

## Repositório ISCTE-IUL

---

**Deposited in *Repositório ISCTE-IUL*:**

2018-10-19

**Deposited version:**

Post-print

**Peer-review status of attached file:**

Peer-reviewed

**Citation for published item:**

Moro, S., Rita, P., Oliveira, C., Batista, F. & Ribeiro, R. (2018). Leveraging national tourist offices through data analytics. *International Journal of Culture, Tourism, and Hospitality Research*. 12 (4), 420-426

**Further information on publisher's website:**

10.1108/IJCTHR-04-2018-0051

**Publisher's copyright statement:**

This is the peer reviewed version of the following article: Moro, S., Rita, P., Oliveira, C., Batista, F. & Ribeiro, R. (2018). Leveraging national tourist offices through data analytics. *International Journal of Culture, Tourism, and Hospitality Research*. 12 (4), 420-426, which has been published in final form at <https://dx.doi.org/10.1108/IJCTHR-04-2018-0051>. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

---

Use policy

Creative Commons CC BY 4.0

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

---

## Leveraging national tourist offices through data analytics

### Structured Abstract

#### *Purpose*

This study proposes a data-driven approach, based on open source tools, that makes it possible to understand customer satisfaction of the accommodation offer of a whole country.

#### *Design/methodology/approach*

The method starts by extracting information from all hotels of Portugal available at TripAdvisor through web scraping. Then, a support vector machine is adopted for modeling TripAdvisor score, which is considered a proxy of customer satisfaction. Finally, knowledge extraction from the model is achieved using sensitivity analysis to unveil the influence of features on the score.

#### *Findings*

The model of TripAdvisor score achieved a mean absolute percentage error of around 5%, proving the value of modeling the extracted data. The number of rooms of the unit and the minimum price are the two most relevant features, showing that customers appreciate smaller and more expensive units, while the location of the hotel does not hold significant relevance.

#### *Originality/value*

National tourist offices can use the proposed approach to understand what drives tourists' satisfaction, helping to shape a country's strategy. For example, licensing new hotels may take into account the unit size and other characteristics that make it more attractive to tourists. Furthermore, the procedure can be replicated at any time and in any country, making it a valuable tool for data-driven decision support on a national scale.

**Keywords:** national tourist offices, web scraping, data mining, sensitivity analysis, data analytics, online reviews.

**Article classification:** Research note

## 1. Introduction

The hospitality industry has developed online review platforms to empower tourists by giving them tools to share their experiences, influencing others through electronic word-of-mouth (eWOM) (Calheiros et al., 2017). The adoption of mobile applications has further emphasized social media use anytime, anywhere (Rita et al., 2018). Research has followed suit with recent studies in hospitality and tourism supporting such trend (Moro and Rita, 2018). However, national tourist offices (NTOs) are falling behind by offering limited data capabilities for the industry to explore in a data-driven world (Lam and McKercher, 2013).

This study proposes a novel approach for NTOs to use state-of-the-art data analysis open source tools for extracting data from TripAdvisor, one of the most prolific online reviews platform, and train a model of the customer score of accommodation units to leverage the knowledge NTOs may offer to tourism and hospitality practitioners and researchers.

## 2. Materials and methods

The proposed approach is exhibited in Figure 1. It consists of three steps represented in rounded rectangles. In the first one, web scraping was performed for automatically extracting information of all hotels available on TripAdvisor of an entire country. Web scraping consists in developing and running a script to efficiently download webpages' contents and to extract the needed information (Mitchell, 2015). The second step consisted in modeling the TripAdvisor score, which may be viewed as a proxy of customer satisfaction, using a classification method, fed with relevant features extracted from the collected data portraying the hotels. The classification method adopted in our experiments is the Support Vector Machine (SVM), an advanced machine learning technique that discerns the intrinsic relations between the input features and the outcome to model. It transforms the input  $x \in \mathcal{R}^M$  space into a high  $m$ -dimensional feature space by using a non-linear mapping that depends on a kernel (Cortes and Vapnik, 1995). Through this transformation, the algorithm identifies the linear hyper plane that best separates the input features space, using specifically selected points that constitute the support vectors, from which the name SVM derives (Steinwart and Christmann, 2008). For the kernel, we adopted the popular Gaussian method, which achieves good performance while requiring less parameters than other alternatives (e.g., Silva et al., 2018).

