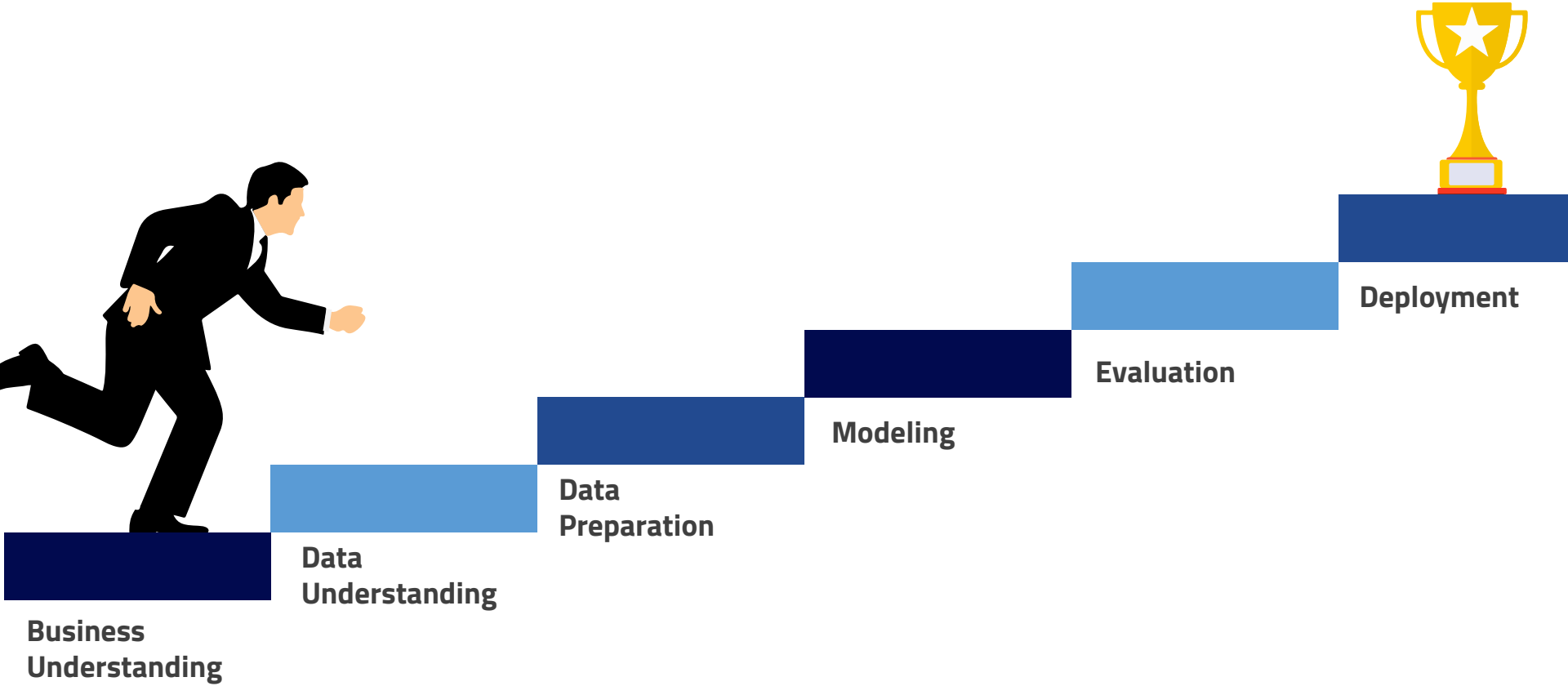
Predictive Analysis on Hotel Cancellations

