

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:

2019-03-29

Deposited version:

Pre-print

Peer-review status of attached file:

Unreviewed

Citation for published item:

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

Further information on publisher's website:

10.1016/j.tmp.2017.04.003

Publisher's copyright statement:

This is the peer reviewed version of the following article: Moro, S., Rita, P. & Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: the case of Las Vegas Strip. *Tourism Management Perspectives*. 23, 41-52, which has been published in final form at <https://dx.doi.org/10.1016/j.tmp.2017.04.003>. 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.

1 **Stripping customers' feedback on hotels through data mining: the case of Las Vegas Strip**

2

3 **Abstract**

4 This study presents a data mining approach for modeling TripAdvisor score using 504 reviews
5 published in 2015 for the 21 hotels located in the Strip, Las Vegas. Nineteen quantitative features
6 characterizing the reviews, hotels and the users were prepared and used for feeding a support
7 vector machine for modeling the score. The results achieved reveal the model demonstrated
8 adequate predictive performance. Therefore, a sensitivity analysis was applied over the model for
9 extracting useful knowledge translated into features' relevance for the score. The findings
10 unveiled user features related to TripAdvisor membership experience play a key role in
11 influencing the scores granted, clearly surpassing hotel features. Also, both seasonality and the
12 day of the week were found to influence scores. Such knowledge may be helpful in directing
13 efforts to answer online reviews in alignment with hotel strategies, by profiling the reviews
14 according to the member and review date.

15

16 **Keywords**

17 Customer feedback; customer reviews; online reviews; knowledge extraction; data mining;
18 modeling; sensitivity analysis; Las Vegas.

19

20

21 **1. Introduction**

22 The Online Travel Agencies (OTA) are now the most used tool of travel booking, both for the
23 means of transport and accommodation (Mauri & Minazzi, 2013) and, consequently, online
24 reviews have been exponentially increasing its use and impact in the hospitality industry over the
25 last years, due to the social media and technological evolution. In fact, nowadays potential hotel
26 customers search for online feedback before travelling and base their purchase decisions on
27 online reviews (Mauri & Minazzi, 2013). Therefore, electronic word-of-mouth (eWOM), which
28 according to Henning-Thurau et al. (2004, pp. 39) is defined as “any positive or negative
29 statement made by potential, actual or former customers about a product or company, which is
30 made available to a multitude of people and institutions via the internet”, has become a huge
31 aspect when travelling, since currently every consumer has access to the internet and can easily
32 express either positive or negative feedback. Most importantly, it is an online tool to be used
33 when others seek for advice as part of the decision-making process, such as where to stay,
34 especially in hospitality industry, as consumers are purchasing an experience and cannot predict
35 its evaluation (Sparks & Browning, 2011). Moreover, holidays can be considered as a high risk
36 and involvement purchase, due to its usual personal importance and also high value of money
37 (Papathanassis & Knolle, 2011). Service quality is a determinant of the customer’s perceptions
38 and their feedback. The ideal would be that the target’s expectations meet the perceptions, which
39 will directly influence a positive word of mouth, contributing for a development of reputation
40 and trust (Corbitt et al., 2003). Hence, research contributions that unveil and provide in-depth
41 understanding on the features that have the most impact on customer feedback are valuable for
42 sustainable decision making.

43 Previous studies have been conducted by various researchers in order to understand and explain
44 the influence and impact of online reviews in the hospitality industry. One of the most common
45 methods used include the analysis of variance (ANOVA) technique, which is offered in many
46 data analysis’ solutions such as the IBM SPSS software. For example, Vermeulen and Seegers
47 (2009) adopted the ANOVA for testing whether or not the user-generated online reviews
48 influence the consumer choice. In a parallel line of research, Jeong and Jeon (2008) also used the
49 ANOVA for analyzing the impact of five relevant features (hotel ownership, stars, number of
50 rooms, room rates, and popularity index) in scoring New York hotels on TripAdvisor’s nine

51 rating items (e.g., location; cleanliness). Their results show that both the number of stars and
52 room rates influence the rating items from TripAdvisor. A similar study focused on analyzing the
53 relationship between the hotel specific rating items used by Expedia (service, condition,
54 cleanliness, and comfort) in the hundred largest US cities. Again, statistical tools and methods
55 were adopted, including the ANOVA (Stringam et al., 2010). Additionally, Sparks and Browning
56 (2011) went further on their research and studied the fact that a consumer generated quantitative
57 rating could be associated together with the actual written review. In a more recent data-driven
58 study, it has been shown through regression models that the financial benefits of an online
59 review from TripAdvisor conceal intrinsic value to the hospitality industry (Neirotti et al., 2016).
60 Nevertheless, the majority of previous recent studies are focused on the impact of the text review
61 itself, applying text mining techniques, which aim to extract meaningful knowledge from a
62 variety of textual data and find relationships and patterns within such unstructured information
63 (Calheiros et al., 2017).

