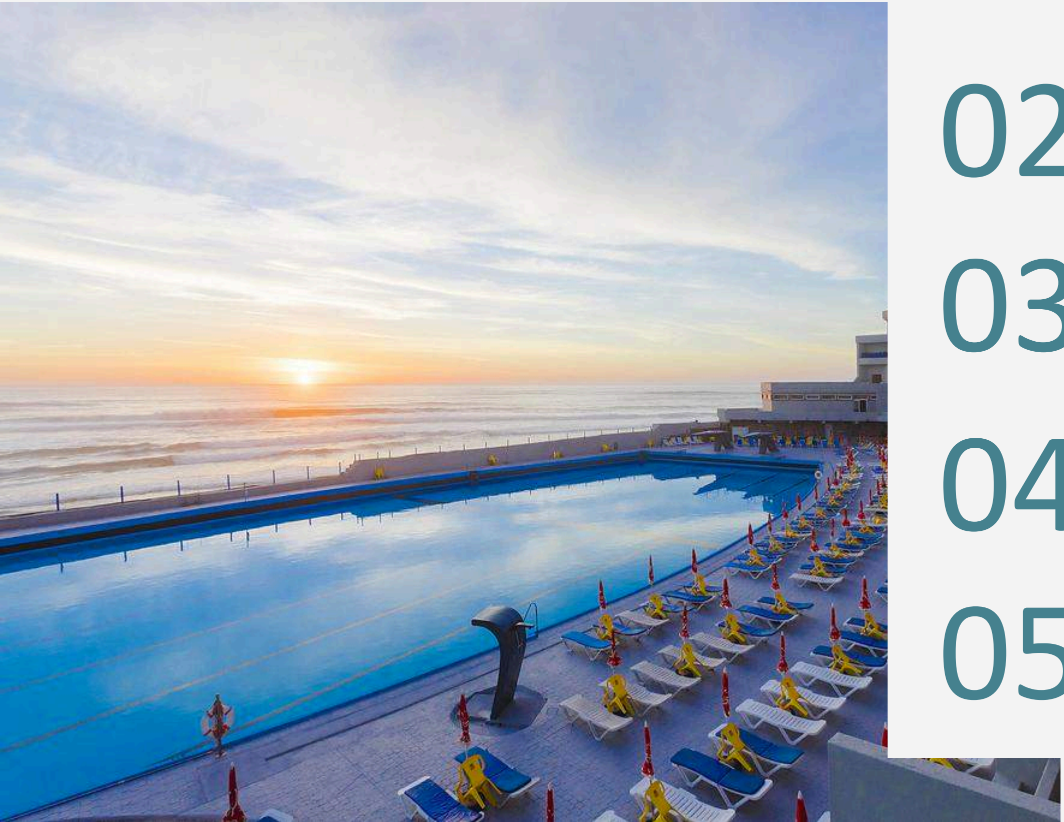
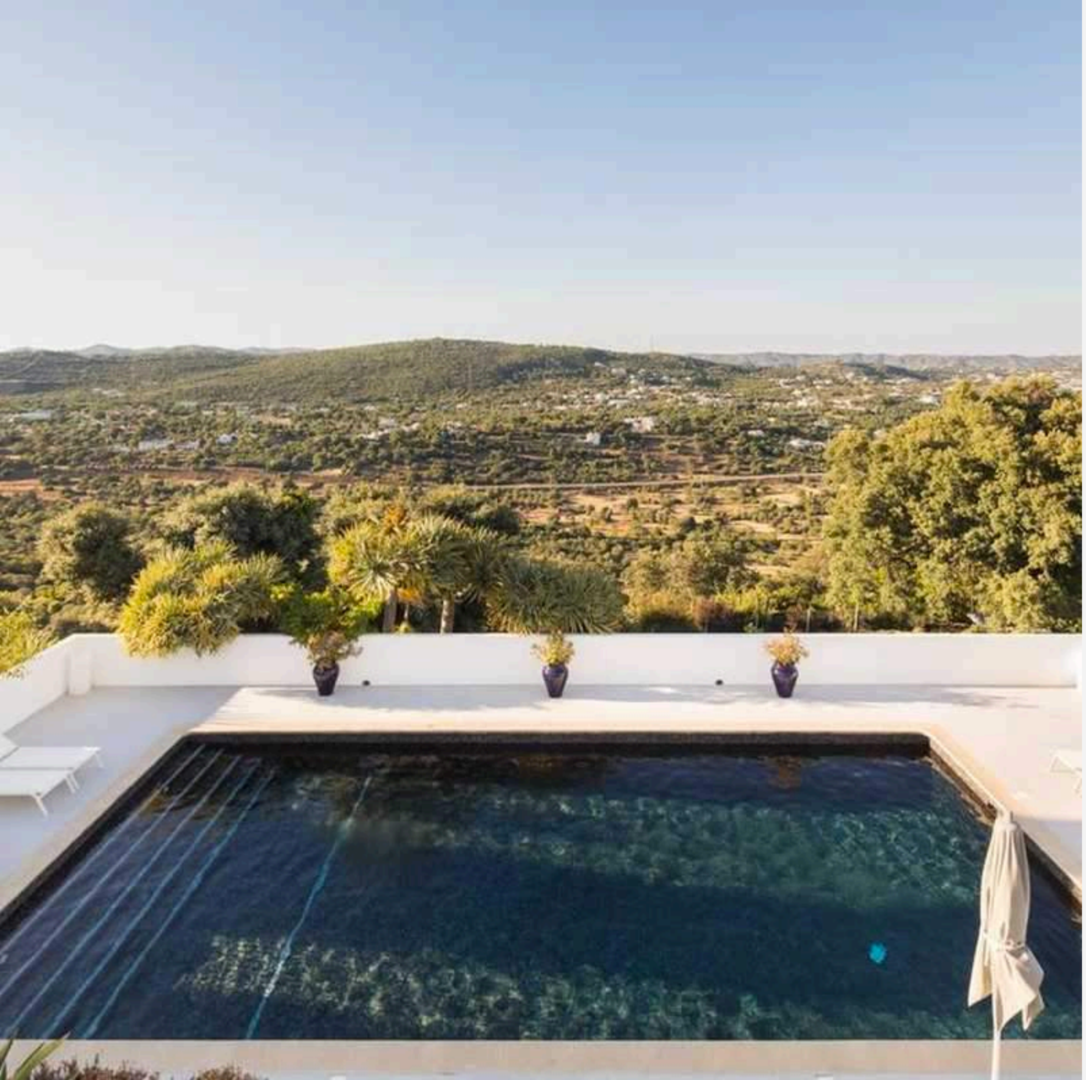
PREDICTING BOOKING CANCELLATIONS