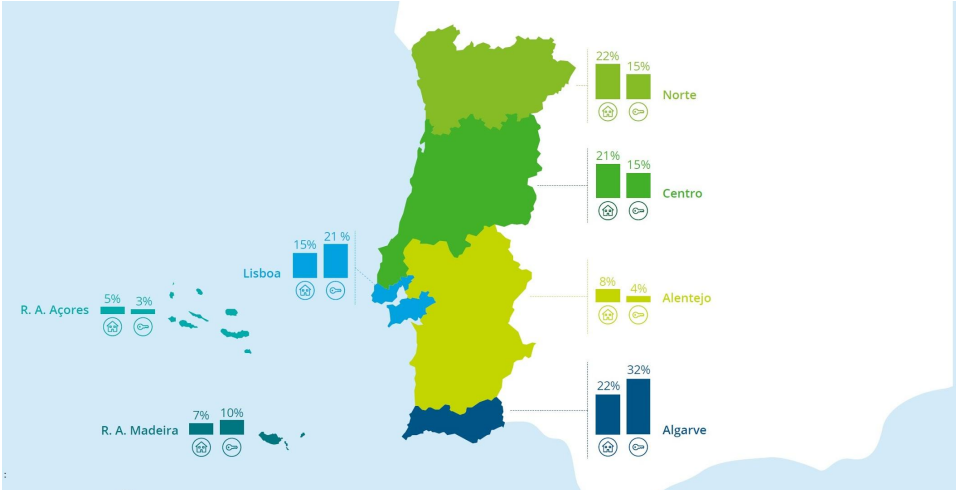
TEAM 8


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment

Business Understanding

Location: Portugal

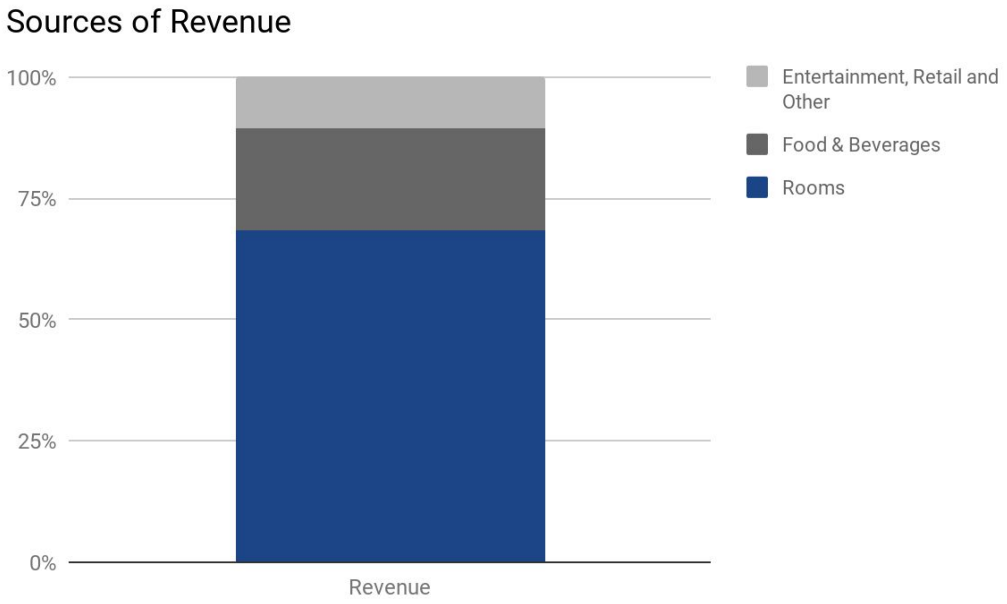
- Lisbon (City Hotel)
- Argarve (Resort Hotel)



Source: Deloitte Industry Report 2017

Business Understanding

- **Industry: Hospitality - Hotel**
- **Performance Metric: RevPAR**
  - Revenue per available room
  - Booking cancellations directly affects RevPAR



Source: Statista

Problem Translation

Business Problem

How can we decrease the room booking vacancy rate and increase profit?

Data Science Problem

At the time of booking, predict the likelihood of the reservation being cancelled at any time until the corresponding check-in date.

