LAPORAN

DATA ANALYTIC COMPETITION

FIND IT! 2020

Tim nakama



Disusun Oleh :

Elvin Nur Furqon  
Sebastian Cahyo Ardhi Iswara

Bandung

2020

Tools yang digunakan :

* Python Language Programming

Library yang digunakan :

* Pandas
* Matplotlib
* Seaborn
* Numpy
* Sklearn ( scikit-learn )

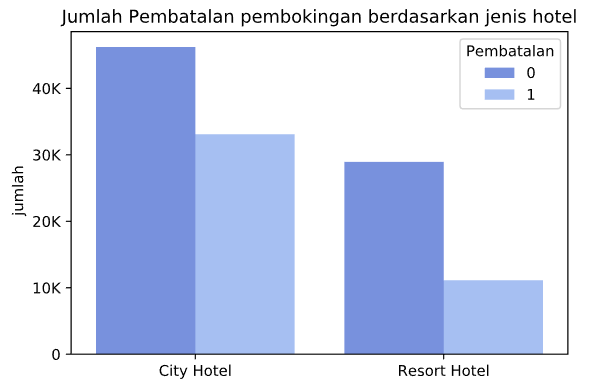
1.EDA ( Exploratory Data Analysis )

- Terdapat data sebanyak 119.390 dengan 31 feature, ukuran data yang terpakai kurang lebih sebesar 29 MB, kemudian terdapat 4 feature yang kosong (NULL)

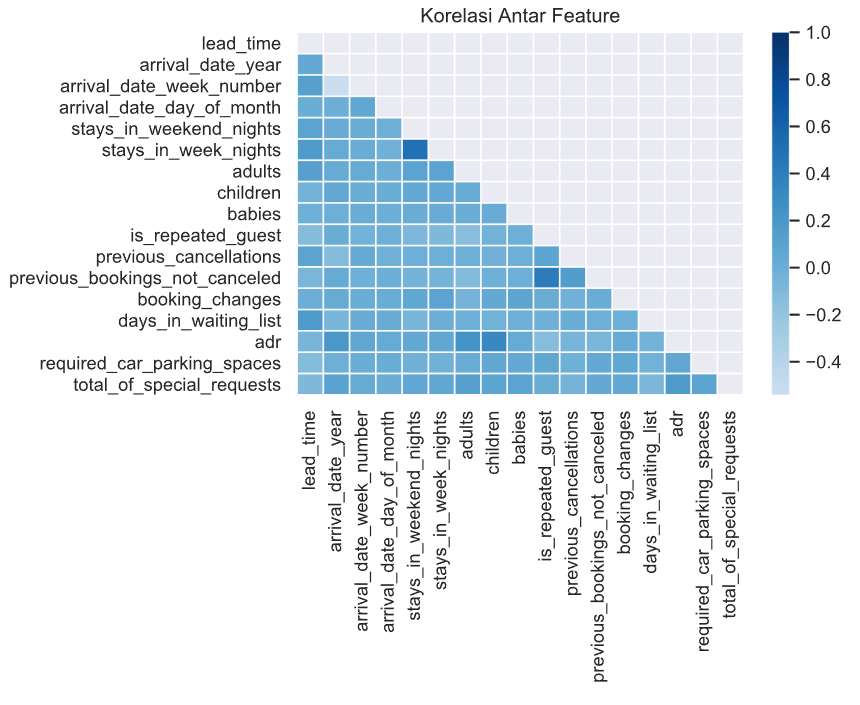
- Data terbanyak yang diberikan dari hotel City yaitu kurang lebih 2x nya dari Resort hotel. Hotel City ( 79330 ) , Resort hotel ( 40060 )

- dari seluruh data hotel disimpulkan untuk yang melakukan pembokingan kebanyakan datang berpasangan

- di City Hotel pembatalan hotel berkisar 33 ribu data dari 79 ribu, kemudian di resort hotel kurang lebih 10 ribu data dari 40 ribu.



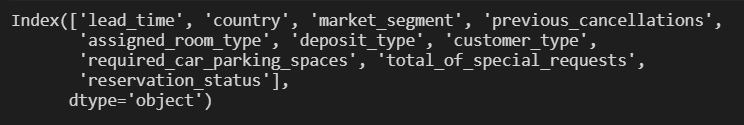
- pada 3 feature yang memiliki nilai kosong persentasi nilai kosong terbesar ke terendah secara berurut company (94%) , agent (13%) , country (0.4%) dan children ( 0.003%)

- Feature yang memiliki dominan korelasi tertinggi yaitu (stay\_in\_week\_nights, stay\_in\_weekend\_nights) dan (previous\_cancellations, is\_repeated\_guest

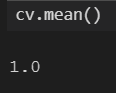
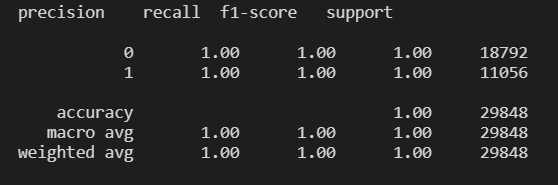
- Pada permasalahan yang diberikan, kami melakukan pengecekan terhadap feature yang memiliki korelasi tinggi menggunakan headmap. dan membuang dua feature yang memiliki korelasi tinggi

2. Pembuatan Model

- Setelah hal tersebut, kami melakukan seleksi feature menggunakan algoritma RFE untuk memilih 10 feature yang paling berpengaruh, antara lain:



- 10 feature tadi kami jadikan sebagai feature yang akan digunakan untuk training, data di split sebesar 25% untuk testing dan 75% untuk training, menggunakaan algoritma random forest kamipun melakukan training dan prediksi, untuk prediksi kami mendapatkan akurasi sebesar 100% dengan cross validation dan F1 pun akurasi tidak berubah tetap sebesar 100%



* Hasil dari prediksi tadi disimpan kedalam file exel dengan nama column predicted\_is\_canceled dan file bernama analyst\_report.xlsx