64 Different studies are aligned through similar conclusions regarding the fact that text mining
65 applications to social media data (i.e. any online platform where customers can exchange
66 information) can provide significant insights on the human behavior and interaction (e.g., He et
67 al., 2013). However, while several studies are known using data mining for sentiment
68 classification and opinion mining (e.g., Schuckert et al., 2015), none was found up to the present
69 adopting a quantitative approach on modeling tourists' reviews through advanced data mining
70 techniques for extracting the influence of hotels' and users' features on the score provided by
71 users. Nevertheless, the quantitative score is the first relevant information users see when they
72 search for feedback information on their next stay (O'Connor, 2010). Understanding which
73 profiles of users are most likely to result in poorer scores may help to shape strategies for
74 choosing the users to whom to answer in TripAdvisor, as answering all users is time-consuming
75 and requires significant human effort (Nguyen & Coudounaris, 2015). Thus, such directed effort
76 can lead to an improvement in positive eWOM, as the responses may be framed for specific
77 users. Additionally, identifying the features influencing scores granted may help to profile users,
78 helping to identify outlier behaviors and possible reputation attacks (Buccafurri et al., 2014).
79 Since users are influenced by hotels (Casalo et al., 2015), including hotel features in a unique
80 model allows to obtain explanatory knowledge intersecting both dimensions. Hence, the present
81 study aims at filling such research gap by focusing on online reviews' quantitative features such

82 as number of stars of the hotel and number of helpful votes the user has received in order to build
83 a predictive model of the tourists' score on the hotels. The knowledge built upon such model
84 may help to shed some light on what drives the rating of a hotel, potentiating meaningful
85 information to support managerial decisions.

86 The proposed data mining approach is an attempt to answer the following research questions:
87 Can the score of an online hospitality review be predicted using as input only quantitative data?
88 What are the features that influence most the review scores in hospitality? How does each of
89 those features affect the score and can this knowledge be useful for hotel managers?

90 Concluding, the main goals and contributions of this study are as follows:

- 91 • Creating a model that predicts the review score based on quantitative features of the
92 user/reviewer and the hotel, as well as the period of time of the specific stay;
- 93 • Contributing to research on customers' feedback and online reviews by providing a novel
94 approach on the used data, the quantitative features, as opposed to the most common
95 analyses of the reviews' text itself;
- 96 • Understanding how users are inherently influenced by hotels' features when submitting
97 numerical scores besides text comments on online platforms, such as TripAdvisor.

98 The next section describes the background concepts, such as the history and evolution of online
99 reviews, as well as the methods for knowledge extraction from data, its dimensions and its use in
100 the industry. Section 3 discusses the materials (e.g. input dataset) and procedures that were
101 applied in the experiment. Then, the results are shown and a critical discussion takes place on the
102 findings section. Finally, the main conclusions of this research are drawn.

103

104 **2. Theory**

105 **2.1. Online reviews**

106 In 2004, Tim O'Reilly coined the term Web 2.0 as the network connecting all devices to which
107 individual users contribute largely by sharing their experiences in numerous ways, therefore
108 becoming one of the most relevant sources of the internet through the so called user-generated
109 contents (O'Reilly & Battelle, 2009). Such internet evolution effectively became a global

110 revolution, including the tourism and hospitality industry by adding new online sources of
111 information to the existing hotel and tourism companies' websites, implying users are becoming
112 key-players in influencing others through their online reviews (Law et al., 2014).

113 Traditional websites have therefore evolved by increasing interactivity level to keep pace with
114 Web 2.0 new demands. However, in this new information-driven era, specialized user-content
115 sites and applications such as wikis, forums, blogs, social networks and especially online
116 reviews' sites for the case of tourism and hospitality have underpinned a new paradigm in which
117 the user is at the center of the network, leading to a mutual exchange and sharing of values
118 (Liburd, 2012). As Zeng and Gerritsen (2014, pp. 27) pointed out, "leveraging off social media
119 to market tourism products has proven to be an excellent strategy".