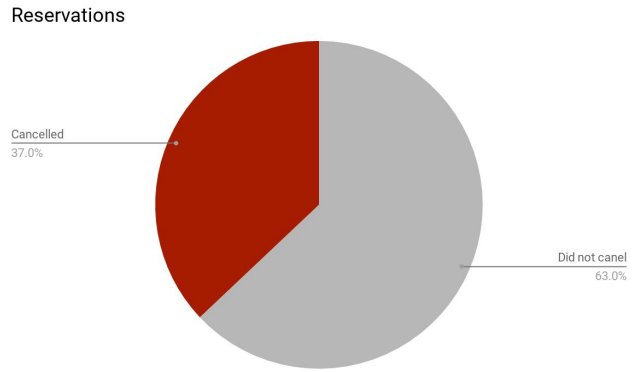
Data Overview

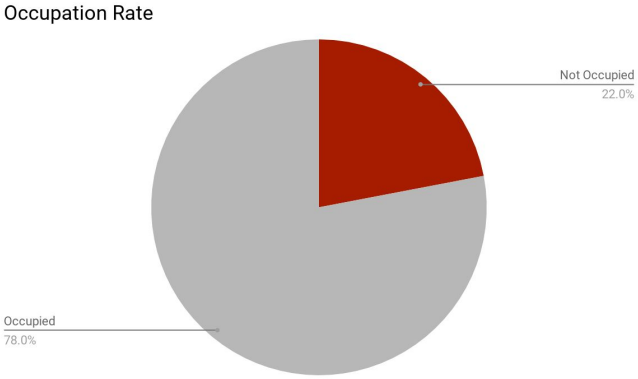
119390 reservations, **32** features

Target Variable: reservation cancellation

Hotels: Resort and City

Types of Variables: Numerical and Categorical





Source: Statista

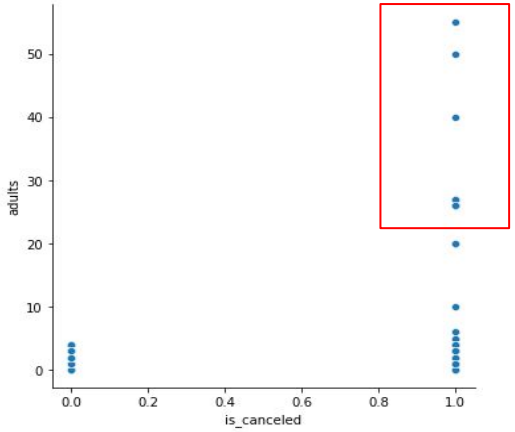
EXPLORATORY DATA ANALYSIS

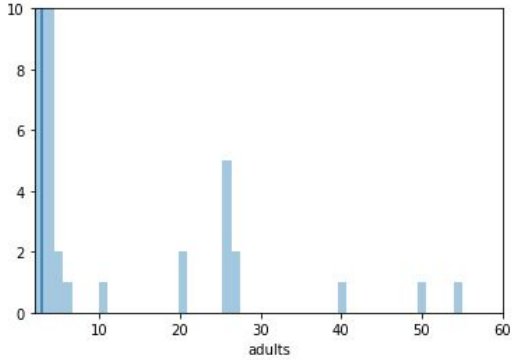
1

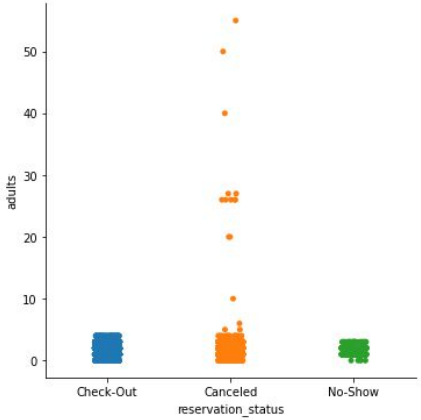
Data Understanding

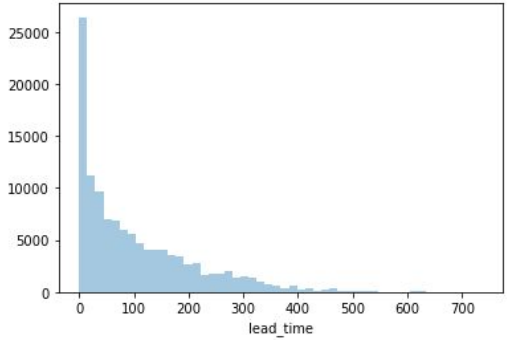
When/How long?	For Whom?	Booking information	What do they need?
Lead time	Adults	Hotel	Reserved room type
Arrival date year	Children	Average daily rate	Meal
Arrival date month	Babies	Booking changes	Total of special requests
Stays in weekend nights	Is repeated guest	Deposit type	Required parking spaces
Stays in week nights	Country	Days in waiting list	
	Customer type	Agent	
	Previous cancels	Distribution channel	

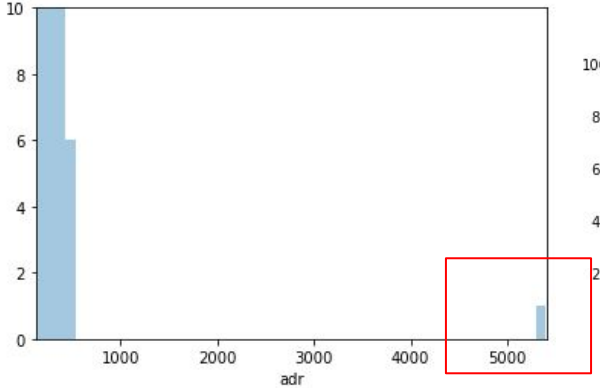
Exploration

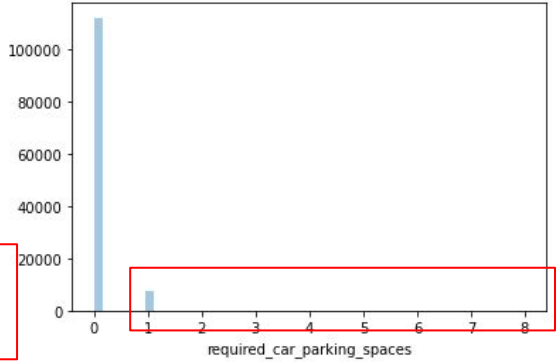










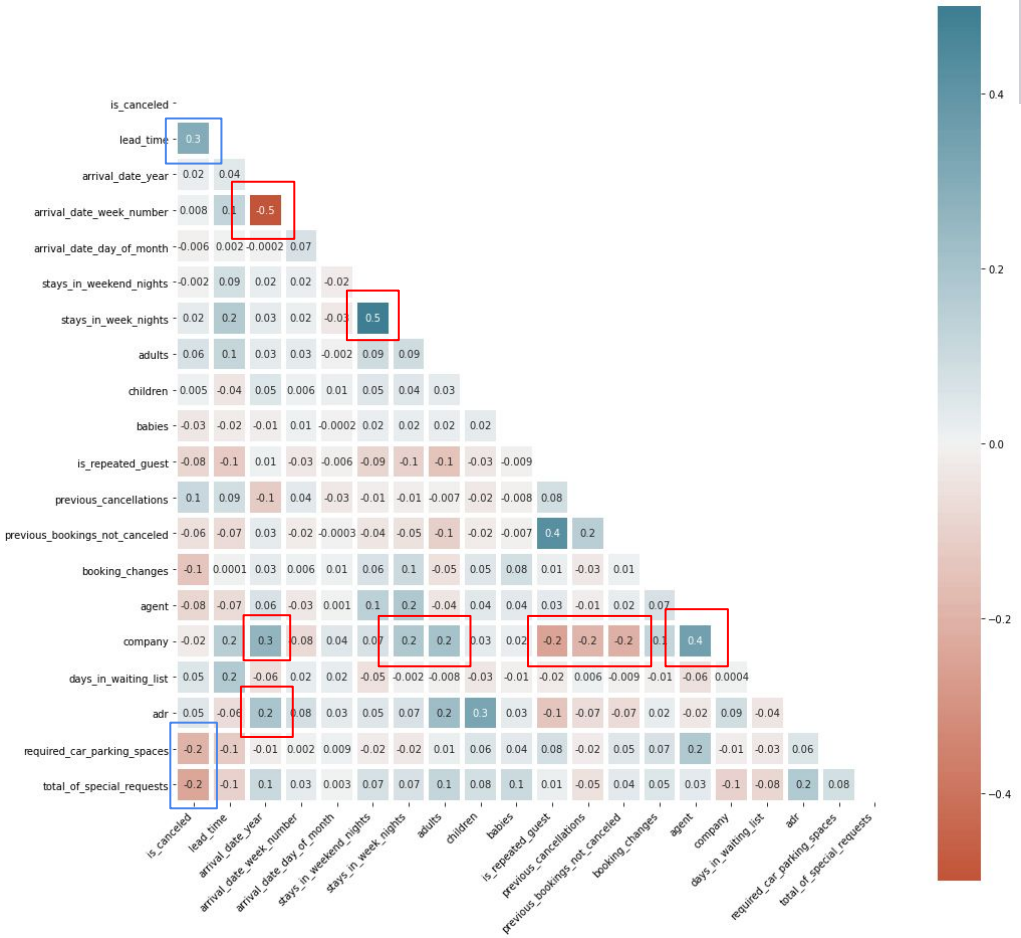


Correlation Matrix

No high correlation

- between the predictors and target variable
- Among most of the predictor variables

