Detecting relevant events through hotel room prices

Albert Navarro Pérez

Final Master Thesis

Master's Degree in Data Science
Universitat Oberta de Catalunya (UOC)

Professor: Jerónimo Hernández González

June 2020

Index

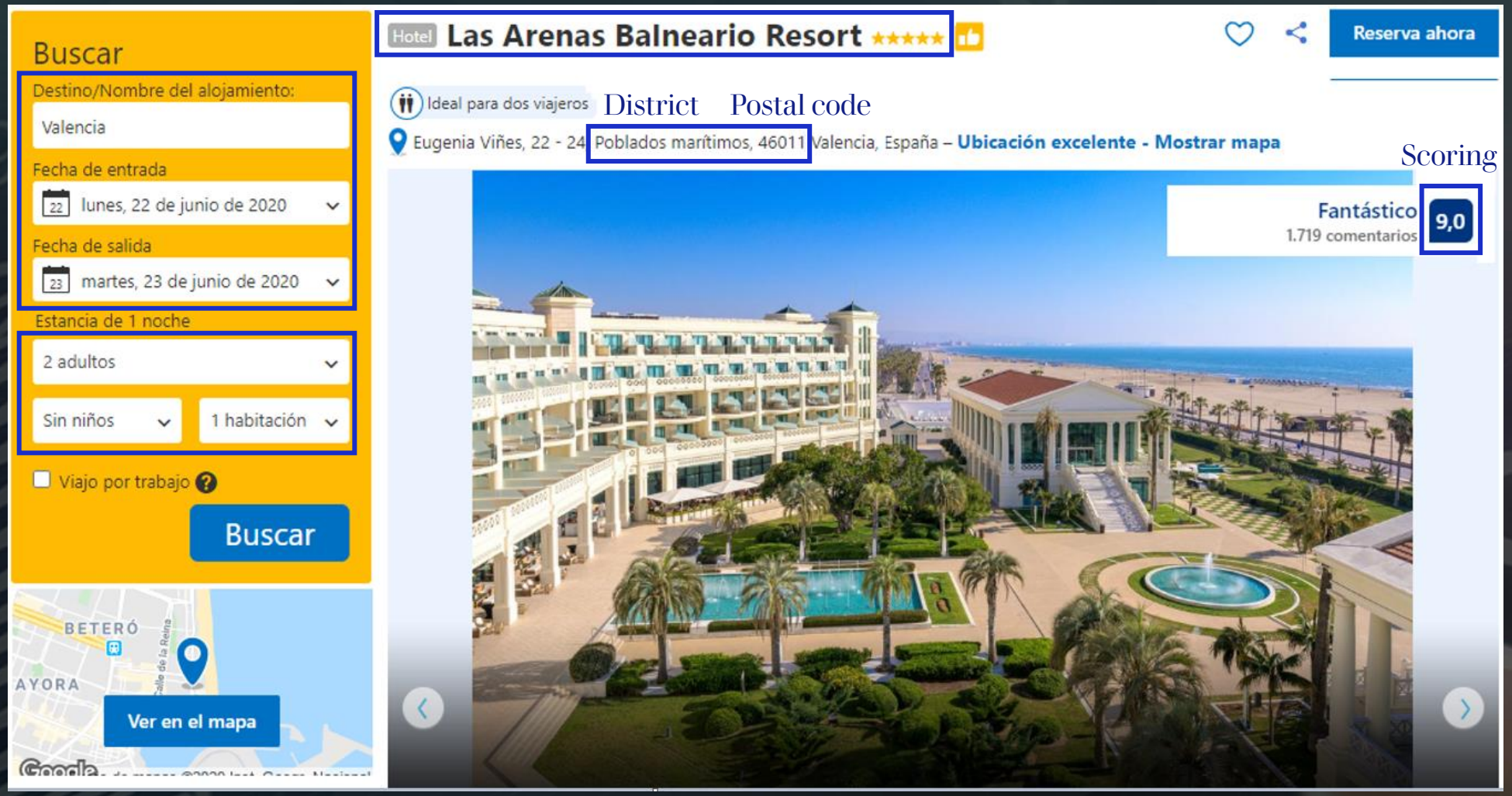
- ✓ Objectives
- ✓ Data domain
- ✓ Design and implementation
- ✓ Data analysis
- ✓ Experiments
- ✓ Results
- ✓ Conclusions
- ✓ Future lines of work