The third step consisted in extracting insightful knowledge from what the model apprehended from data. Sensitivity analysis applied to a data mining model enables to understand how each feature contributes to modeling the outcome feature (Barraza et al., 2018). Such task is accomplished by varying each of the input features through their range of possible values to assess model's sensitivity, i.e., if the outcome is little or highly affected by varying the input features. The data-based sensitivity analysis (DSA) is a specifically tuned method for assessing relations between the input features and their influence on the outcome through a fast-computational algorithm that uses a randomly selected sample from the dataset used to train the model (Cortez and Embrechts, 2013).

The techniques used in the three steps are well-established in the tourism literature. Yet, those are more often adopted for analyzing individual online reviews, especially, web scraping and SVM. This research takes as a baseline Moro et al. (2017)'s study but at a higher granularity level, i.e., instead of analyzing individual reviews, it focuses on hotels as units which have received scores from travelers. Furthermore, while the abovementioned study used 504 manually collected reviews, we included web scraping to benefit from the required volume to analyze an entire country's offer.

To test the approach, Portugal was the chosen destination country, as it is an attractive western European nation with a strong tourism market, with most of the hotels available in TripAdvisor (Moro et al., 2018). The data was extracted on the 27<sup>th</sup> of June 2017, with the web scraping procedure, taking around four hours to collect information on the 1,011 hotels registered in Portugal. The total number of reviews for those hotels is above 584k, validating the representativeness of TripAdvisor as a proxy to evaluate customer satisfaction on a large scale. The 22 features gathered which characterize the hotels are displayed in Table 1. Figures 2 to 4 show snapshots from TripAdvisor identifying the locations in the webpage from where each feature was extracted (the numbers correspond to column # in Table 1). Thereafter, the data cleaning procedure identified 44 hotels with missing values, which were discarded, considering those only represented 4.35% of the whole dataset, and leaving a total of 967 for training the model.

Features with the same value for the whole dataset were discarded (e.g. the country was always set as Portugal); likewise, features that are different for all cases (e.g., hotel name;

address) were also discarded; in both cases, the features do not hold patterns, being useless for modeling, as argued by Moro et al. (2016). The data preparation procedure enabled to identify that while TripAdvisor GreenLeader has four different levels, it is an award hardly granted; therefore, it was converted into a binary feature: 6% hotels hold it, while the remaining do not. Five of the hotel features (highlights; top.amenities; amenities; room.amenities; and things.to.do) contained lists of amenities and attractions the units offer, such as “Air Conditioning”. However, a large number of these characteristics overlap (see Figures 2 to 4), justifying the conversion of those five features into each individual amenity (with binary value: (Y)es, if the hotel offers the amenity; (N)o, otherwise), adding a total of 63 features to the dataset and discarding the original five. Further data analysis found that several of those 63 new features were highly unbalanced (threshold considered of 10%: one of the two categories of the binary value is represented in less than 10% of the cases), providing justification to discard 19 underrepresented features. The procedures undertaken left a dataset of 967 hotels and 55 features tuned for modeling TripAdvisor score.

For the implementation of the experiments, the two most sound and widely used open source scripting languages for data analysis were adopted, namely Python and R (Xia et al., 2010). The vast number of online communities and enthusiasts of both languages ensures a large number of open source packages for data analysis tasks. The experimental setup consisted in developing web scraping with the *wget* package and Python, while the “rminer” package from R was used for modeling and knowledge extraction (Figure 1) (Cortez, 2010). However, both of them could be used interchangeably for all the tasks, proving the versatility of open source scripting languages. Therefore, this study contributes also to highlight the usefulness of these tools for research.

### **3. Results and discussion**

The modeling step took as input the  $967 \times 56$  dataset (rows  $\times$  columns, i.e., 967 hotels characterized by a total of 55 features plus the score) for training the model, which was validated through a 10-fold cross validation scheme to ensure independency, as described by Moro et al. (2017). Model’s accuracy was assessed using two metrics: the mean absolute error (MAE), which measures the absolute average deviation between the predicted and the

real values (TripAdvisor score); and the mean absolute percentage error (MAPE), which is computed similarly to MAE, but relatively to the real value, thus represented in percentage (Hyndman and Koehler, 2006). The model achieved a MAE of 0.219, and a MAPE of 5.30%, a significant lower error than what was achieved by Moro et al. (2017), who modeled TripAdvisor review's score instead of hotel's score, thus justifying the reliability of the model for knowledge extraction. Nevertheless, results cannot be directly compared, not only because different target variables with different granularities are being modeling, but also because different datasets are being used.