Abélia PETELLE
February 2020

TABLE OF CONTENTS





01

CONTEXT

- About us
- Our client
- Missions

ABOUT US

DATAWORLD
CONSULTING GROUP

Dataworld is an international consulting firm, specialized in analysing data for the the travel industry since 2008

OUR CLIENT

Booking.com

Booking.com is one of the leading online accommodation booking websites

MISSIONS

PROBLEM	SOLUTION
Approach clients based on their characteristics	Cluster clients with unsupervised machine learning algorithms
Reduce the number of cancellations via their website	Predict customers cancellations with supervised machine learning algorithms

02

PROCESS

- 4 key steps

PROCESS

DATA COLLECTION

- Data : booking information about two hotels located in Portugal (2015-2017)
- Source : Kaggle

DATA CLEANING

DATA EXPLORATORY & MANIPULATION

MODELING

- Unsepervized learning: clustering
- Supervized learning: classification

03

OVERVIEW

- Data
- Data cleaning
- Data distribution

DATA

Bookings information of a city hotel and a resort hotel based in **Portugal**

From the 1st of July of **2015** to the 31st of August **2017**

119.368 x 32

	hotel	object
Target	is_canceled	int64
Period	lead_time	int64
	arrival_date_year	int64
	arrival_date_month	object
	arrival_date_week_number	int64
	arrival_date_day_of_month	int64
	stays_in_weekend_nights	int64
	stays_in_week_nights	int64
Client	adults	int64
	children	float64
	babies	int64
	meal	object
	country	object
	market_segment	object
	distribution_channel	object
	is_repeated_guest	int64
	previous_cancellations	int64
	previous_bookings_not_canceled	int64
Reservation	reserved_room_type	object
	assigned_room_type	object
	booking_changes	int64
	deposit_type	object
	agent	int64
	company	int64
	days_in_waiting_list	int64
	customer_type	object
	adr	float64
	required_car_parking_spaces	int64
	total_of_special_requests	int64
	reservation_status	object
	reservation_status_date	object