Objectives

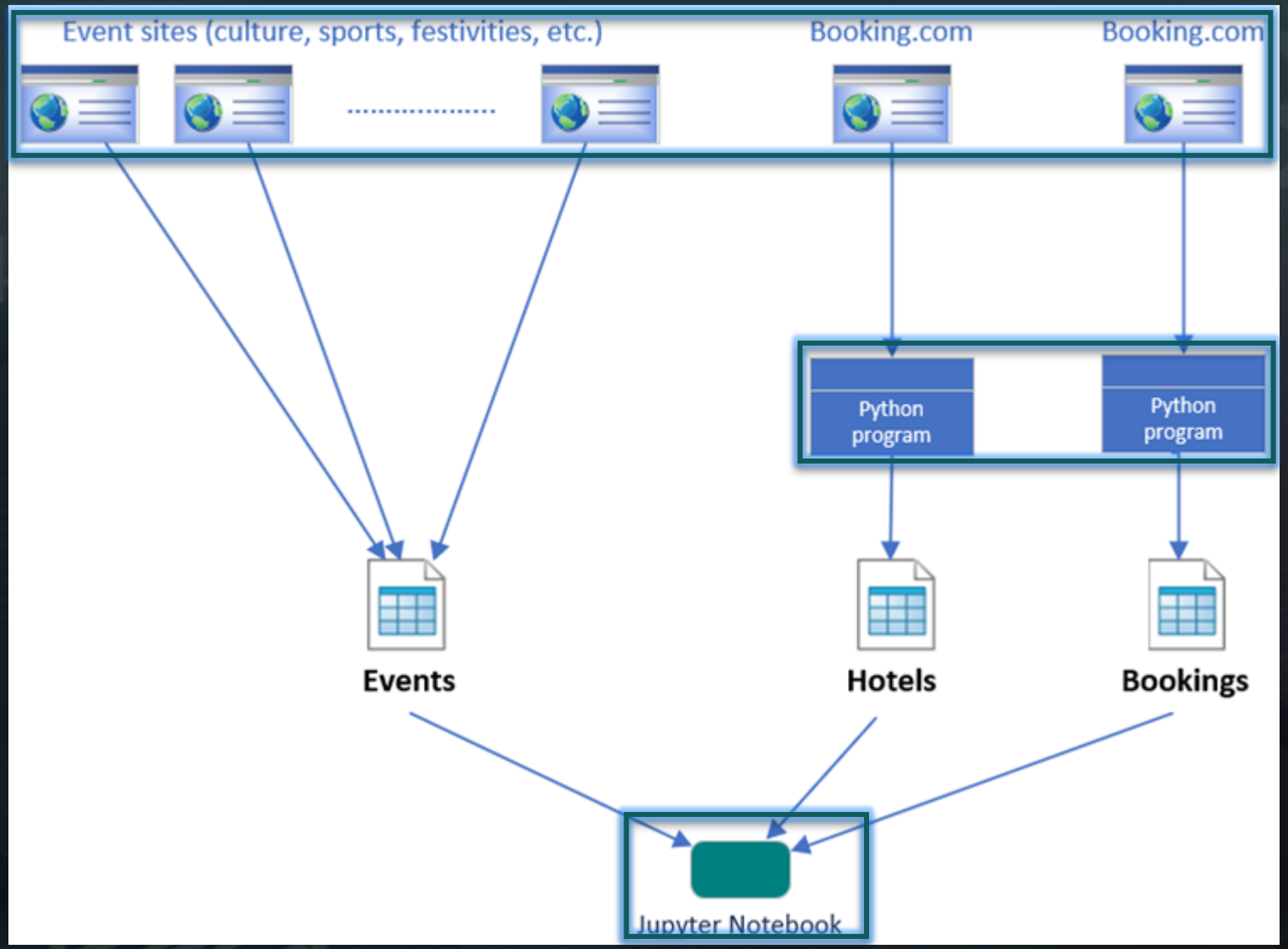
- ✓ Main objective:
  - Rise or fall of hotel room prices to predict relevant events the next day after a check-in date

- ✓ Secundary objectives:
  - Booking price over Days of advance search
  - Booking price Star rating

