Hotel Booking Cancellation Prediction

by Ruba Alnashwan

PROPLEM UNDERSTANDING 1

BACKGROUND

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.

DATA DESCRIPTION

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 from the data.

Features	Description	Type
hotel	Hotel (H1 = Resort Hotel or H2 = City Hotel)	object
is_canceled	Value indicating if the booking was canceled (1) or not (0)	int
lead_time	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date	int
arrival_date_month	Month of arrival date	object
arrival_date_week_number	Week number of year for arrival date	int
arrival_date_day_of_month	Day of arrival date	int
stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	int
stays_in_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	int
adults	Number of adults	int
children	Number of children. Sum of both payable and non-payable children.	float
babies	Number of babies.	int
meal	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), and FB (Full board: breakfast, lunch and dinner).	Object

PROPLEM UNDERSTANDING 2

country	Country of origin. Categories are represented in the International Standards Organization (ISO)	Object
distribution_channel	Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators".	Object
previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking. In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and canceled.	Object
reserved_room_type	Code of room type reserved. Code is presented instead of designation for anonymity reasons.	Object
booking_changes	Number of changes/amendments made to the booking from the moment the booking was entered on the Property Management System until the moment of check-in or cancellation. Calculated by adding the number of unique iterations that change some of the booking attributes, namely: persons, arrival date, nights, reserved room type or meal.	int
deposit_type	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), and Refundable (a deposit was made with a value under the total cost of stay).	Object
required_car_parking_spaces	Number of car parking spaces required by the customer.	int

SCOPE

In this project,We used data for booking hotel, 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 19 columns.

TECHNOLOGIES AND LIBRARIES

- Python
- LinearRegression
- sklearn
- imblearn.over_sampling
- xgboost
- Pandas
- NumPy
- Seaborn
- Matplotlib
- heatmap
- StandardScaler

ALGORITHMS 3

FEATURE ENGINEERING

- Combine several columns into one column.
- Dummies varible for categorical Feature
- Scaling: StandardScaler

MODELS

Main Metric used for evaluation : F1
Secondary Metrics: Precision, Recall and accuracy
Baseline Model used: Logistic Regression
Model used: Random Forest،Decision Tree,GridSearch, XGBClassifier

