Hotel Booking Cancellation Prediction

by Ruba Alnashwan

METHODOLOGY

- Proplem understanding
- Data collection
- Data cleaning
- Exploratory data analysis (EDA)
- Feature Engineering and selection
- Data modeling

Proplem understanding

Overview:
In this project, we will use data from the kaggle website, which provides information hotel and the label (cancel or not) . Our goal from this project is to build classification models that predict if the customer will cancel the booking or not.

Scope:
observation represents a hotel booking between the 1st of July 2015 and 31st of August 2017, including booking that effectively arrived and booking that were canceled. The dataset contains 119390 rows and 10 columns.

Data cleaning

- check nulls

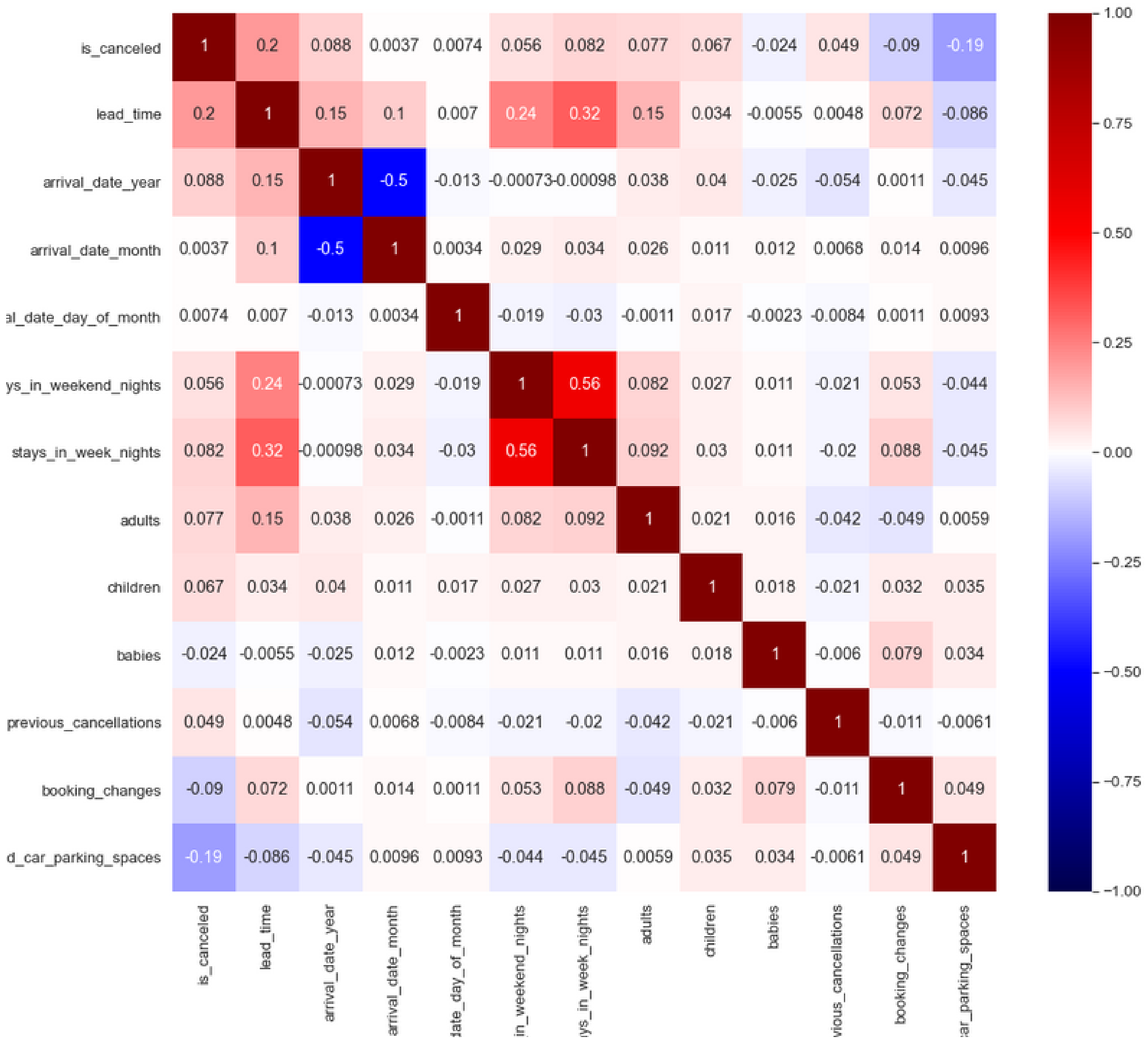
  I filled in the null values in the country ,children features with mode and mean.