Data domain: Booking website

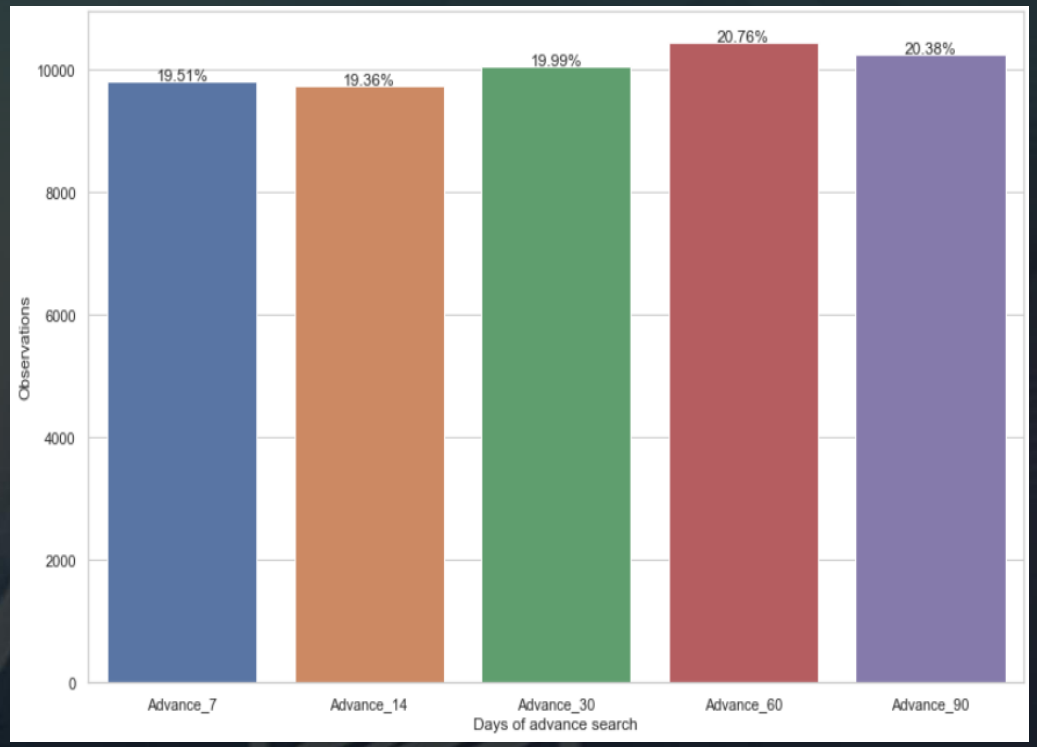


Design and implementation: overall view

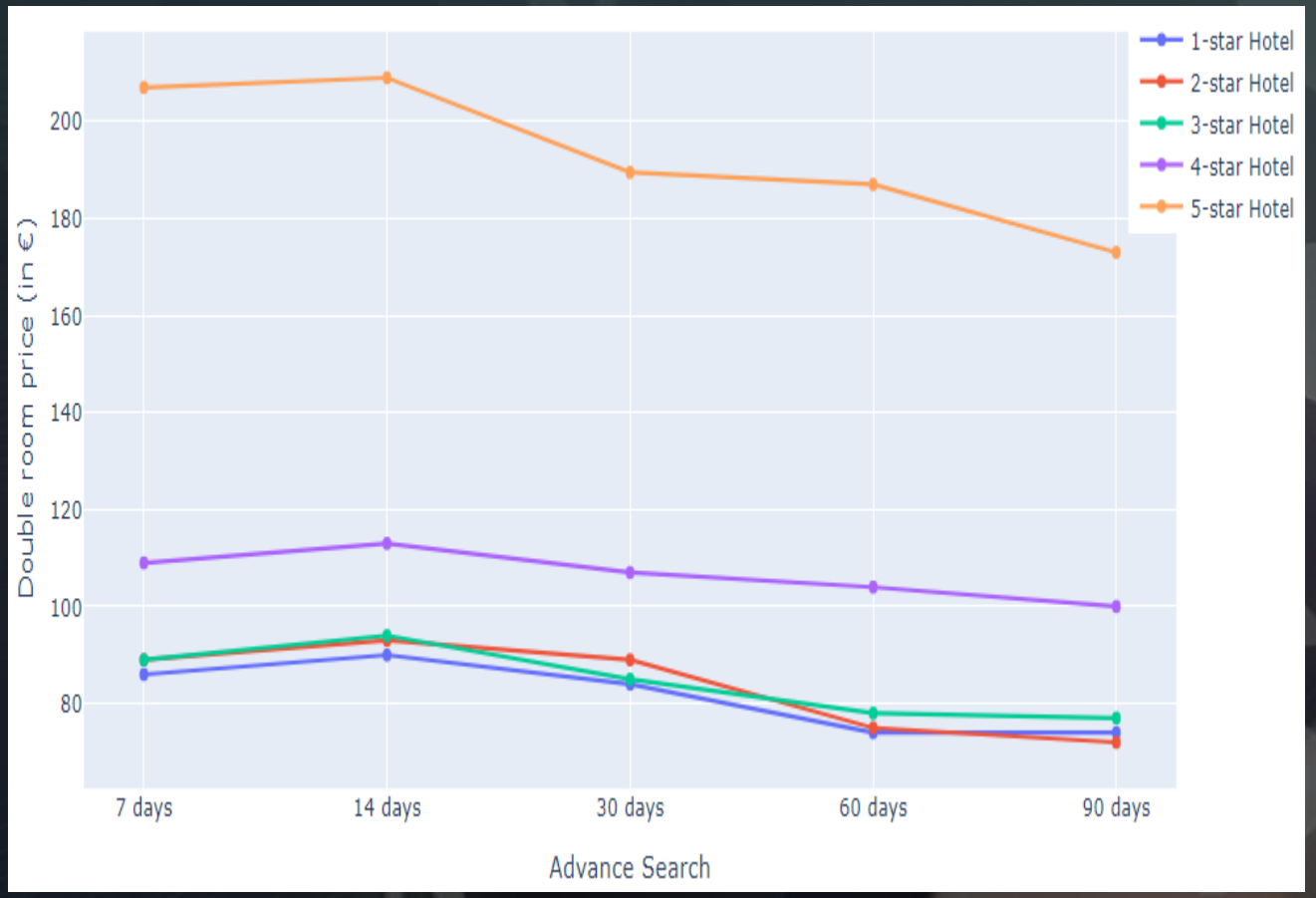


Data analysis: days of advance search

Distribution of bookings



Booking price over days of advance search

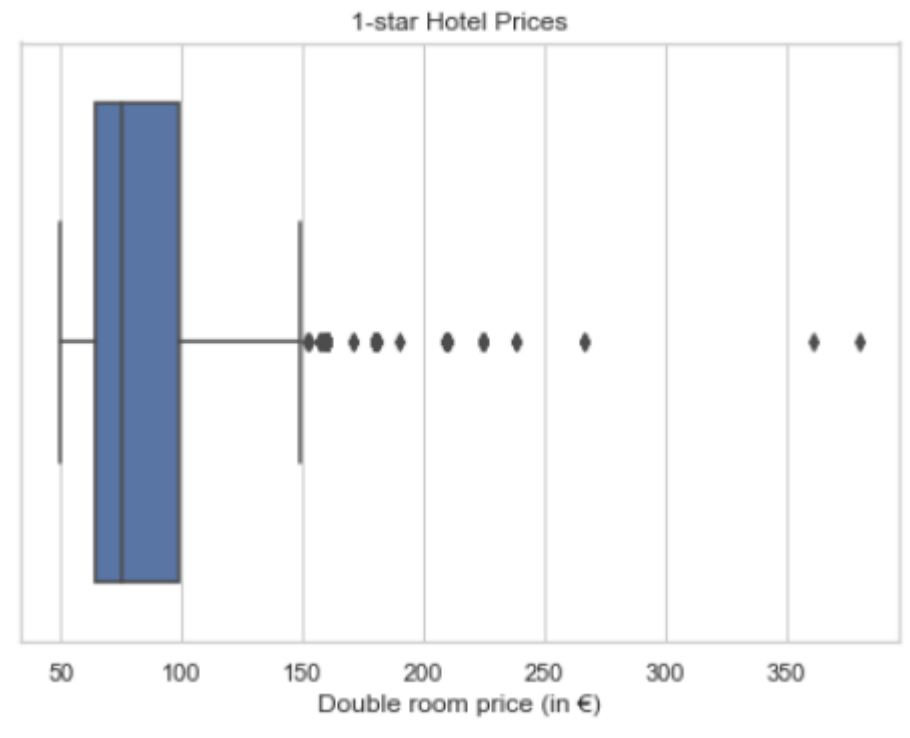


Data analysis: outliers / anomalies

Z-score: It describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units

$$z = \frac{x - \mu}{\sigma}$$

1 star Hotel prices boxplot



Bookings dataset

name	checkIn	stars	advance	netPrice	weekday	month	zscore
MD Modern Hotel - Jardines	30/11/2019	1-star	60 days	€ 380.00	Saturday	November	8.4866
MD Modern Hotel - Jardines	30/11/2019	1-star	90 days	€ 361.00	Saturday	November	7.9306
MD Modern Hotel - Jardines	16/11/2019	1-star	90 days	€ 266.00	Saturday	November	5.1506
MD Modern Hotel - Jardines	16/11/2019	1-star	60 days	€ 266.00	Saturday	November	5.1506

