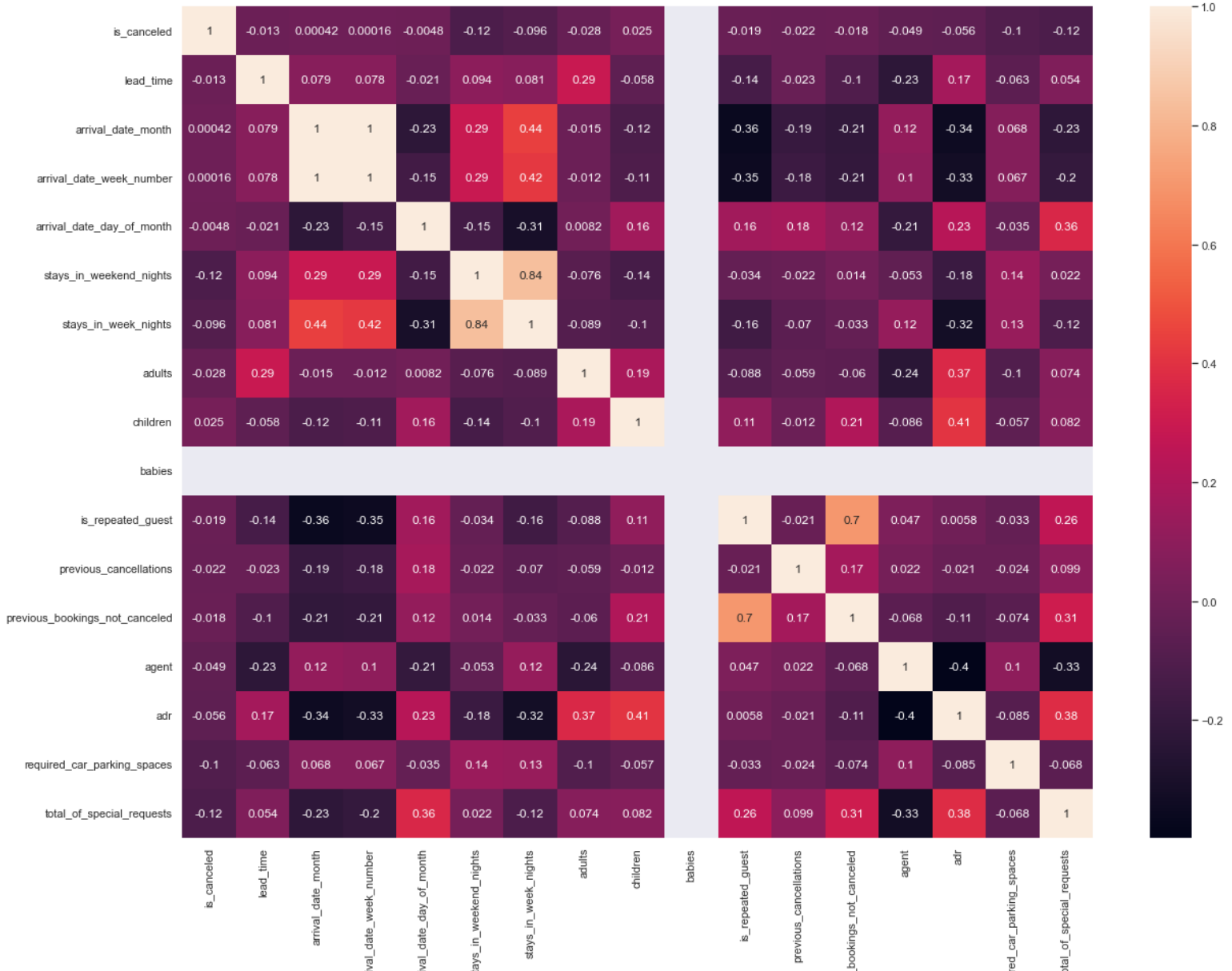
- As you seen **Portuguese** have the most no. of guest.