The DSA assesses the relevance of each feature in terms of its contribution to modeling TripAdvisor score. Figure 5 displays the relative relevance of each of the ten most relevant features. It should be noted that each of the remaining 45 features holds an individual relevance below 2.5%. Results highlight two main features responsible for explaining 23.1% of the score granted, based on the SVM model, namely: the number of rooms and the minimum price for booking a reservation. The former was also considered relevant for the model achieved by Moro et al. (2017), confirming the relevance of accommodation units' size to customer experience, a result also highlighted in previous studies (e.g., Ariffin and Maghzi, 2012). Price is also an issue to which consumers tend to give high importance, especially in accommodation, due to the high level of competition and transparency in social media (María Munar, 2011). Therefore, the minimum price holds significant relevance to hotel's score. It is important to stress that these two features also emerge in the literature as highly relevant, confirming the value of the proposed approach for measuring customer satisfaction about accommodation offer.

The third most relevant feature belongs to the TripAdvisor brand and it is an Excellence reward based on several dimensions such as the score, the quantity and recency of reviews (Diappi, 2018). This shows the influence the TripAdvisor holds on the hospitality industry, a result corroborated by Moro et al. (2017), who found that TripAdvisor membership years also influence the score granted by users. The number of reviews also plays a role in influencing hotel score, emphasizing the power the TripAdvisor brand exerts on hotels. The number of stars within the star international system closes the top five ranking, with relevance slightly below 4%.

The remaining five features shown on Figure 5 represent amenities offered by the hotels, with relevance between 2.5% and 2.9%. The fact that none of the amenities appears among the five most relevant features points out that there is not an isolated amenity standing out; instead, each individual amenity plays a small role, but the combined relevance of all of them corresponds to 51.6% of the score, which is a highly significant contribution. Hence, hotel managers need to have a holistic perspective over the amenities offered to ensure high levels of customer satisfaction. Also interesting is the fact that the Portuguese region where the hotel is located plays little role in the score, unveiling a homogeneous distribution of hotels throughout the country when it comes to pleasing customers.

Finally, the influence of each of the top five specific features over the TripAdvisor score is analyzed individually (Figure 6). In a similar result to the study by Moro et al. (2017), smaller units tend to result in higher TripAdvisor scores. Current literature highlights that small scale hotels can offer personalized hospitality services, which are harder to implement in large hotels where economy of scale is used while serving to mass number of guests rather than tailoring for individual guests (Kurgun et al., 2011).

The main findings can be summarized as follows:

- Minimum price influences TripAdvisor score, with tourists visiting Portugal acknowledging the value for money ratio offered by more expensive hotels;
- TripAdvisor Excellence reward is recognized by hosts as a quality badge, resulting in higher scores for hotels which have been awarded;
- Units graded with more stars are also better valued by visitors (Jeong and Mindy Jeon, 2008).

#### **4. Conclusions**

This study contributes to enhance data analysis' capabilities of NTOs by introducing an automated approach for analyzing an entire country's accommodation offer from a customer satisfaction perspective. The proposed approach is based on a combination of known methods implemented using open source tools to model TripAdvisor score of individual hotels, helping to shed light about the main drivers for satisfying customers.

The experimental procedure was set in Portugal, a western leading European country as a tourist destination, although the method is directly generalizable to other countries. The managerial implications include an increase in awareness of customer sentiments towards accommodation units by tourist offices, helping to shape national tourism strategies. Additionally, since the method is automated, it may be replicated periodically to refresh the analysis with latest data reflecting the most updated trends. For the case of Portugal, both the number of rooms and the minimum price were found the most relevant features. Such acquired knowledge can be useful when deciding to legislate on the current or possible new units, helping to align demand and supply toward meeting customers' expectations.

The contributions underlined in this study can set roots for future research, including the application of the method to larger countries and the creation of separated and more tuned models based on other dimensions such as the hotel type to assess feature relevance on different perspectives. Nevertheless, there are some limitations that must be pointed out. Web scraping is highly dependent on the source webpages' format. This implies that the extraction process will need to be revised every time TripAdvisor changes the HTML tags and style sheets used by the process. Furthermore, some of the extracted features may change or even be removed by TripAdvisor. Yet, while this does not happen, the procedure remains directly replicable. Thus, monitoring each execution will help in identifying source changes and subsequently adapt the extraction script. Another limitation is related to the localized nature of the study. Further research is needed in different destinations to assess if the findings can be generalized.