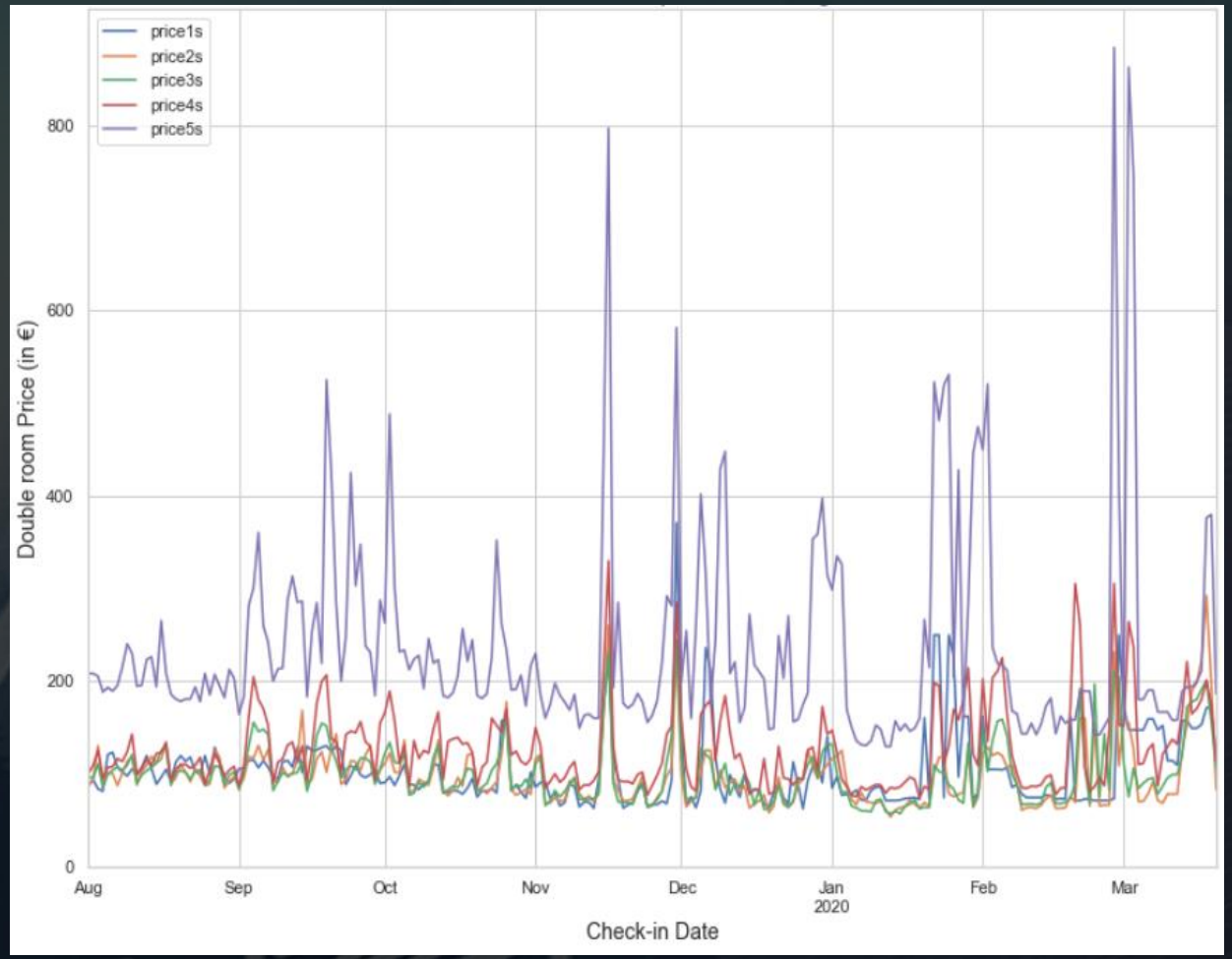
30/11/19: On 01/12/19 took place **Maratón Valencia**
16/11/19: On 17/11/19 took place **Gran Premi MotoGP** and **Dos Ruedas + VLC Bike** fair

name	checkIn	stars	advance	netPrice	weekday	month	zscore
The Westin Valencia	03/09/2019	5-star	14 days	€ 5,559.00	Tuesday	September	11.2781
The Westin Valencia	04/09/2019	5-star	14 days	€ 5,559.00	Wednesday	September	11.2781
The Westin Valencia	04/09/2019	5-star	7 days	€ 5,559.00	Wednesday	September	11.2781
The Westin Valencia	05/09/2019	5-star	14 days	€ 5,559.00	Thursday	September	11.2781