Price (time-series/Autocorrelation)
Average Daily Rate Mean:
2015: 87.17
2016: 98
2017: 114
Price is increasing each year
so we transformed variable to remove inflation



Information Gain

Information gain for each predictor is low (at .10s)
Still, we get rank order on which predictors are more informative

Top 3:

- Deposit type
- Average daily rate
- Lead time

	Columns	Info_gain
6	deposit_type	0.130952
21	adr	0.121398
12	lead_time	0.083678
11	agent	0.080440
8	country	0.067132
17	previous_cancellations	0.041190
3	market_segment	0.036687
23	total_of_special_requests	0.036600
22	required_car_parking_spaces	0.029941
19	booking_changes	0.019892
4	distribution_channel	0.017106
20	days_in_waiting_list	0.015232
7	customer_type	0.009796
0	hotel	0.009541
18	previous_bookings_not_canceled	0.008935
14	stays_in_week_nights	0.007235
10	is_repeated_guest	0.004130
9	adults	0.003895
5	reserved_room_type	0.002776
1	arrival_date_month	0.002489
2	meal	0.001264
16	babies	0.000660
13	stays_in_weekend_nights	0.000457
15	children	0.000417

DATA PREPARATION

2

Dropped Features

Leakage
Assigned room type: assigned once customer checks in
Reservation status: changes to either cancelled or check-out
Reservation status date: used in conjunction with reservation status

Redundancy
Information gained can be acquired from another attribute
Arrival date week number
Arrival date day of month
Arrival date year

Outliers/High leverage points
E.g. $5000 room, company trip

Features with too many levels
E.g. Countries

	Columns	Unique count
0	hotel	2
1	is_canceled	2
2	lead_time	479
3	arrival_date_year	3
4	arrival_date_month	12
5	arrival_date_week_number	53
6	arrival_date_day_of_month	31
7	stays_in_weekend_nights	17
8	stays_in_week_nights	35
9	adults	14
10	children	5
11	babies	5
12	meal	5
13	country	178
14	market_segment	8
15	distribution_channel	5
16	is_repeated_guest	2
17	previous_cancellations	15
18	previous_bookings_not_canceled	73
19	reserved_room_type	10
20	assigned_room_type	12
21	booking_changes	21
22	deposit_type	3
23	agent	333
24	company	352
25	days_in_waiting_list	128
26	customer_type	4
27	adr	8879
28	required_car_parking_spaces	5
29	total_of_special_requests	6
30	reservation_status	3
31	reservation_status_date	926

Country Transformation

Country attribute tells the customer's country of origin
There were **178** different categories

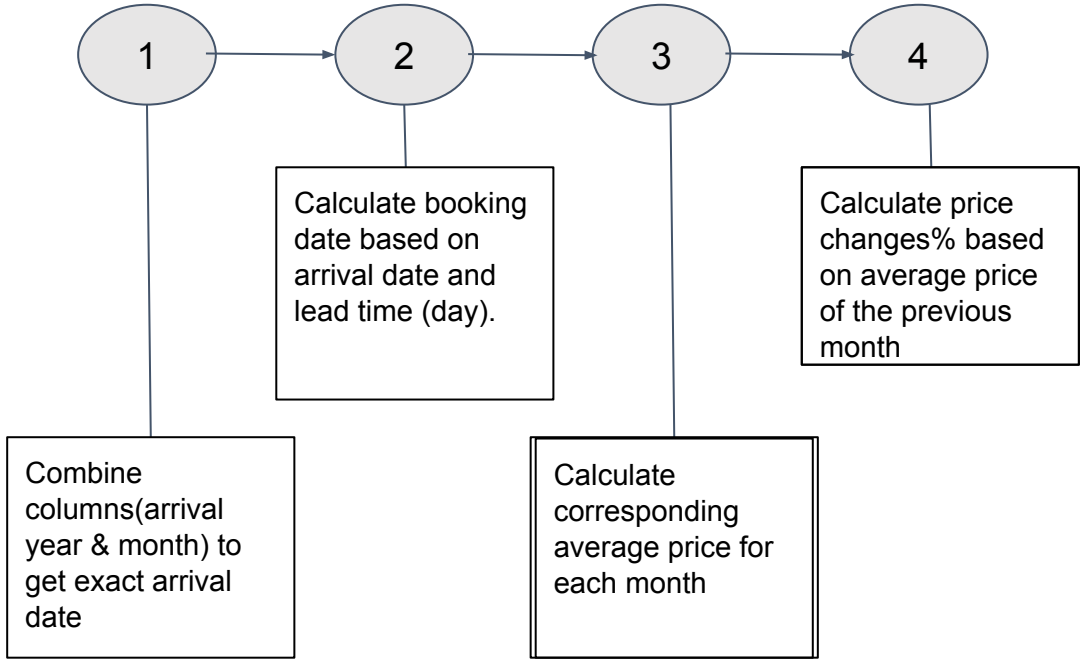
To address the large dimensions we imported a country codes index which corresponds sub-regions to each country code