## **References**

- Ariffin, A.A.M., and Maghzi, A. (2012), "A preliminary study on customer expectations of hotel hospitality: Influences of personal and hotel factors", *International Journal of Hospitality Management*, Vol.31 No.1, pp.191-198.
- Barraza, N., Moro, S., Ferreyra, M., & de la Peña, A. (2018). Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study. *Journal of Information Science*, DOI:10.1177/0165551518770967.



Calheiros, A.C., Moro, S., and Rita, P. (2017), "Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling", *Journal of Hospitality Marketing & Management*, Vol.26 No.7, pp.75-693.

Cortes, C., and Vapnik, V. (1995), "Support vector machine", *Machine Learning*, Vol.20 No.3, pp.273-297.

Cortez, P. (2010), Data mining with neural networks and support vector machines using the R/rminer tool, *Advances in data mining. Applications and theoretical aspects*, pp.572-583.

Cortez, P., and Embrechts, M.J. (2013), "Using sensitivity analysis and visualization techniques to open black box data mining models", *Information Sciences*, Vol.225, pp.1-17.

Diappi, L. (2018). The Tourism as Local Development Leverage: The Restaurant/Guest house of Olga's and the Professional School YCTC in Livingstone, Zambia. In *Sustainable Urban Development and Globalization* (pp.197-208). Springer, Cham.

Hyndman, R.J., and Koehler, A.B. (2006), "Another look at measures of forecast accuracy", *International Journal of Forecasting*, Vol.22 No.4, pp.679-688.

Jeong, M., and Mindy Jeon, M. (2008), "Customer reviews of hotel experiences through consumer generated media (CGM)", *Journal of Hospitality & Leisure Marketing*, Vol.17 No.1-2, pp.121-138.

Kurgun, H., Bagiran, D., Ozeren, E., and Maral, B. (2011), "Entrepreneurial marketing-The interface between marketing and entrepreneurship: A qualitative research on boutique hotels", *European Journal of Social Sciences*, Vol.26 No.3, pp.340-357.

Lam, C., and McKercher, B. (2013), "The tourism data gap: The utility of official tourism information for the hospitality and tourism industry", *Tourism Management Perspectives*, Vol.6, pp.82-94.

María Munar, A. (2011), "Tourist-created content: rethinking destination branding", *International Journal of Culture, Tourism and Hospitality Research*, Vol.5 No.3, pp.291-305.

Mitchell, R. (2015), *Web scraping with Python: collecting data from the modern web*, O'Reilly Media, Inc.

Moro, S., Rita, P. and Vala, B. (2016), "Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach", *Journal of Business Research*, Vol.69, pp.3341-3351.

Moro, S., Rita, P., and Coelho, J. (2017), “Stripping customers' feedback on hotels through data mining: the case of Las Vegas Strip”, *Tourism Management Perspectives*, Vol.23, pp.41-52.

Moro, S., and Rita, P. (2018), “Brand strategies in social media in hospitality and tourism”, *International Journal of Contemporary Hospitality Management*, Vol.30 No.1, pp.343-364.