- check duplicates and drop

- check outliers using EDA and remove

- replace some character to numbers like month

Exploratory data analysis (EDA)

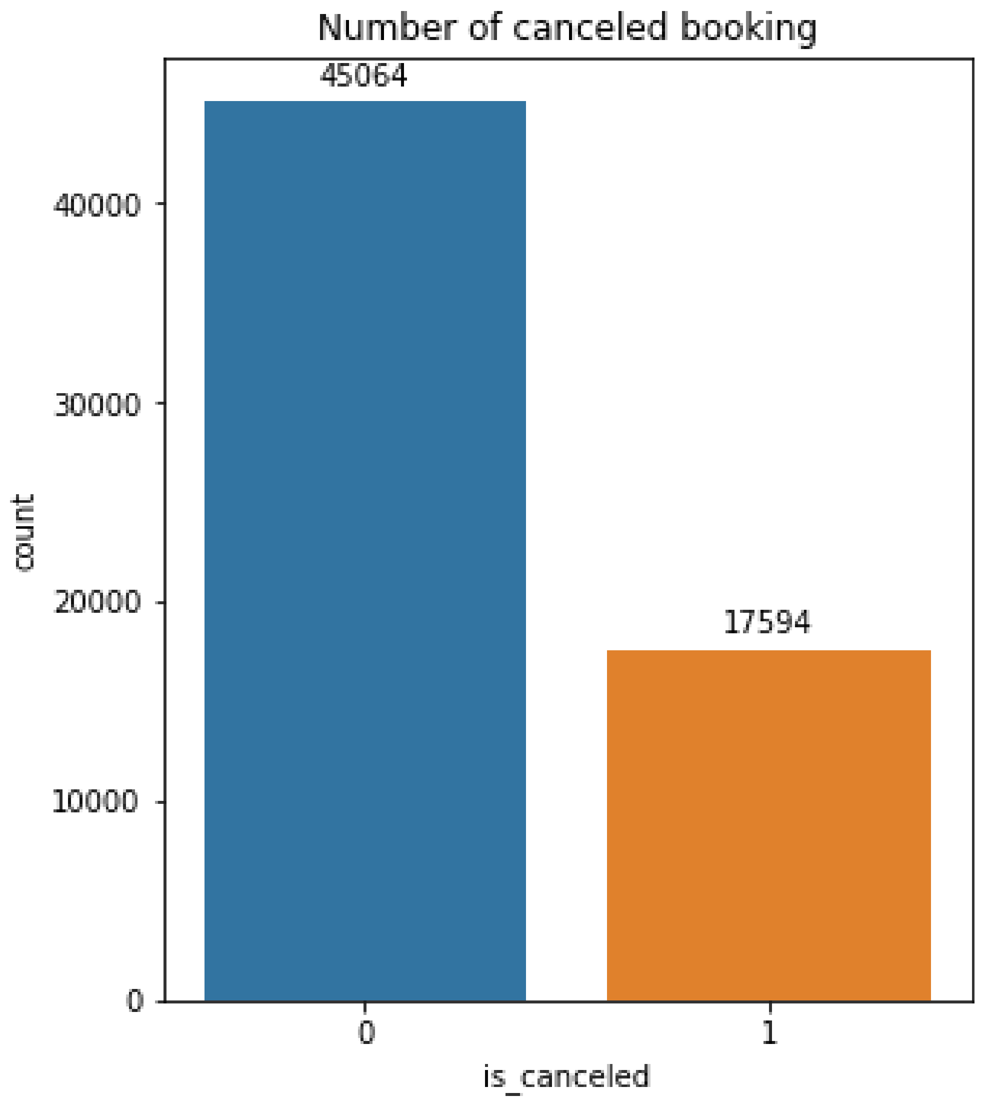


	is_canceled	lead_time	arrival_date_year	arrival_date_month	date_day_of_month	in_weekend_nights	ys_in_week_nights	adults	children	babies	vious_cancellations	booking_changes	ar_parking_spaces
is_canceled	1	0.2	0.088	0.0037	0.0074	0.056	0.082	0.077	0.067	-0.024	0.049	-0.09	-0.19
lead_time	0.2	1	0.15	0.1	0.007	0.24	0.32	0.15	0.034	-0.0055	0.0048	0.072	-0.086
arrival_date_year	0.088	0.15	1	-0.5	-0.013	-0.00073	-0.00098	0.038	0.04	-0.025	-0.054	0.0011	-0.045
arrival_date_month	0.0037	0.1	-0.5	1	0.0034	0.029	0.034	0.026	0.011	0.012	0.0068	0.014	0.0096
al_date_day_of_month	0.0074	0.007	-0.013	0.0034	1	-0.019	-0.03	-0.0011	0.017	-0.0023	-0.0084	0.0011	0.0093