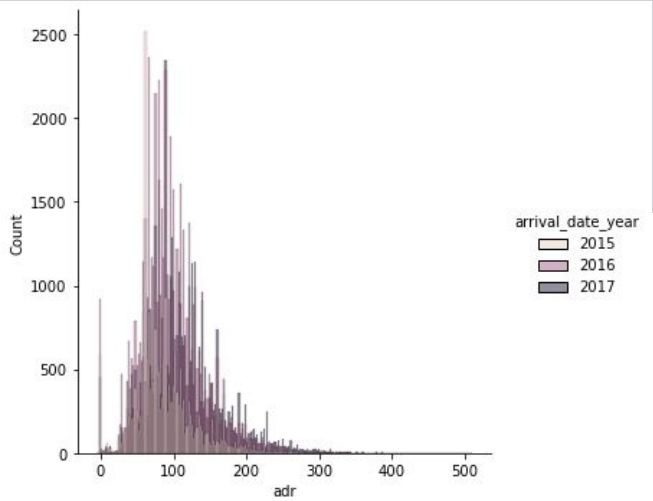
For countries that had less than 500 we remapped their value to corresponding sub-region

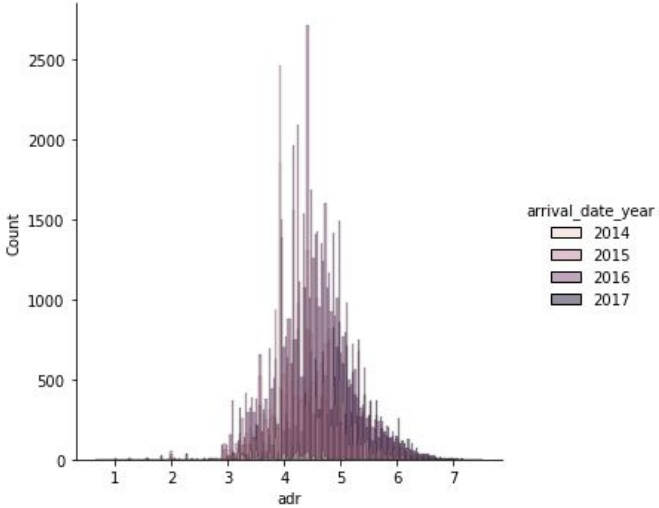
For sub-regions less than 500 occurrences we further group them into regions or "others"

We were able to get the country categories down to 28 dimensions

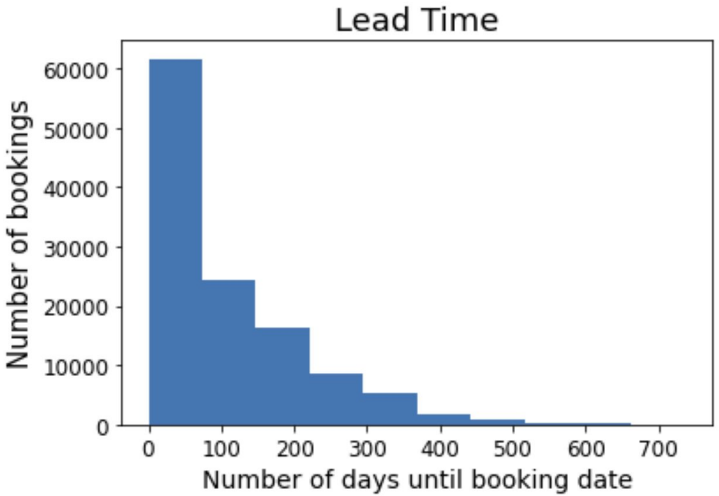
Average Daily Rate Transformation





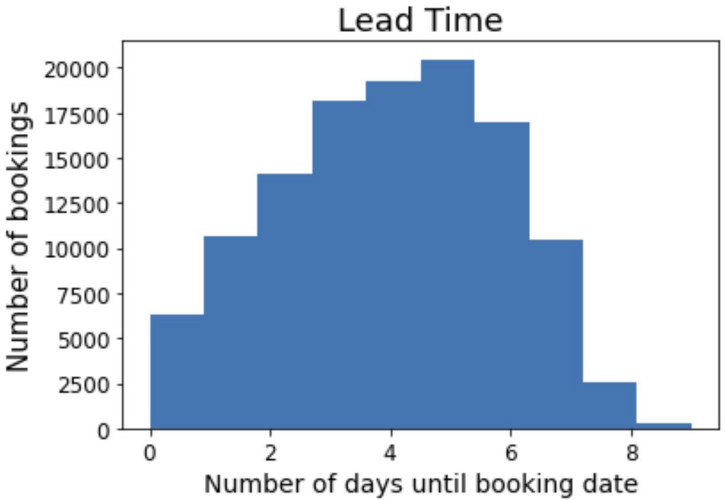


Normalizing distributions

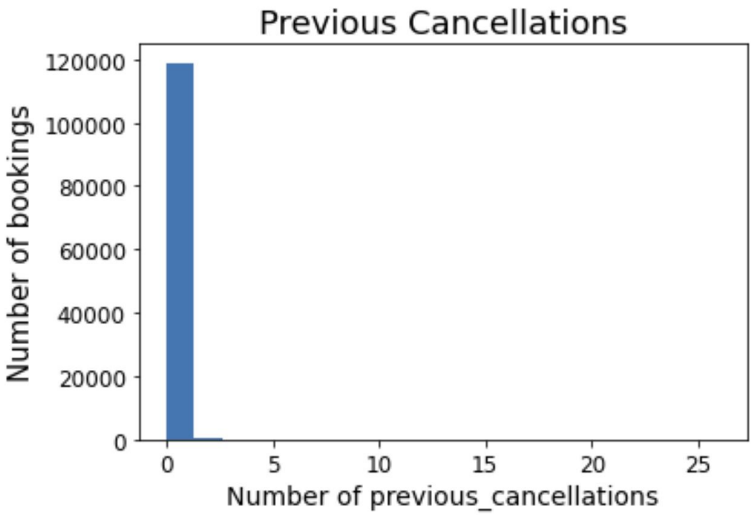


Lead Time

Stays in Weekend Nights



Regrouped Variables



- Agents
- Children
- Babies
- Adults
- Previous Cancellations
- Previous Bookings Not Canceled
- Days on Waiting List
- Booking Changes
- Required Car Parking Spaces
- Reserved Room Type