MODEL EVALUATION AND SELECTION

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

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

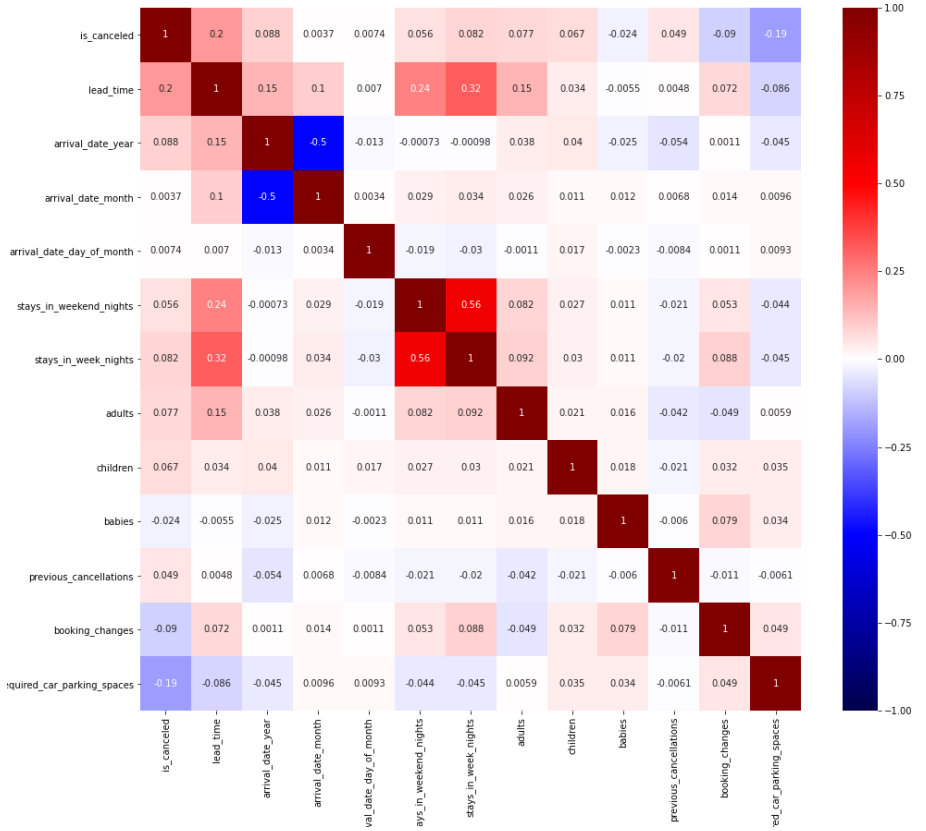
SMOTE

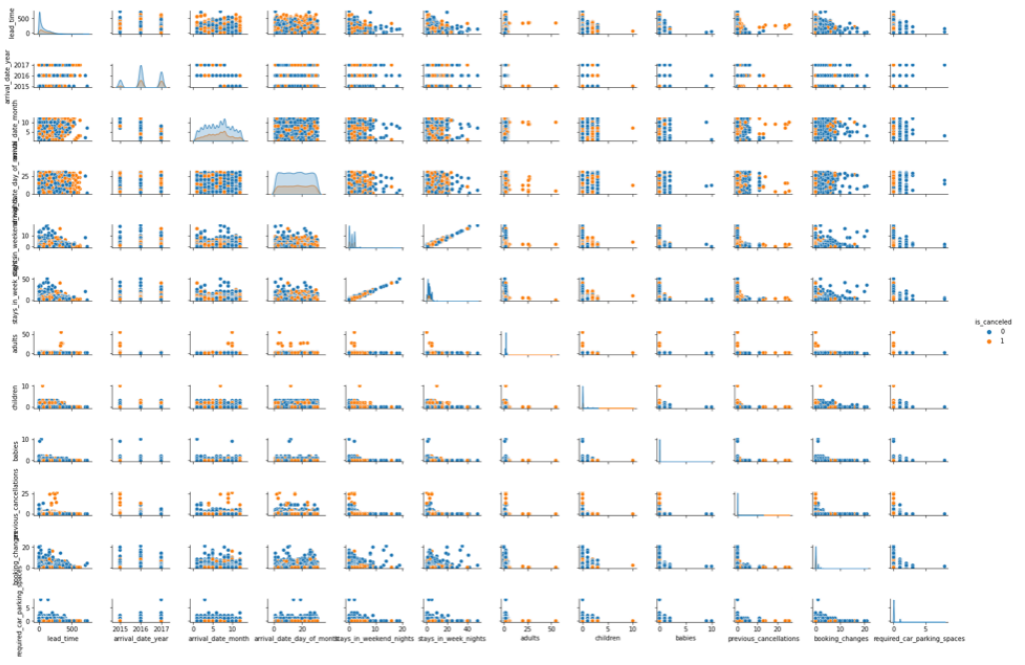
TRAIN SCORES:
F1 SCORE IS 0.682
PRECISION SCORE IS 0.763
RECALL SCORE IS 0.616

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

COMMUNICATION 4

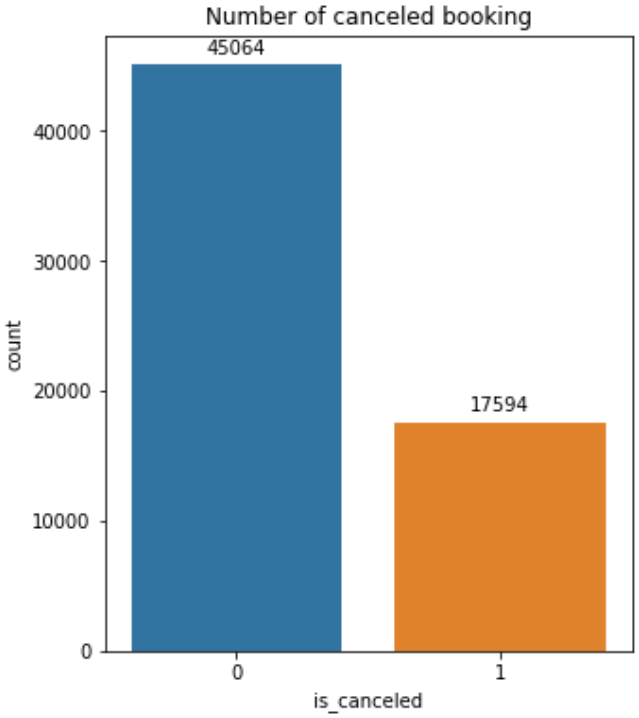
This heat map shows that there is no strong relationship between the dependent variable and features .



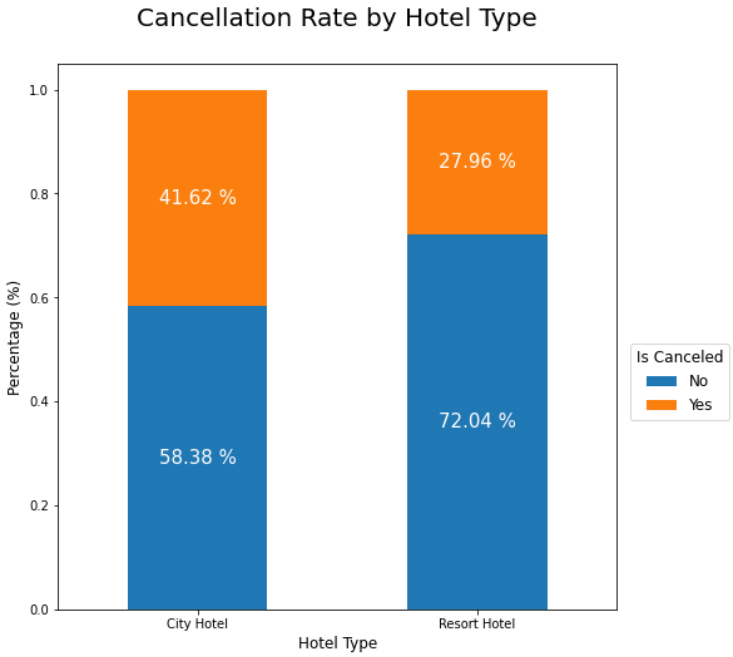


COMMUNICATION 5

The par plot shows the Number of canceled booking .



The par plot shows the Cancellation Rate by Hotel Type.



The best confusion matrix