DATA CLEANING

MULTIPLE CATEGORIES FEATURES

- Countries: grouping the Top 5, “Other Europe” and “Other” countries
- Agent & Company: replacing their ID by 1 and 0 if not
- Meal: merging ‘Undefined’ and ‘SC’ (Self Catering) as both mean ‘No Meal’

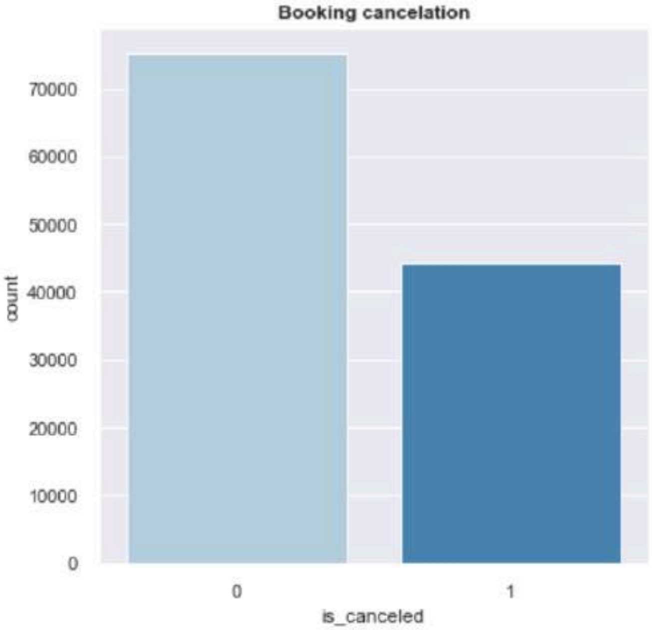
MISSING VALUES

- Children: replacing with 0
- Country: placing in “other” category
- Agent & Company: replacing with 0

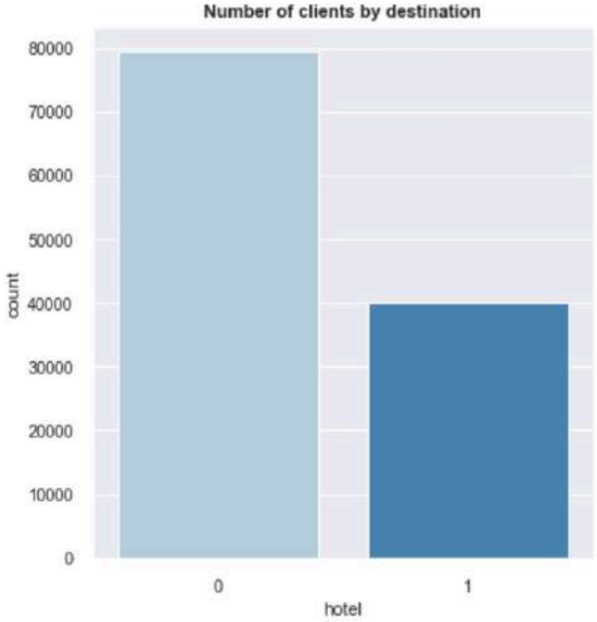
POSSIBLE SYSTEM ERRORS

- Dropping rows with booking containing more than 10 people
- Dropping bookings with 10 children & 8 babies
- Dropping negative prices (‘adr’)