	Country	No.of Guests
0	PRT	77
1	AUT	23
2	FRA	20
3	GBR	19
4	ITA	15
5	ESP	8
6	DEU	7
7	CZE	4
8	BEL	3
9	MEX	2
10	CHN	2
11	NLD	2
12	USA	2
13	CHE	1
14	AUS	1
15	BGR	1
16	ROU	1

Building Machine Learning Models



Find out the highest relative correlated parameter for use as input for training ..



Modeling

I used the following *linear & classification algorithms :*

Random Forest Classifier , Decision Tree Classifier , Logistic Regression

- These are the Report scores of all the models I did:

Logistic Regression:

	precision	recall	f1-score
0	0.95	1.00	0.97
1	0.00	0.00	0.00
accuracy			0.95

Decision Tree Classifier

	precision	recall	f1-score
0	0.96	0.91	0.93
1	0.17	0.33	0.22
accuracy			0.88

Random Forest Classifier

	precision	recall	f1-score
0	0.96	1.00	0.98
1	1.00	0.33	0.50
accuracy			0.96

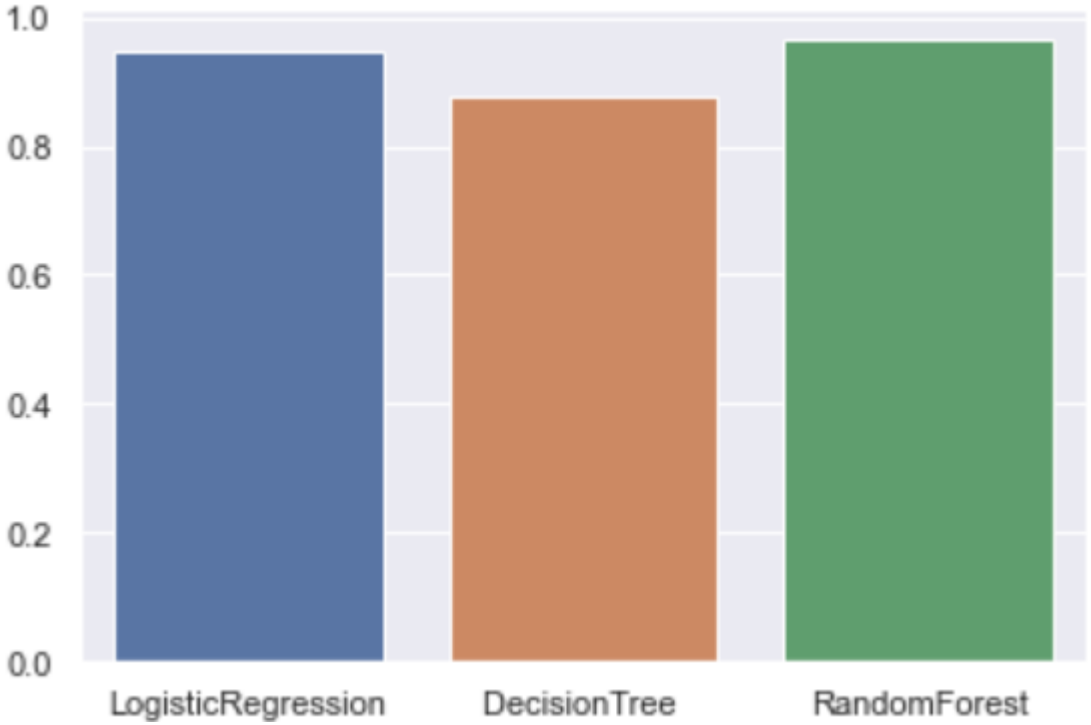
Modeling

- These are the comparison for accuracy test scores of all the models :

	Model	Score
2	Random Forest Classifier	0.964912
0	Logistic Regression	0.947368
1	Decision Tree Classifier	0.859649

The best Model to predict the target:
Its the Random Forest Classifier with accuracy test score 96%
in simple train and test techniques .

visualize the accuracy comparison between models



Thank you
I hope you like my work
And enjoying this presentation