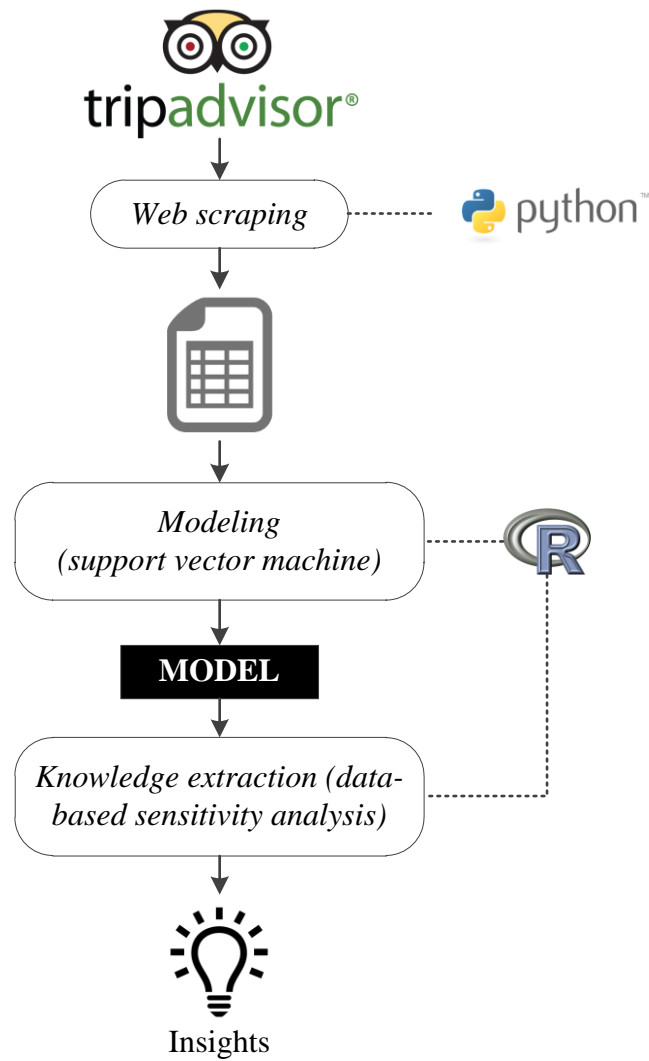
Moro, S., Rita, P., and Oliveira, C. (2018), “Factors influencing hotels’ online prices”, *Journal of Hospitality Marketing & Management*, Vol.27, No.4, pp.443-464.

Rita, P., Oliveira, T., Estorninho, A., and Moro, S. (2018), “Mobile services adoption in a hospitality consumer context”, *International Journal of Culture, Tourism and Hospitality Research*, Vol.12 No.1, pp.143-158.

Silva, A.T., Moro, S., Rita, P., and Cortez, P. (2018), “Unveiling the features of successful eBay smartphone sellers”, *Journal of Retailing and Consumer Services*, Vol.43, pp.311-324.

Steinwart, I., and Christmann, A. (2008), *Support vector machines*, Springer Science & Business Media.

Xia, X.Q., McClelland, M., and Wang, Y. (2010), “PypeR, A Python package for using R in Python”, *Journal of Statistical Software*, Vol.35 No.2, pp.1-8.



**Figure 1** - Methodological approach.

1

# As Janelas Verdes

2 3 4 5 6

887 Reviews 14 #6 of 247 Hotels in Lisbon

Rua das Janelas Verdes, 47 Lisbon 1200-690 Portugal 21 396 8143 Hotel website E-mail hotel


22

Best prices for your stay

07/16/2017 07/17/2017

1 room 2 adults 0 children

Expedia	€280 €236	View Deal >
PRESTIGIA	€280 €236	View Deal >
Booking.com	€280 €236	View Deal >



15

Figure 2 – Information extracted from TripAdvisor’s main section.

## About

**Awards & Recognition** 12

Travelers' Choice 2017 Winner

GreenLeaders Silver level 16

Certificate of Excellence 13

**Details** 11

PRICE RANGE 10 €133 - €599 (Based on Average Rates for a Standard Room)

HOTEL CLASS 7 ★★★★★

HOTEL STYLE

- #6 Luxury Hotel in Lisbon
- #7 Romantic Hotel in Lisbon
- #8 Business Hotel in Lisbon
- #10 Green Hotel in Lisbon
- #18 Family Hotel in Lisbon

ROOM TYPES

Non-Smoking Rooms , Family Rooms , Suites

NUMBER OF ROOMS 9 29

RESERVATION OPTIONS

TripAdvisor is proud to partner with Booking.com, TripOnline SA, Expedia, Prestigia, Traveltool, S.L.U., ACCOR, Odigeo, Hoteis.com and HotelQuickly so you can book your As Janelas Verdes reservations with confidence. We help millions of travelers each month to find the perfect hotel for both vacation and business trips, always with the best discounts and special offers.

LOCATION 4

Portugal > Central Portugal > Lisbon District > Lisbon

**Amenities** 18 19 20 21


TOP AMENITIES	HOTEL AMENITIES	ROOM AMENITIES	THINGS TO DO
Bar/Lounge	Dry Cleaning	Air Conditioning	Bar/Lounge
Room Service	Laundry Service		
Free High Speed Internet ( WiFi )	Multilingual Staff		
Breakfast included	Room Service		
	Breakfast included		
	Concierge		
	Babysitting		
	Breakfast Available		






**Hotel description**

As Janelas Verdes is a small charming boutique hotel in the historic center of Lisbon. This 18th-century palace is next to the National Ancient Art Museum. With its small garden and the top floor library overlooking the river, this mansion transmits a romantic and welcoming feeling. The rooms are sunny and cheerful and the Tagus is close by.