DATA



→ Well balanced data

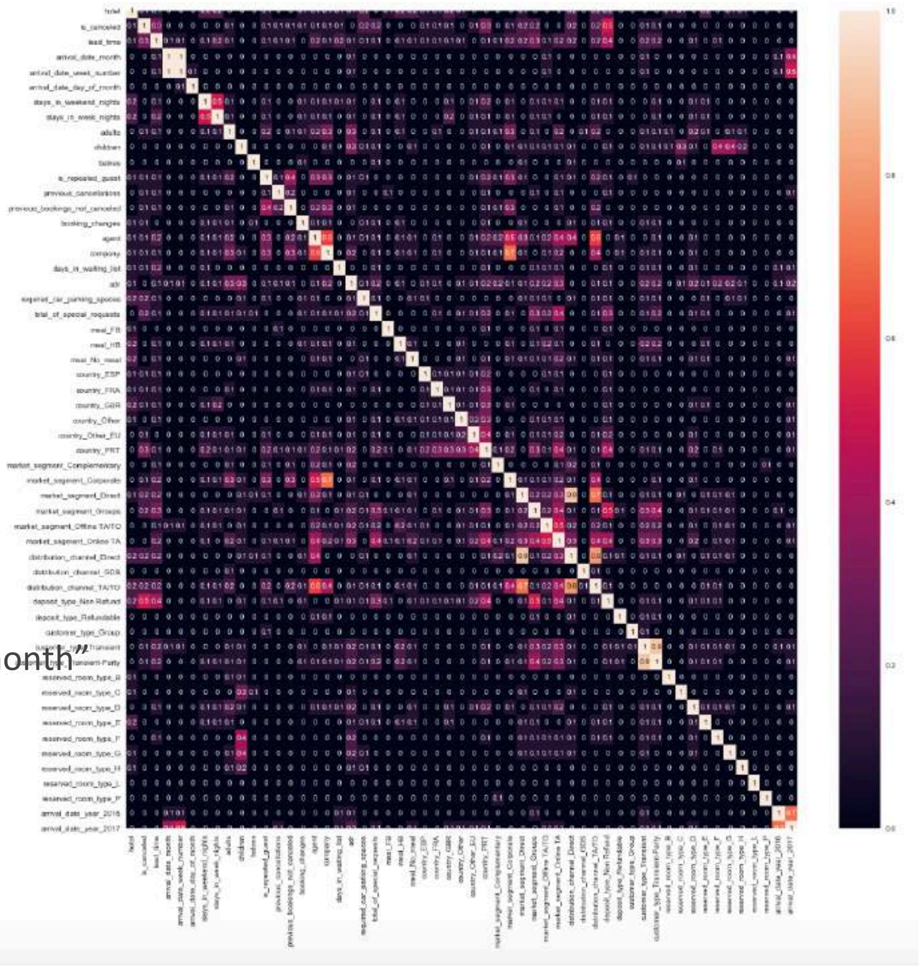


→ More clients from the resort hotel than the city hostel

CORRELATION MATRIX

6 correlated features:

- “distribution_channel_Direct” with “market_segment_Direct”
- “customer_type_Transient” with “customer_type_Transient-Party”
- “arrival_date_week_number” with “arrival_date_month”