Quadratic Features for Linear Models

	children	is_canceled	count
0	0.0	0	0.624601
1	0.0	1	0.375399
2	1.0	0	0.675265
3	1.0	1	0.324735
4	2.0	0	0.574156
5	2.0	1	0.425844
6	3.0	0	0.773333
7	3.0	1	0.226667
8	10.0	1	1.000000

Children

Number of Week Nights

Data Preparation

24 Features & Eliminate 500+ Meaningless Dimensions

When/How long?

- Lead time
- Arrival date year
- Arrival date month
- Stays in weekend nights
- Stays in week nights

What do they need?

- Reserved room type
- Meal
- Total of special requests
- Required parking spaces

Previous cancels

Distribution channel

Modeling

3

Applied Techniques

Supervised Classification

One Hot Encoder

Nested Cross Validation

Standardization

Grid Search

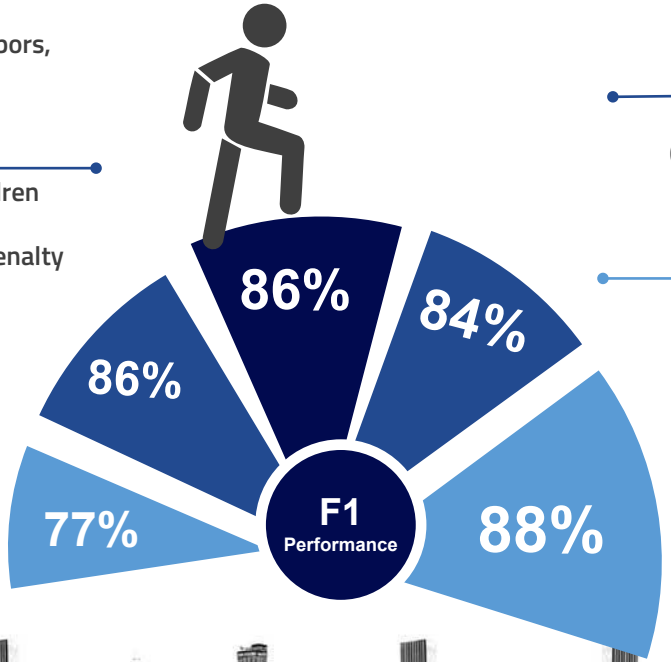
Generalization Performance

kNN

With Standardization
Optimal parameters: 21 neighbors, p:1, weights: distance, f1:86%

Logistic Regression

Used the quadratic of the children and weeknights predictors
Optimal parameters: C = 10, penalty = l2, F1: 77%



Support Vector Machine

Optimal Parameter: C: 0.01, class_weight: balanced, penalty: l2
F1: 77%

Decision Tree

Optimal Parameters:
Criterion: Entropy, Max depth = 15
Min samples leaf = 7, F1: 84%

Random Forest

Optimal parameters:
Criterion: 'gini', Max depth = 25 ,
Number of trees = 500, F1: 88%

Why these models?

01

Supervised Learning vs Unsupervised Learning
Supervised Learning is appropriate in this context because we have historical data and have a specific target variable for which we want to enhance prediction.

02

Classification vs Numeric Prediction
This is a classification task because the target variable is binary - whether the booking was canceled or not. Therefore, we can choose models that are suitable for classification prediction.

03

Pros and Cons of Linear Models
Linear models such as logistic regression and support vector machine can capture nonlinear relationships by incorporating polynomial features. We engineered two quadratic features in our model. A drawback is that with large dataset, these models are not able to capture complexities in the data as much as tree models.

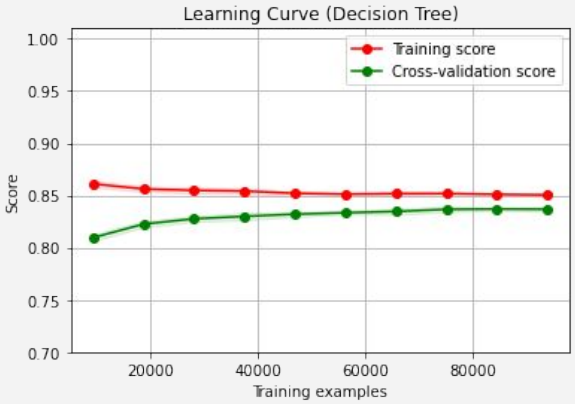
04

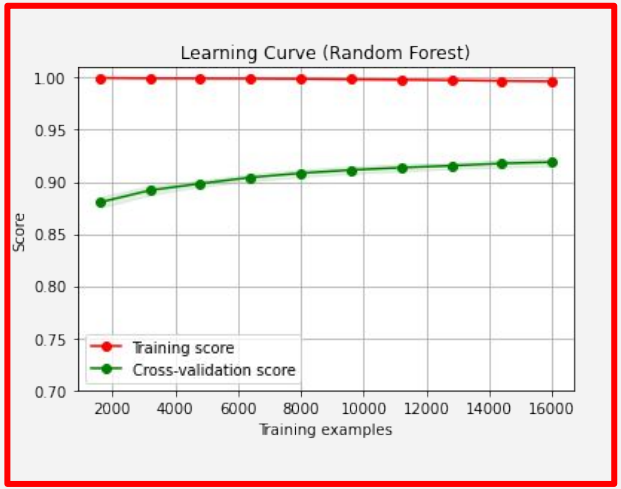
Pros and Cons of Tree Models
Tree models give flexible representation and are great at capturing high complexities when the dataset is large. Needing a good data size can be a drawback, but we have enough data points in this case.