Figure 3 – Information extracted from TripAdvisor’s “about” section.




## Overview

4.5  887 reviews

Excellent		78%
Very good		18%
Average		2%
Poor		1%
Terrible		1%

TRAVELERS TALK ABOUT

-  "honesty bar" (89 reviews)
-  "third floor" (58 reviews)
-  "top floor" (56 reviews)

Free Wifi   

Breakfast included

Air Conditioning

Non-Smoking Hotel 17

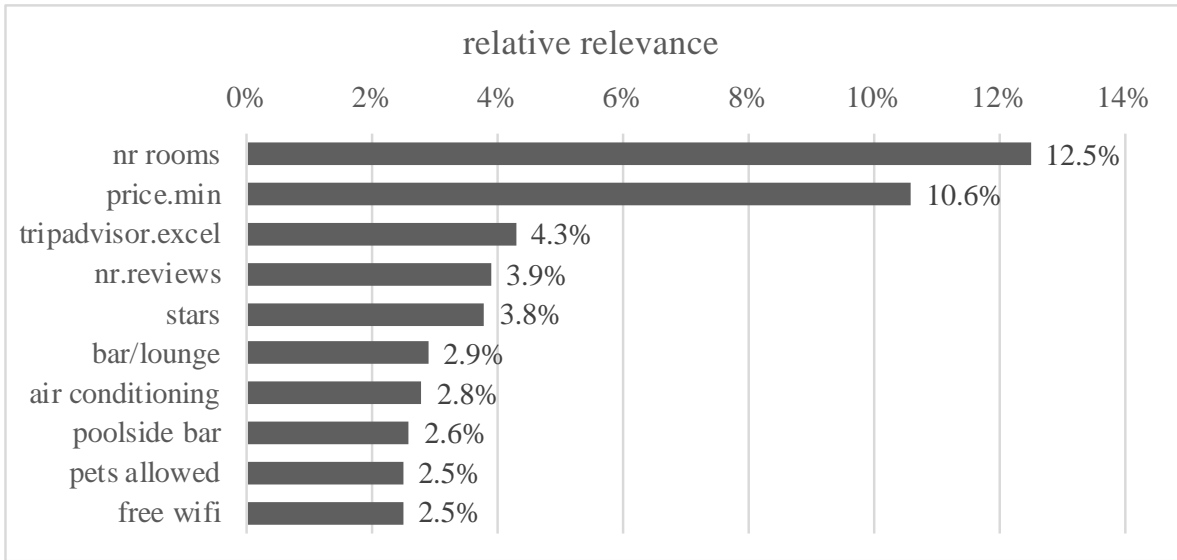
★★★★★  
4.0 Star Hotel

[All hotel details](#)