04

MODELING

- Clustering clients
- Predicting cancellations

DATA MANIPULATION

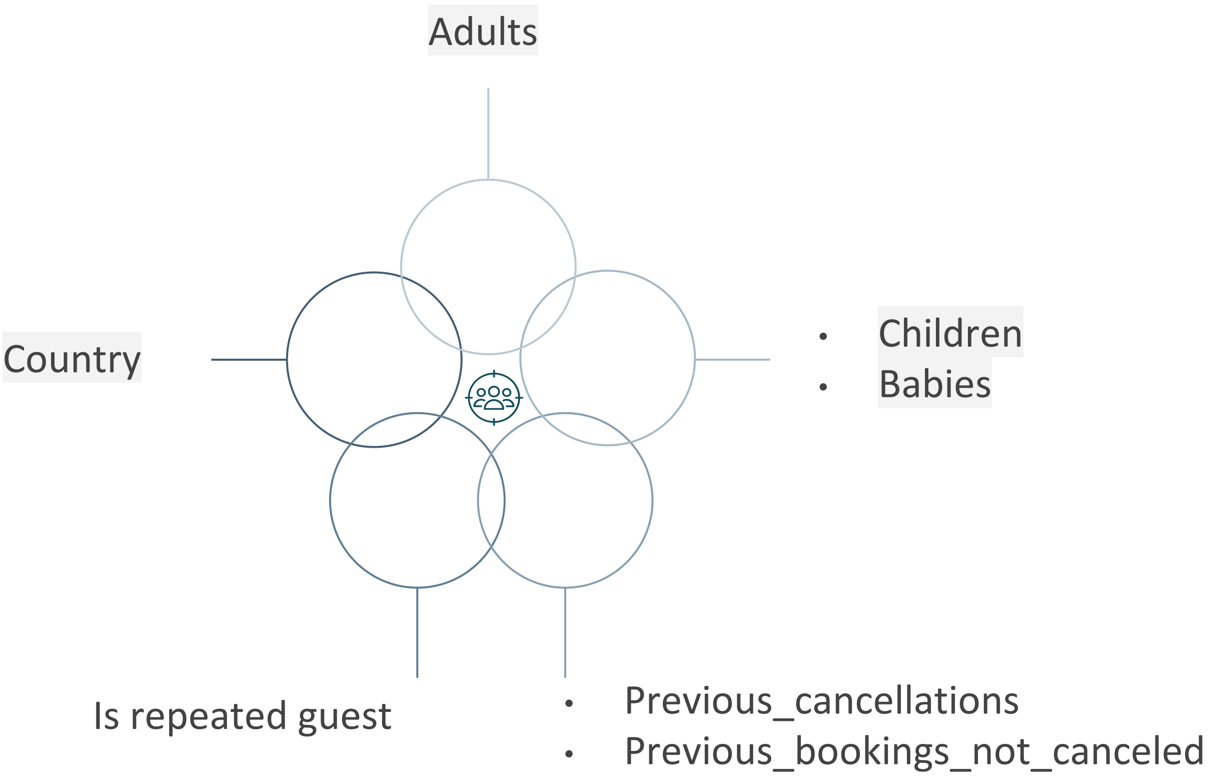
TRANSFORMING FEATURES IN NUMERICAL DATA

- **Month into interger values**
- **Creating dummies for multi-value features** : ‘hotel’, 'meal','country','market_segment','distribution_channel','deposit_type’, 'customer_type','reserved_room_type','arrival_date_year’

STANDARDIZING FEATURES HAVING DIFFERENT SCALES

- 'previous_cancellations’,’previous_bookings_not_canceled’