120 Several studies are found based on online reviews for tourism and hospitality, especially to
121 analyze how exchanges of information influence directly the consumer choices regarding a
122 certain hotel (e.g., Park & Nicolau, 2015), with most of them concluding that an exposure to an
123 online hotel positive review will increase the average probability of that consumer to book a
124 room in the same hotel. Features such as the number of stars have shown to positively influence
125 the score granted by users on online reviews (Hu & Chen, 2016). In fact, users expect higher
126 rated hotels (i.e., with a higher number of stars) to have more positive reviews, according to
127 Phillips et al. (2015). The latter study goes further on the analysis by revealing that larger hotel
128 units with higher number of rooms do not directly translate into high revenue. By building an
129 artificial neural network model, Phillips et al. (2015), managed to obtain a unique and valuable
130 model explaining the intersection of a few hotel and regional characteristics, with the number of
131 reviews. However, the same study did not include in its model the features of each individual
132 user, as it was aimed for a granularity at the hotel level. Fang et al. (2016) confirmed through an
133 econometric model that user/reviewer characteristics affect the perceived value of the reviews
134 made, proving that user features should also be accountable when modeling online reviews'
135 scores.

136 The recent study by Kim et al. (2017), comparing both TripAdvisor scores and traditional
137 customer satisfaction through travel intermediaries, found out that online reviews play a more
138 significant role in explaining hotel performance metrics than traditional feedback. Such finding
139 can be linked to users' perceptions, as a vast majority of them believe in online reviews

140 published on platforms such as TripAdvisor, being directly influenced by scores granted by other
141 users, even though reputation attacks seem to occur often in the hospitality industry (Filieri et al.,
142 2015). Kwok et al. (2017) presented an analysis of 67 online reviews' articles published between
143 2000 and 2015. The same study reveals most of research focuses on TripAdvisor and,
144 specifically, on hotel reviews, with a significant increase in the number of publications after
145 2012. Nevertheless, most of the quantitative research analyzed by the aforementioned study
146 employs active user participated methods such as surveys; on the opposite, qualitative research
147 based on textual comments adopts passive data collection and analysis methods. The present
148 research aims at filling such gap by adopting a passive data analysis through advanced data
149 mining modeling of the score based on quantitative features characterizing both users and hotels,
150 which have proven to affect the review score.

151

152 **2.2. Data mining in tourism and hospitality**

153 A large amount of studies by different authors were conducted where data mining procedures
154 were undertaken on tourism and hospitality data. Min et al. (2002) studied the application of data
155 mining, more specifically using decision tree modeling in order to develop the profile of a certain
156 group of customers within different hotels. In another paper, data mining has also been studied
157 regarding its importance and influence in a hotel's marketing department and how it may help in
158 providing a way where companies can reach to their potential customers, know them and their
159 behavior (Magnini et al., 2003). Song and Li (2008) analyzed tourism and hospitality literature
160 published between 2000 and 2007 for modeling tourism demand and identified several data
161 mining techniques that have started to be adopted alongside with traditional models such as the
162 integrated autoregressive moving-average models (ARIMA). From the articles they analyzed,
163 there is a general impression that advanced techniques such as support vector machines
164 outperform traditional ARIMA models, although there is not a single technique that achieves
165 always better results than the others, thus the accuracy is dependent on the specific context and
166 data that defines the problem. However, as Moro and Rita (2016) discussed after analyzing fifty
167 recent articles published between 2013 and 2016, most of the data analysis procedures conducted
168 on tourism and hospitality data are still based on ARIMA models.