ys_in_weekend_nights	0.056	0.24	-0.00073	0.029	-0.019	1	0.56	0.082	0.027	0.011	-0.021	0.053	-0.044
stays_in_week_nights	0.082	0.32	-0.00098	0.034	-0.03	0.56	1	0.092	0.03	0.011	-0.02	0.088	-0.045
adults	0.077	0.15	0.038	0.026	-0.0011	0.082	0.092	1	0.021	0.016	-0.042	-0.049	0.0059
children	0.067	0.034	0.04	0.011	0.017	0.027	0.03	0.021	1	0.018	-0.021	0.032	0.035
babies	-0.024	-0.0055	-0.025	0.012	-0.0023	0.011	0.011	0.016	0.018	1	-0.006	0.079	0.034
previous_cancellations	0.049	0.0048	-0.054	0.0068	-0.0084	-0.021	-0.02	-0.042	-0.021	-0.006	1	-0.011	-0.0061
booking_changes	-0.09	0.072	0.0011	0.014	0.0011	0.053	0.088	-0.049	0.032	0.079	-0.011	1	0.049
d_car_parking_spaces	-0.19	-0.086	-0.045	0.0096	0.0093	-0.044	-0.045	0.0059	0.035	0.034	-0.0061	0.049	1

Observations:

This heat map shows that there is no strong relationship between the dependent variable and features .