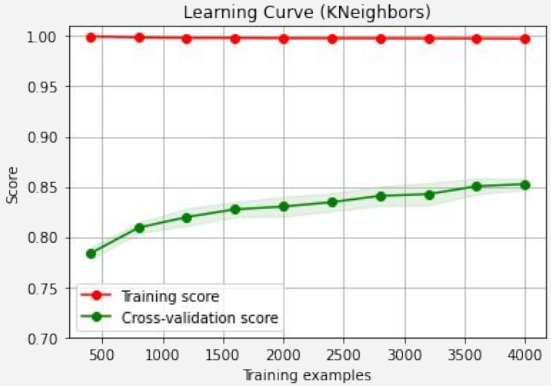
Model Evaluation

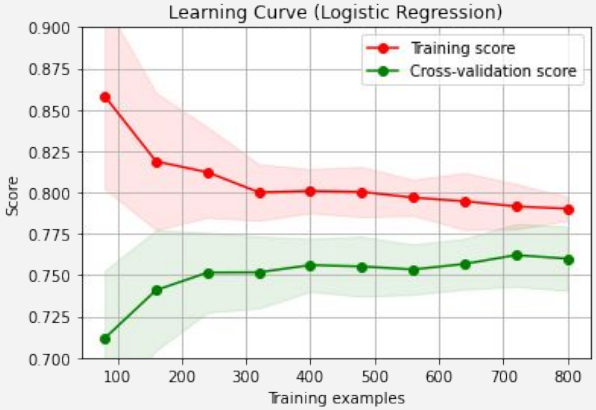
4

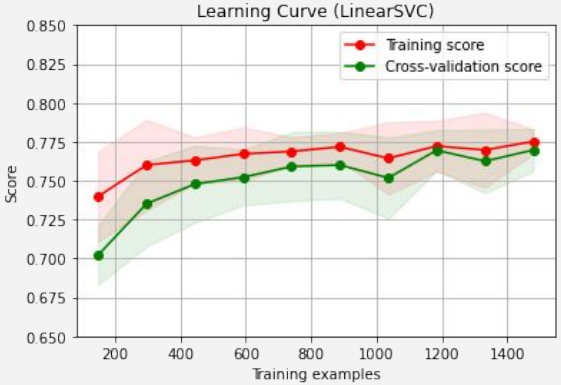
Learning Curves





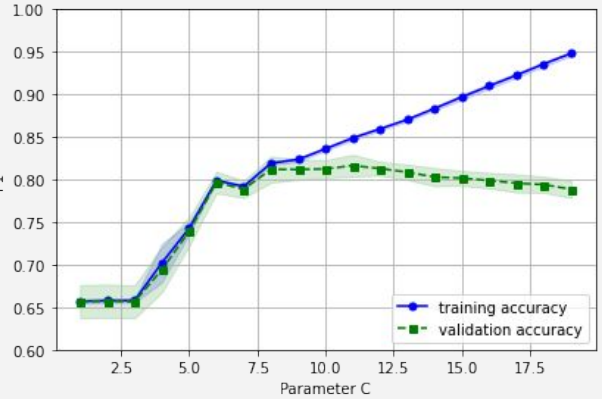




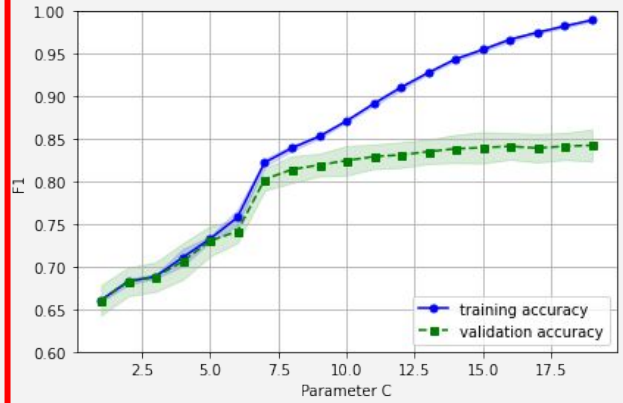


Fitting Curve

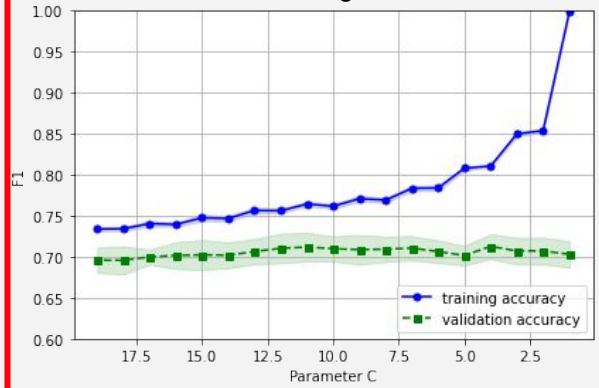
Decision Tree



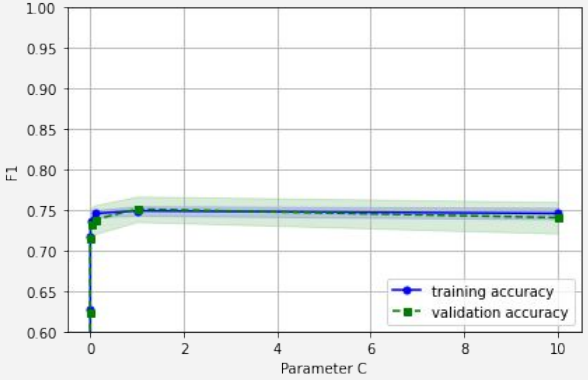
Random Forest



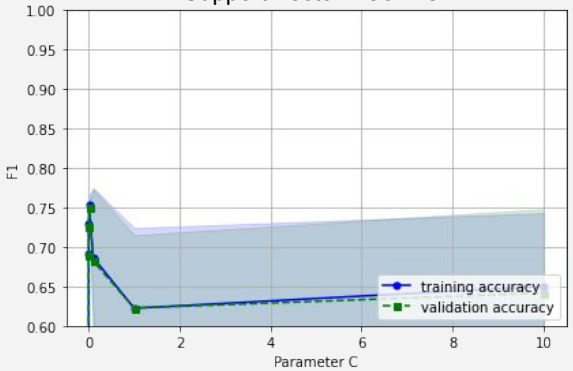
K Neighbors



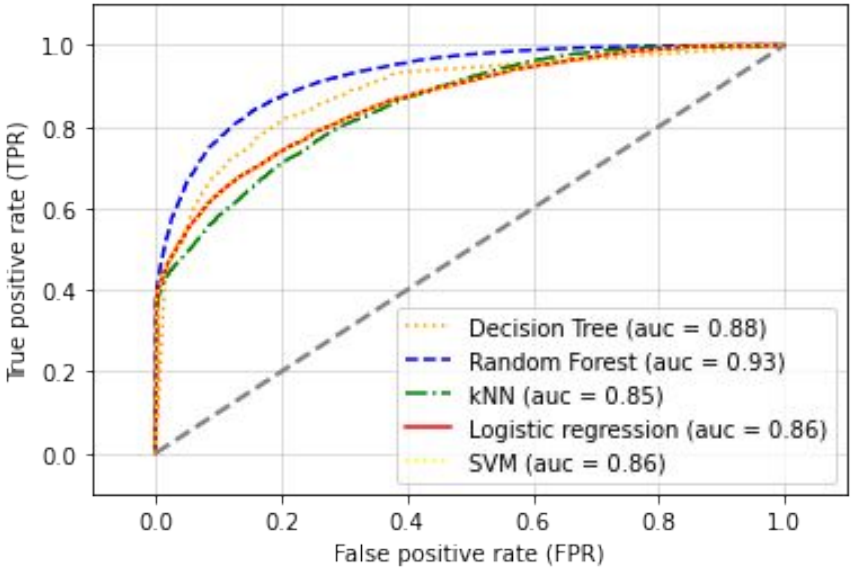
Logistic Regression



Support Vector Machine



ROC CURVE


1.0
0.8
0.6
0.4
0.2
0.0
True positive rate (TPR)
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate (FPR)
Decision Tree (auc = 0.88)
Random Forest (auc = 0.93)
kNN (auc = 0.85)
Logistic regression (auc = 0.86)
SVM (auc = 0.86)

TRUE POSITIVE:
Correctly predicted cancellations

Benefit: price of filling the room

FALSE POSITIVE:
Correctly predicted non-cancellations

Cost: refund room price, pay price difference for customer to find another room

FALSE NEGATIVE:
Predicted customer will cancel but did not

No cost/benefit

TRUE NEGATIVE:
Predicted customer will not cancel but canceled

No cost/benefit

Classification Costs and Benefits

Assumptions:

- If able to identify cancellation, hotel will fill vacancy
- Hotel is fully booked

Back to our business problem..

RevPAR	59.18€	47.85€	46.69€	35.05€	32.26€	26.83€	21.15€
	Lisboa	R. A. Madeira	Algarve	Norte	R. A. Açores	Alentejo	Centro