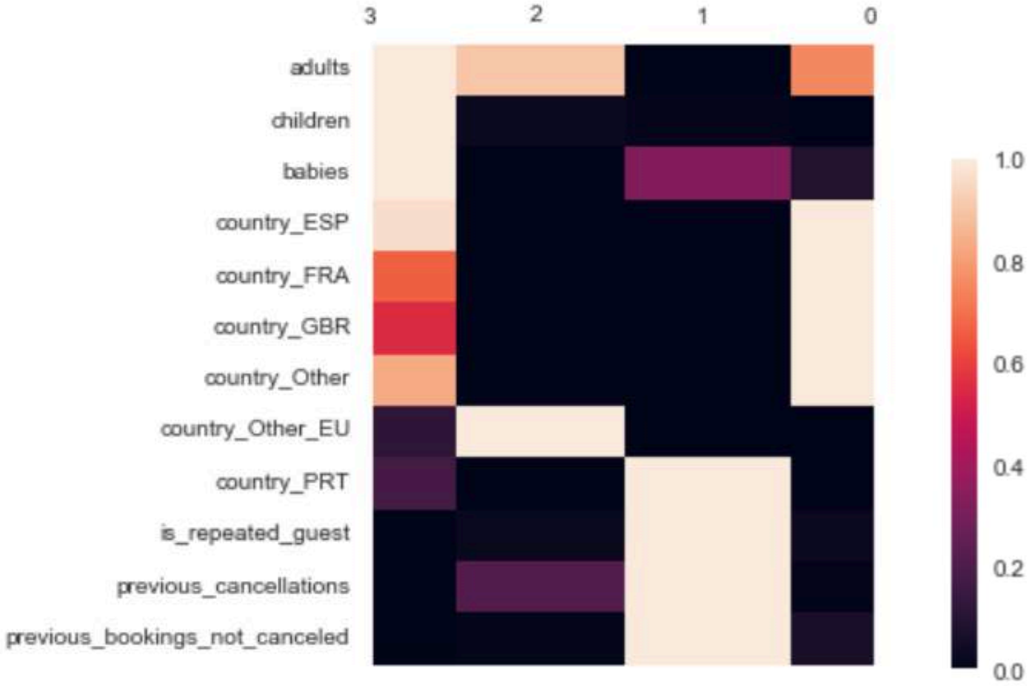
CLUSTERING CLIENTS



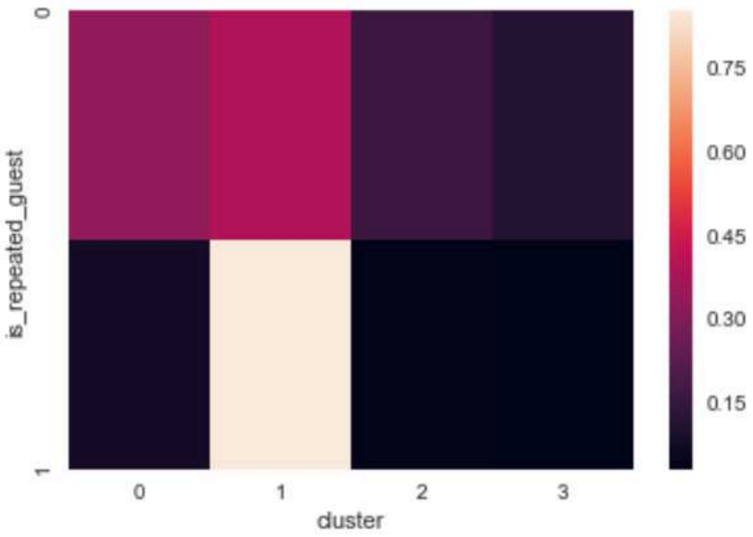
KMEANS WITH PCA

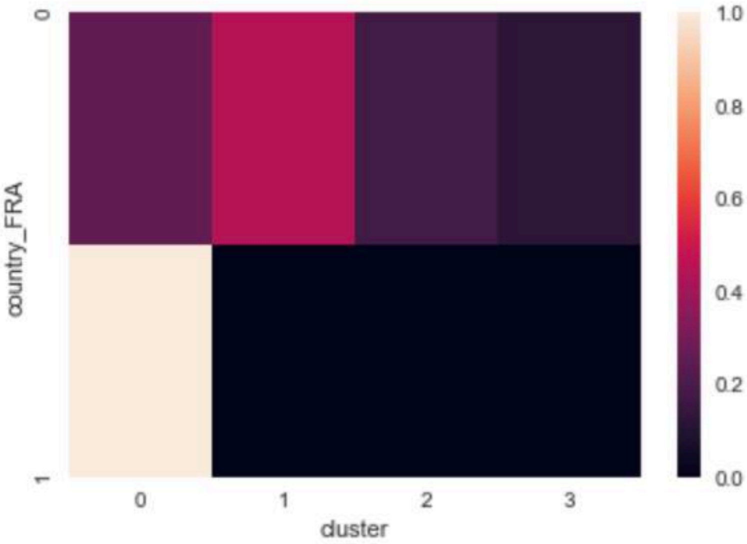
4 clusters:

- 1 : 48 568
- 3 : 38 399
- 2 : 19 029
- 0 : 13 372



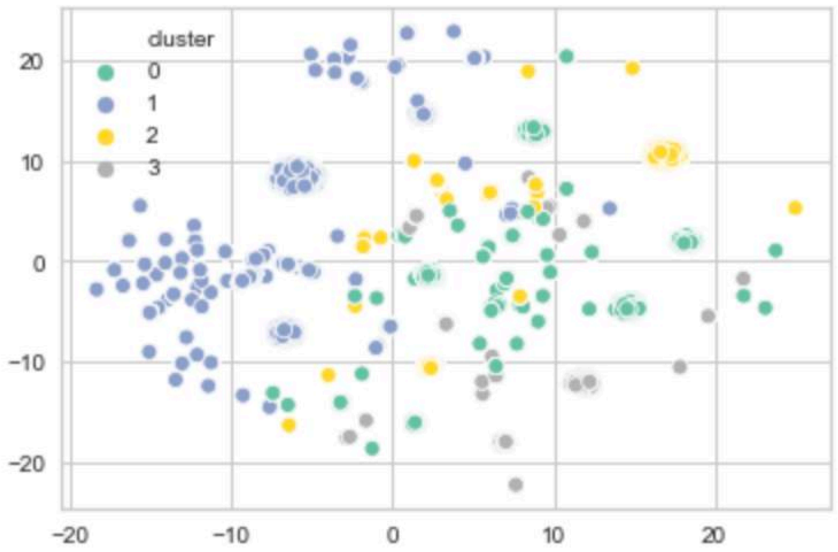
CLUSTERING WITH PCA





VISUALIZATION WITH UMAP

- **Silhouette score : 0.48**
- **Davies Bouldin score : 1.3**



→ Even if metrics are not bad, we do not see clear separations between clusters

PREDICTING BOOKING CANCELLATIONS

- Objective
- Comparison of models
- Contribution of features

OBJECTIVE

TRUE NEGATIVE

MODEL PREDICTED « NOT CANCELLED»

AND IT WAS TRUE

FALSE POSITIVE

MODEL PREDICTED « CANCELLED »

BUT THE CLIENT DIDN'T CANCEL

= RISK OF OVER - BOOKING

FALSE NEGATIVE

LACK OF RESERVATION =

MODEL PREDICTED « NOT CANCELLED»

BUT THE CLIENT CANCEL

TRUE POSITIVE

MODEL PREDICTED « CANCELLED»