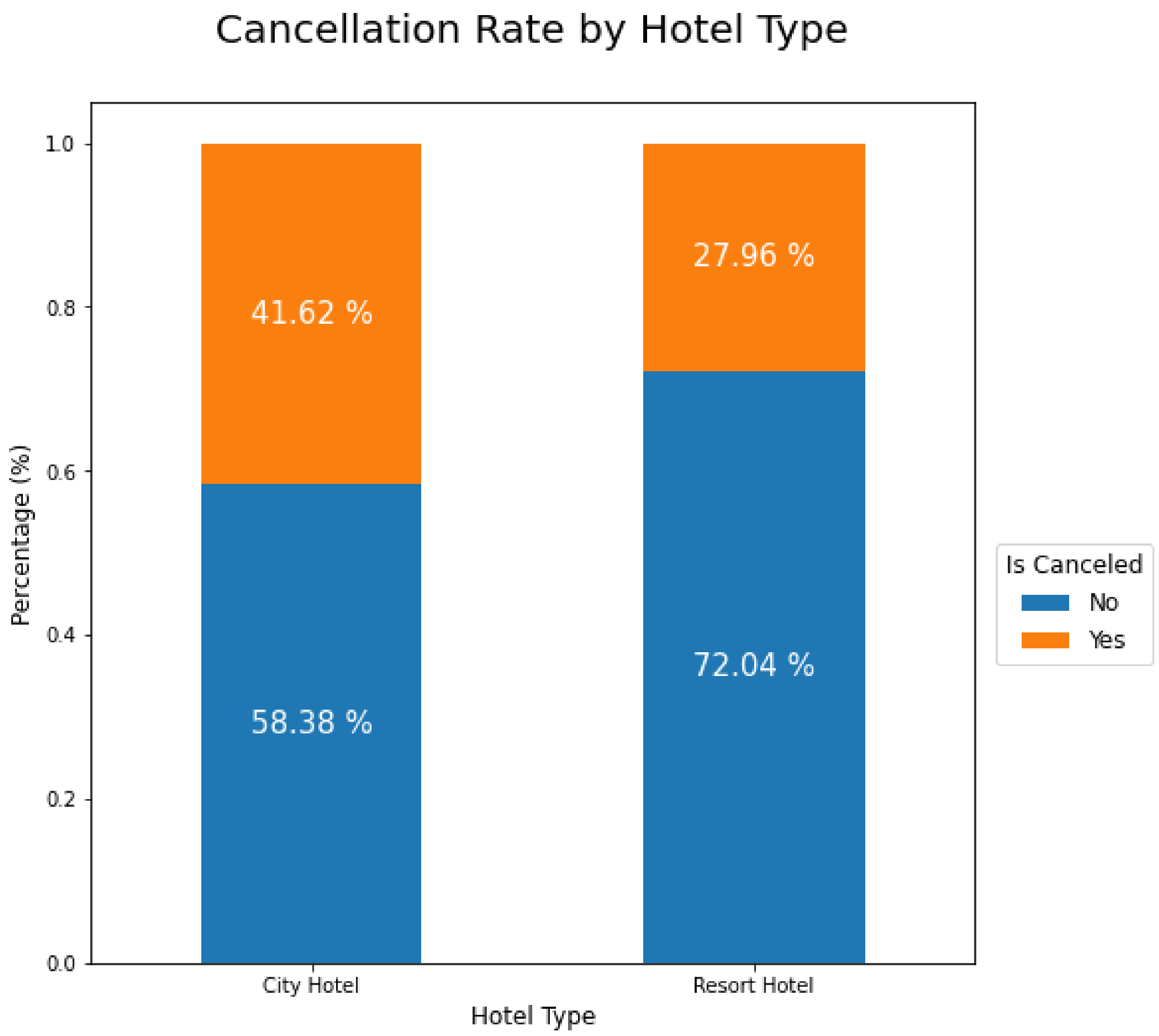
Exploratory data analysis (EDA)



Observations:

The par plot shows the Number of canceled booking

Exploratory data analysis (EDA)



Observations:

The par plot shows the Cancellation Rate by Hotel Type

Feature Engineering and selection

Combine several columns into one column

Dummies varible for categorical Feature

Scaling: StandardScale

Data modeling

MAIN METRIC USED FOR EVALUATION : F1

SECONDARY METRICS: PRECISION, RECALL AND ACCURACY

BASELINE MODEL USED: LOGISTIC REGRESSION

MODEL USED: RANDOM FOREST،DECISION TREE

Data modeling

basline

TRAIN SCORES:
F1 SCORE IS 0.457
PRECISION SCORE IS 0.669
RECALL SCORE IS 0.346

VALIDATION SCORES:
F1 SCORE IS 0.459
PRECISION SCORE IS 0.665
RECALL SCORE IS 0.350

dummies

TRAIN SCORES:
F1 SCORE IS 0.573
PRECISION SCORE IS 0.901
RECALL SCORE IS 0.421

VALIDATION SCORES:
F1 SCORE IS 0.575
PRECISION SCORE IS 0.903
RECALL SCORE IS 0.422

Remove outliers

TRAIN SCORES:
F1 SCORE IS 0.582
PRECISION SCORE IS 0.880
RECALL SCORE IS 0.4351

VALIDATION SCORES:
F1 SCORE IS 0.583
PRECISION SCORE IS 0.859
RECALL SCORE IS 0.441

Data modeling

RandomOverSampler

TRAIN SCORES:
F1 SCORE IS 0.674
PRECISION SCORE IS 0.777
RECALL SCORE IS 0.594

VALIDATION SCORES:
F1 SCORE IS 0.625
PRECISION SCORE IS 0.656
RECALL SCORE IS 0.596

TomekLinks

TRAIN SCORES:
F1 SCORE IS 0.597
PRECISION SCORE IS 0.862
RECALL SCORE IS 0.456

VALIDATION SCORES:
THE F1 SCORE IS 0.593
THE PRECISION SCORE IS 0.828
THE RECALL SCORE IS 0.4621

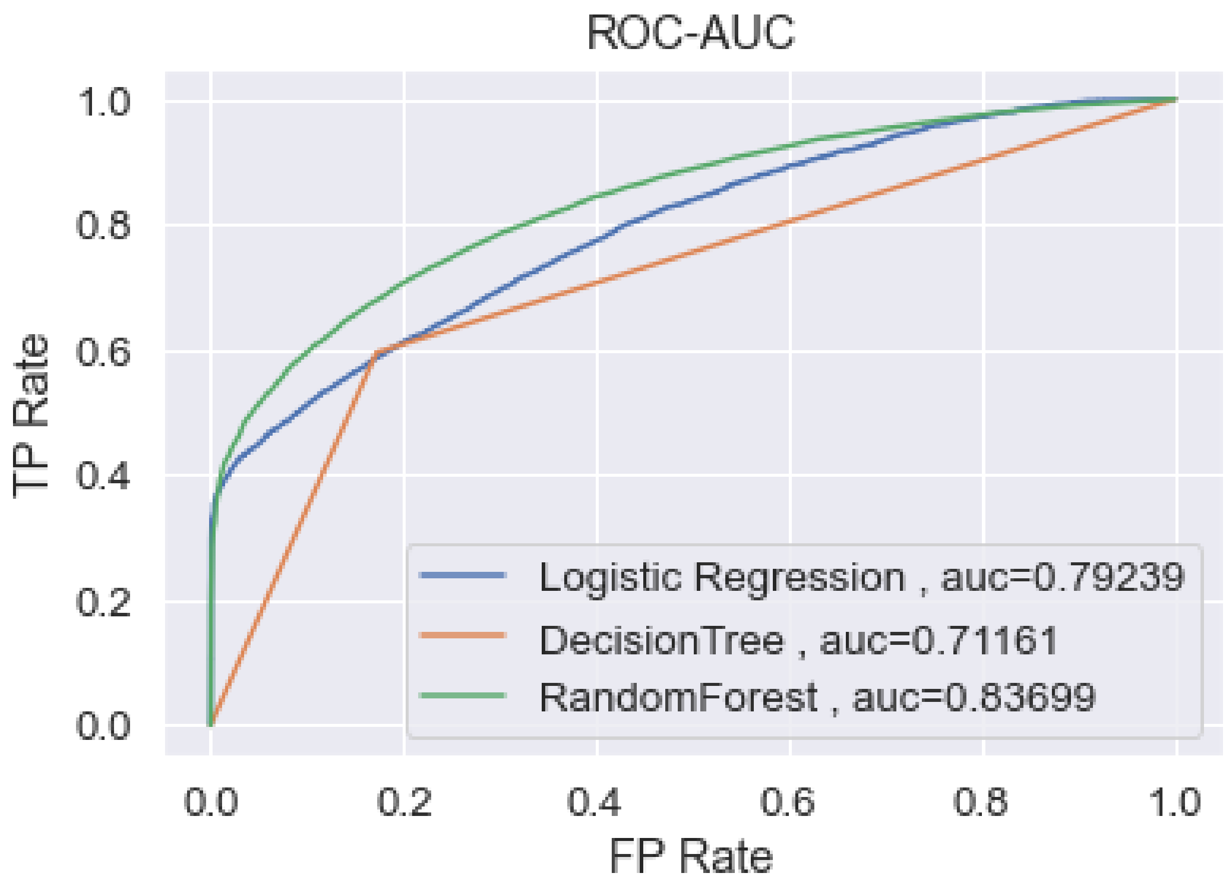
SMOTE

TRAIN SCORES:
F1 SCORE IS 0.682
PRECISION SCORE IS 0.763
RECALL SCORE IS 0.616

VALIDATION SCORES:
F1 SCORE IS 0.627
PRECISION SCORE IS 0.640
RECALL SCORE IS 0.615

CONCLUSION

Result the best model



Difficulties:

No strong relationship between the dependent variable and features
The data volume is large, so the training time is long

Thank you

Any questions?