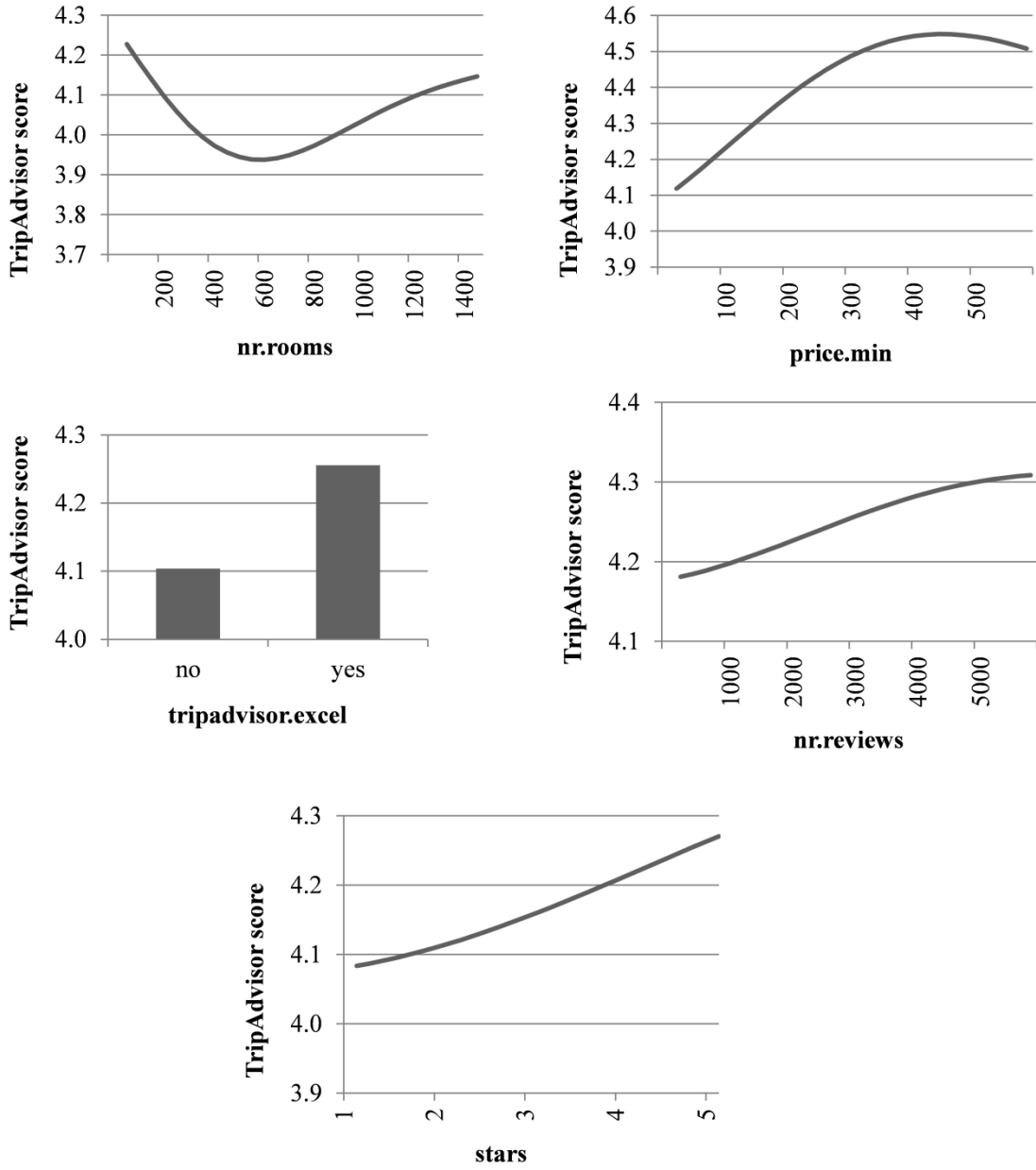
OFFERS FROM AS JANELAS VERDES

Offers & Announcements: [Free Breakfast until 12h](#)

Figure 4 – Information extracted from TripAdvisor’s “Overview” section.



**Figure 5** – Features' relevance.



**Figure 6** – Influence of the top five features on TripAdvisor score.

**Table 1** - Features extracted.

#	Feature	Type and description	Included
1	name	Name of the hotel	N
2	addr.street	Street address	N
3	addr.local	Town	N
4	addr.region	{Northern Portugal; Algarve; Alentejo; Madeira Islands; Central Portugal; Azores }	Y
5	addr.postal	Postal zipcode	N
6	addr.country	Country="Portugal"	N
7	stars	Stars in the international ranking system (1 to 5)	Y
8	has.website	If the hotel has website ("Y"es / "N"o)	Y
9	nr.rooms	Number of rooms of the unit	Y
10	price.min	Minimum price	Y
11	price.max	Maximum price	Y
12	trav.choice	TripAdvisor's travellers' choice award ("Y"es / "N"o)	Y
13	tripadvisor.excel	TripAdvisor's certificate of excellence ("Y"es / "N"o)	Y
14	nr.reviews	Number of reviews the hotel has on TripAdvisor	Y
15	nr.photos	Number of photos published by the hotel on TripAdvisor	Y
16	eco.level	TripAdvisor GreenLeaders' level {Bronze, Silver, Gold, and Platinum}. This feature was transformed into binary since only 6% of the units were GreenLeaders: if the hotel is GreenLeader, then "Y"; else "N"	
17	highlights	Each of these five features held a list of individual hotel characteristics (63 different characteristics for the whole dataset, with several overlaps). Therefore, 63 different features were added as columns to the dataset (e.g., "Free Wifi", "Air Conditioning", "Room Service")	
18	top.amenities		
19	amenities		
20	room.amenities		
21	things.to.do		
22	score	TripAdvisor's score of the unit (3 to 5 for the Portuguese units)	Y