AND IT WAS TRUE

The worst case

= **risk of over-booking**

The main objective

= **decrease the False Positive**

Metrics to increase

= **precision score**

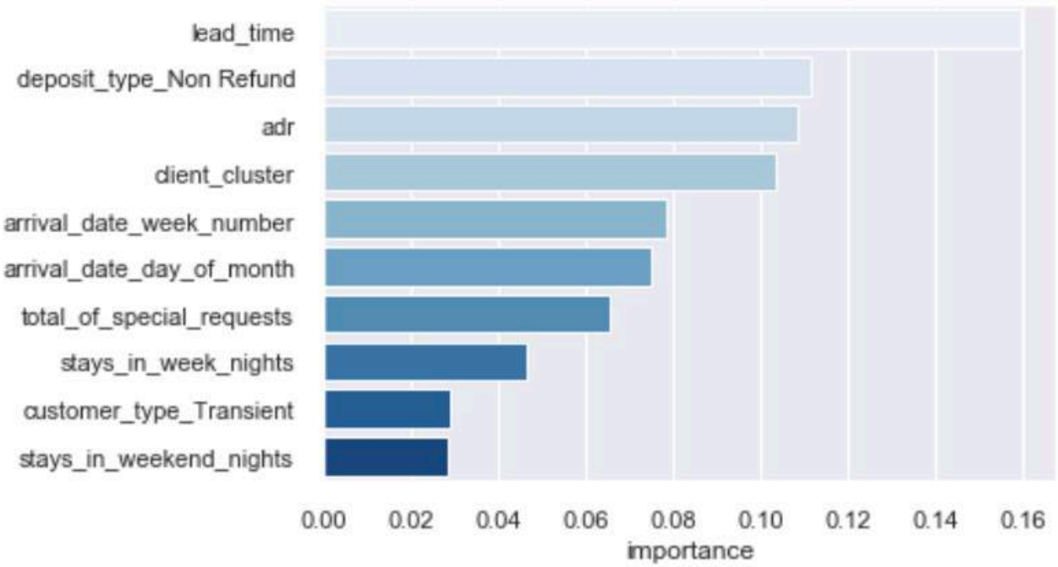
COMPARISON

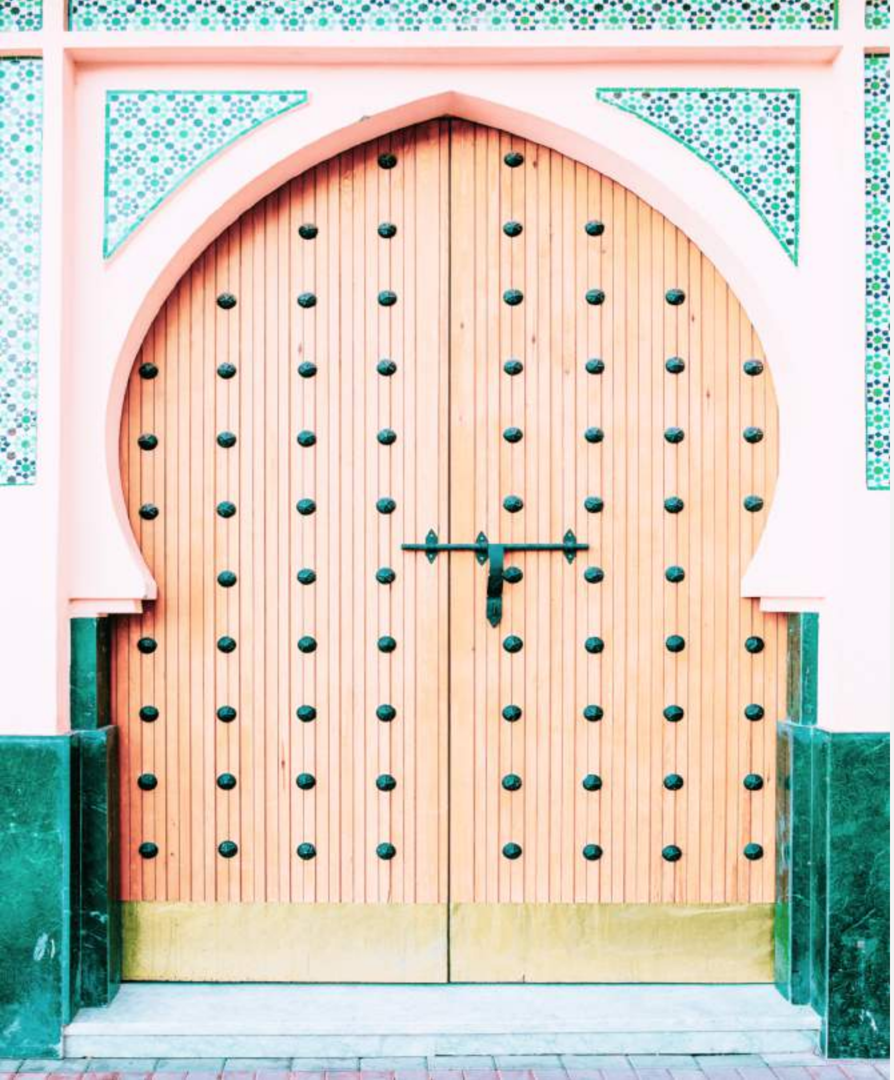
	Model	Roc Auc	Accuracy	Recall	Precision	F1 score
0	Logistic Regression	0.759	0.771	0.713	0.683	0.698
1	K Nearest Neighbors	0.812	0.833	0.734	0.798	0.765
2	Decision Tree	0.830	0.841	0.786	0.786	0.786
3	Random Forest	0.861	0.879	0.792	0.869	0.829
4	Naive Bayes	0.637	0.578	0.864	0.463	0.602
5	Catboost	0.844	0.862	0.777	0.839	0.806
6	Voting Classifier	0.863	0.880	0.798	0.866	0.831

→ Since we know that Random forest tends to overfit, we will keep the **Voting Classifier** (the ensemble of our best models) as our final model

CONTRIBUTION OF FEATURES

Features that have a **high impact** for **predicting a cancellation** are:





05

SUMMARY

RESULTS

1. With the Voting Classifier model, we are able to **predict a booking cancellation by 86%**
2. Voting Classifier works well and **guarantees us a model that will tend to be less over-fitted** than the Random Tree Classifier
3. While we weren't confident about our **clusters**, it turns out to be **the 4th most important feature** for predicting a cancelation

DIFFICULTY

Hard to see the result of our clustering and understand what the model did

IMPROVMENTS

Building a more universal model getting more data from various hotels

Building a more specialized model with better results focusing only on a specific hotel

THANKS

Do you have any questions?