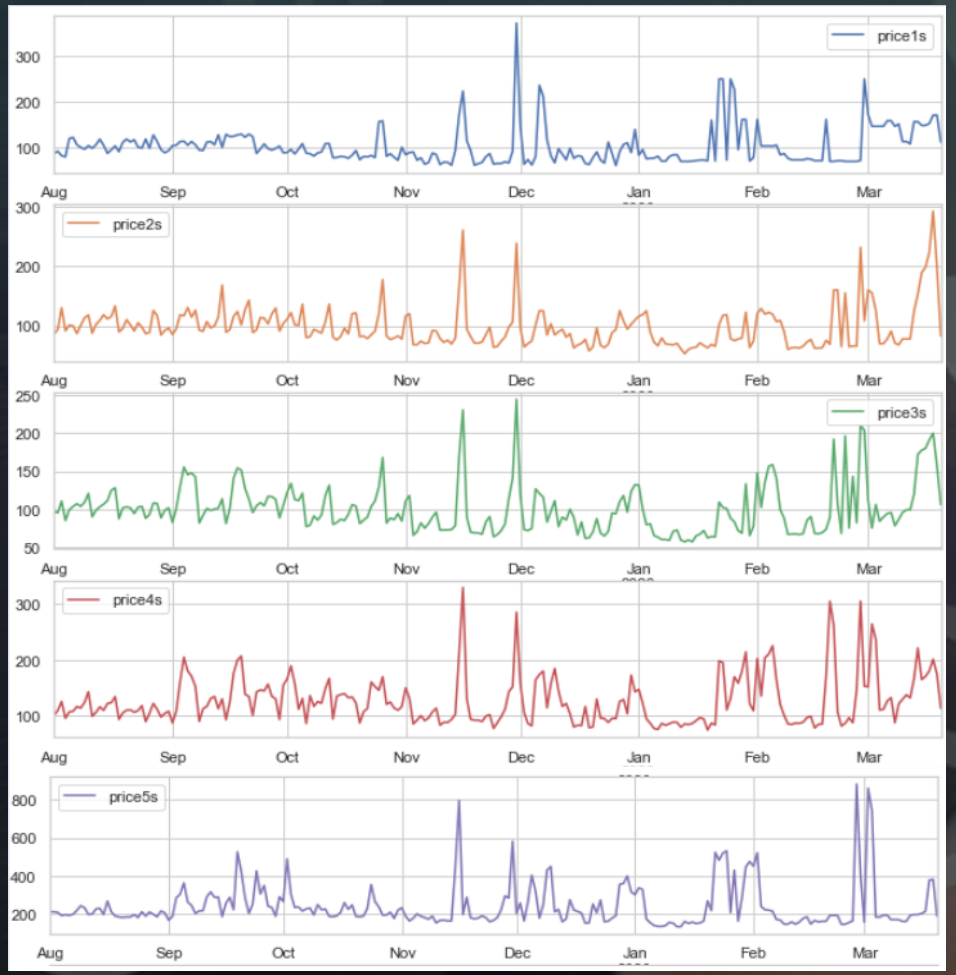
The Westin Valencia hotel has several kind of rooms.
This one refers to the most expensive one, a 80m2 **suite executive room**.

Data analysis: star rating

Booking price over Star rating



Overlapped by star



Separated by star

Data analysis: selection & manipulation

Events dataset

date	event	place	city	category
2019-08-10	Trofeo Naranja: Valencia-Inter Milan (21:30h)	Estadio de Mestalla	Valencia	Sports
2019-08-15	Día de la Asunción de la Virgen	Valencia	Valencia	Festivity
2019-09-04	EFIC Congress-Pain in Europe XI	Feria Valencia	Valencia	Congress
2019-09-05	EFIC Congress-Pain in Europe XI	Feria Valencia	Valencia	Congress
2019-09-05	IX Volta Provincia de València	Valencia	Valencia	Sports

Bookings dataset

name	propertyId	checkIn	stars	score	advance	netPrice	zipcode	district
El Globo	24498	2019-08-01	1-star	7,8	7	180.0	46011.0	Poblados marítimos
El Globo	24498	2019-08-04	1-star	7,8	7	160.0	46011.0	Poblados marítimos
El Globo	24498	2019-08-05	1-star	7,8	7	160.0	46011.0	Poblados marítimos
El Globo	24498	2019-08-06	1-star	7,8	7	160.0	46011.0	Poblados marítimos
El Globo	24498	2019-08-07	1-star	7,8	7	128.0	46011.0	Poblados marítimos

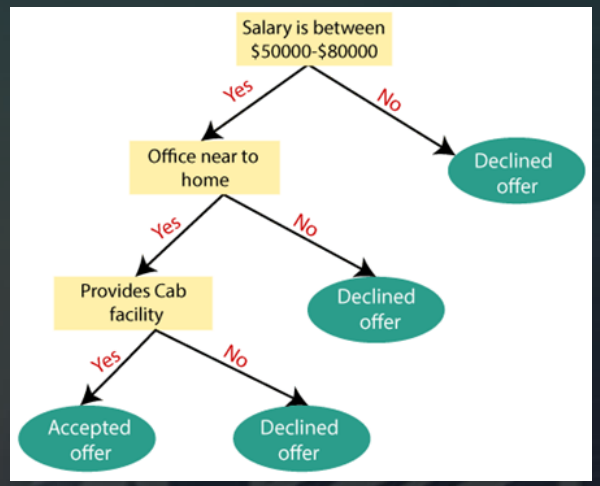
Predictors | Class label

checkIn	price1s	price2s	price3s	price4s	price5s	eventNextDay
2019-08-01	88.0	86.0	97.0	100.0	208.0	0
2019-08-02	93.0	94.0	95.0	109.0	208.0	0
2019-08-03	83.5	130.0	110.5	125.0	205.0	0
2019-08-04	81.0	91.5	85.0	95.0	188.0	0
2019-08-05	120.5	101.0	99.0	106.5	193.0	0
...	...	...	...	...	...	...
2020-03-16	149.0	198.0	180.0	170.0	199.0	1
2020-03-17	153.0	223.0	191.0	180.0	211.0	1
2020-03-18	171.0	292.0	200.0	201.0	376.0	1
2020-03-19	171.0	198.0	157.0	174.0	379.5	1
2020-03-20	113.0	83.0	106.0	113.0	186.0	0