ARR [1]	80.65€	59.67€	63.31€	55.69€	52.52€	54.32€	46.25€
Occupancy rate	72.5%	77.5%	64.9%	60.8%	58.4%	47.1%	43.5%
Lodging capacity (no. of rooms)	30,646	14,280	44,466	21,762	4,615	6,283	21,112
Lodging revenues (000€)	654,482	241,147	666,404	269,071	51,702	58,623	155,060
Overnight stays (000#)	13,146	7,307	18,110	6,884	1,544	1,584	4,942
Guests (000#)	5,643	1,355	4,035	3,806	509	875	2,835
Average stay	2.33	5.39	4.49	1.81	3.03	1.81	1.74

[1] ARR (Average Room Rate) = room revenue / (# rooms x Occ. rate x 365).
Source: TP (turismo de Portugal) - Travel BI, INE (December 2016 data - Report data: March 2017), ATL and Turismo do Porto; Analysis: Deloitte.

© 2017. For further information, contact Deloitte Consultores, S.A.

Source: Deloitte Industry Report 2017

Deployment

5

Deployment Considerations

How

Incorporation in the online booking platform

Real-time prediction

Allow overbook

Be careful

False positives are costly:
cost of double booking, bad customer experience

Ethics

Race, Gender, Age
Nationalities

Risks

Costly overbooks: presidential suites, penthouses

Mitigation: alert system for overbook of certain type of rooms


Thank you

Sources

Country Code Dataset
https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv

Data Source: Kaggle - 'Hotel Booking Demand Dataset' by Nuno Antonio
https://www.sciencedirect.com/science/article/pii/S2352340918315191