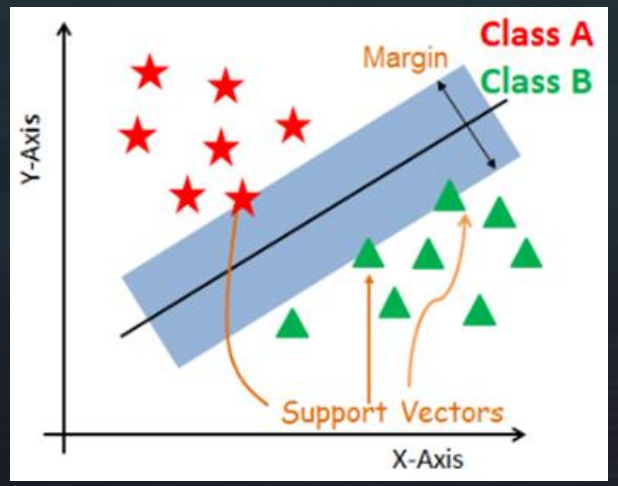
Final dataset

Experiments: binary classification models

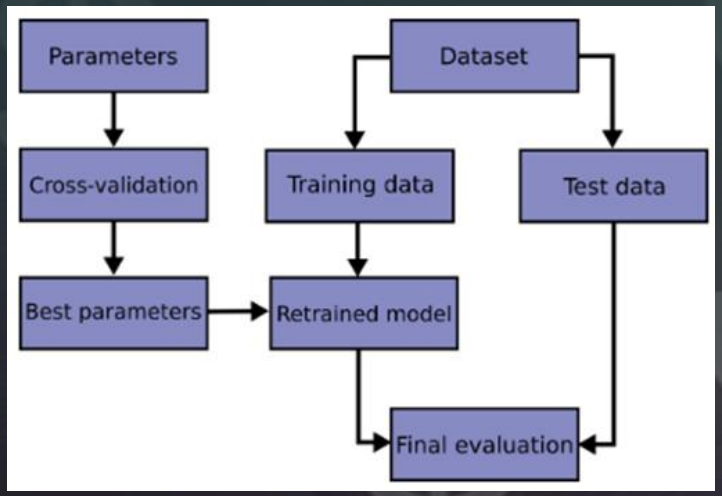
Decision Tree



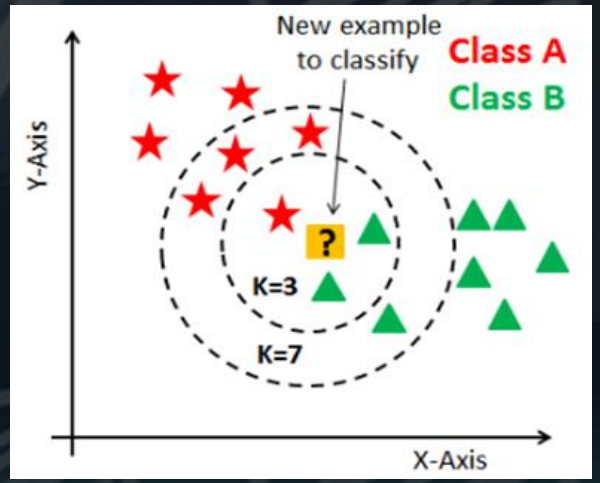
Support Vector Machine



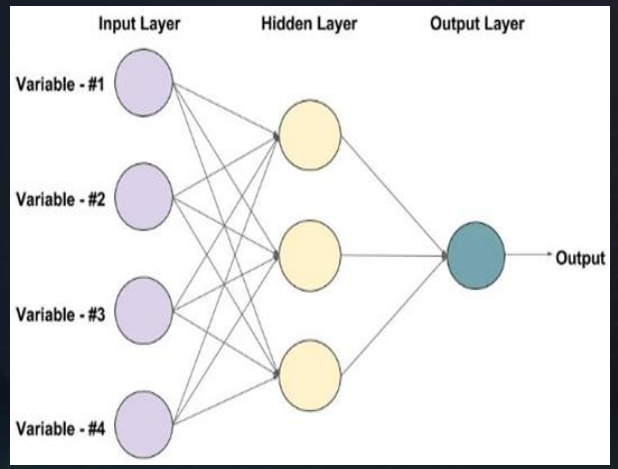
Cross validation + Hyperparameters



K-Neighbours



Neural Network



Model	*Best hyperparameters*
K-Nearest Neighbors	n_neighbors: 3 weights: uniform metric: euclidean
Support Vector Machine	kernel: linear C: 0.1 gamma: 0.001
Neural Network	hidden layers: 1 epochs: 100
Decision Tree (CART)	criterion: entropy max_depth: 2

Results: Accuracy

Accuracy responds to ‘What percent of predictions were correct calculated?’

Formula: TP + TN / (TP + TN + FP + FN)

where:

- TP (True Positives): instances where the model correctly predicts the positive class.
- TN (True Negatives): instances where the model correctly predicts the negative class.
- FP (False Positives): instances where the model incorrectly predicts the positive class when it is actually negative.
- FN (False Negatives): instances where the model incorrectly predicts the negative class when it is actually positive.

Accuracy of our Binary Classification Models

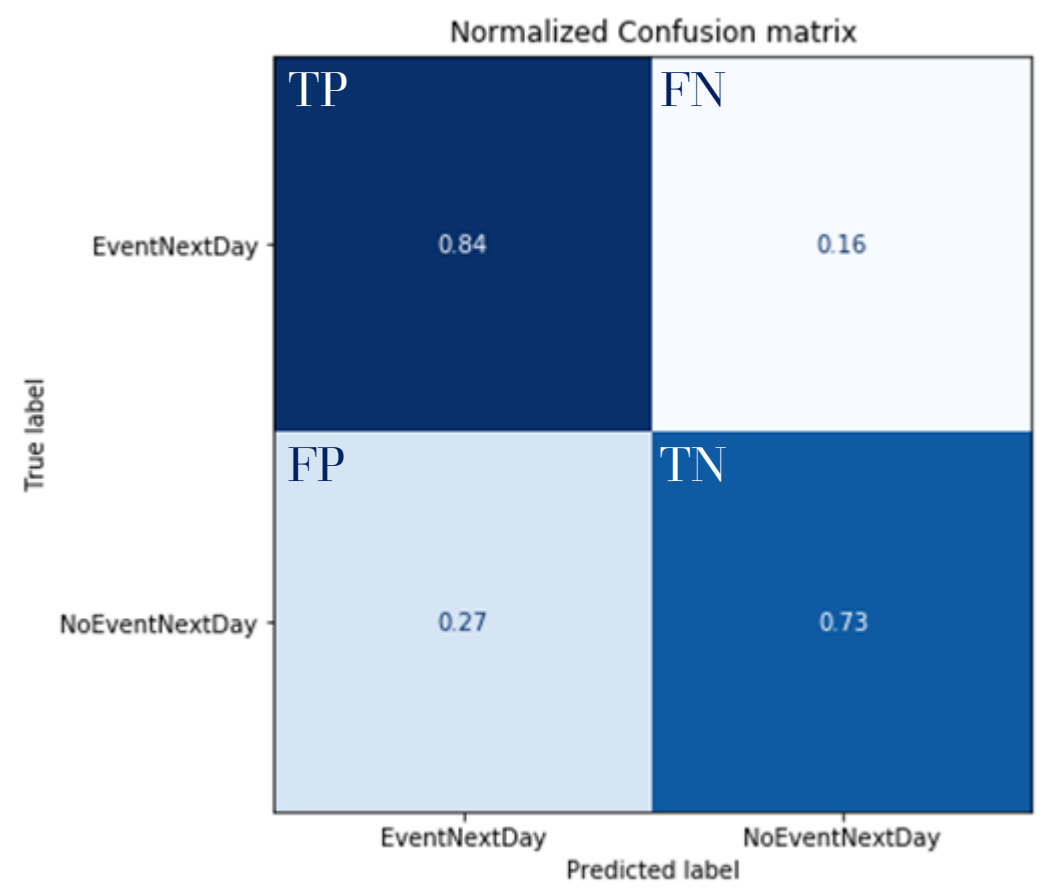
Model	*Accuracy (%)*
K-Nearest Neighbors	70,21
Support Vector Machine	74,47
Neural Network	72,34
Decision Tree (CART)	78,72

Best accuracy →

Results: Confusion matrix

Confusion matrix allow us to compare predicted values with actual values.

Decision Tree matrix



TP → in 84% of the instances there was a relevant event the next day to a check-in date and the model predicted that effectively there was

FN → in 16% of the instances there was a relevant event the next day to a check-in date and the model predicted that there was not.

FP → in 27% of the instances there was not a relevant event the next day to a check-in date and the model predicted that there was.

TN → in 73% of the instances there was not a relevant event the next day to a check-in date and the model predicted that effectively there was not.

Conclusions

Best classification model: Decision Tree

Precision:

78% right predictions when there was a relevant event

80% right predictions when there was not a relevant event

Difficulties:

Hotel policies in vacation days

Future lines of work

- ✓ Related to Events
  - Weighted events
  - Number of events per day
- ✓ Related to Bookings
  - Type of property: apartments, hostels, etc.
  - District and postal code
- ✓ Related to ML models
  - Other classification models: Naïve Bayes, Random Forest, etc.
  - With more data: LSTM

Thank you for your attention

Albert Navarro Pérez

alnape@uoc.edu