169 As stated previously, a large number of the published research based on customer feedback and,
170 in particularly, in tourism and hospitality, focus on the analysis of the textual contents from
171 users' reviews through techniques based on text mining and sentiment analysis. As an example,
172 Ye et al. (2009b) applied sentiment classification techniques in various online reviews from
173 diverse travel blogs, comparing them with three different supervised machine learning
174 algorithms. In a different line of research, Cao et al. (2011) investigated the impact of online
175 review features hidden in the textual content of the reviews on the number of helpful votes of
176 such review texts by applying text mining for extracting the review's characteristics, while Guo
177 et al. (2017) applied text mining and topic modeling for unveiling several dimensions that
178 hoteliers need to control for managing interactions with visitors. However, several issues and
179 challenges are brought up when it comes to use text mining. The most widely discussed are
180 context specificities associated with the user and problem being dealt with, language barriers,
181 and human communication issues such as sarcasm and irony (Aggarwal & Zhai, 2012; Ampofo
182 et al., 2015). For example, many of the reviews published in TripAdvisor are made in each user's
183 native languages. Also, syntactic errors are common on this platform, as users are not concerned
184 with typing errors. Despite some advances in these domains, the intrinsic linguistic subjectivity
185 is still a challenge yet to be overcome. Such difficulty does not exist when only quantitative data
186 based on numerical or categorical features are used for feeding a model based on a data mining
187 technique.

188 In TripAdvisor, users are able to rank hotel units by providing a quantitative score (O'Connor,
189 2010). While a few recent studies have adopted data mining techniques for discovering the
190 influence of online reviews (e.g., Qazi et al., 2016, modeled the helpfulness of online reviews),
191 none considered using an advanced modeling technique encompassing dimensions such as hotel,
192 user, and review features. Therefore, the contribution and innovation to the hospitality industry
193 and literature brought by the present paper is the application of data mining to all the quantitative
194 features that can be collected from TripAdvisor, in order to model the score given by the
195 reviewers, based on their experience as TripAdvisor users and the hotel's characteristics, instead
196 of the common text mining applied to the written comments published by users.

197

198

199 **3. Materials and methods**

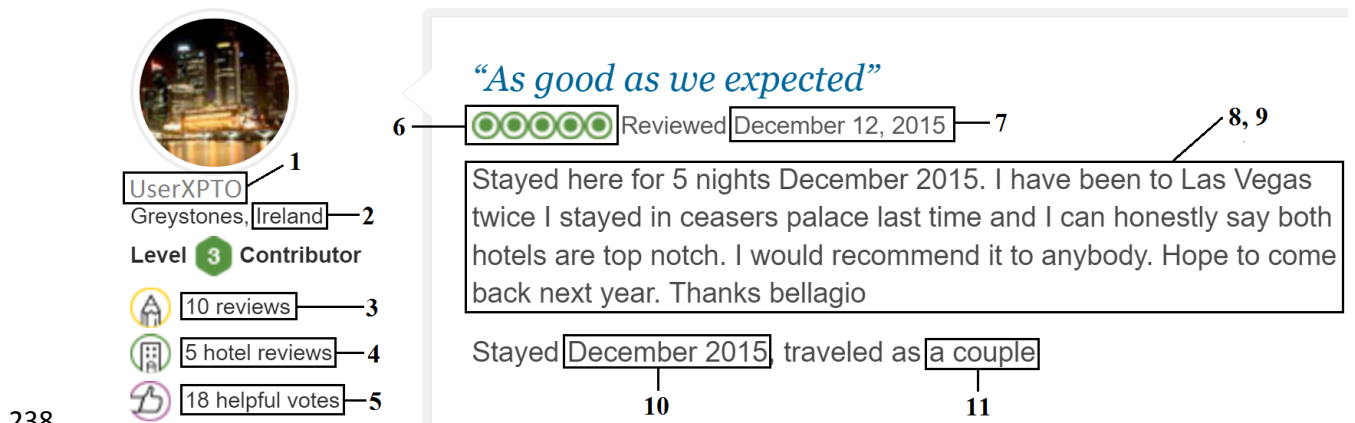
200 **3.1. Data collection and preparation**

201 After defining the problem, data collection and preparation is the next key step for compiling a
202 dataset that serves as input for modeling. Such dataset is the building block essential for
203 unveiling knowledge through a data mining modeling technique. Moreover, the dataset needs to
204 be composed of a table where each row represents an instance of the problem being addressed
205 and each column represents a feature that characterizes that instance (Witten & Frank, 2005).

206 Since TripAdvisor owns several domains to cover suffixes from several countries, the data was
207 collected from the TripAdvisor.com website, as the .com is considered the base site where there
208 are reviews belonging to users from every part of the world. Then, it was necessary to filter the
209 information by location, i.e. Las Vegas, Nevada, and more specifically filtering by hotels in the
210 Strip avenue. Las Vegas, the so called city of sin, born eighty years ago over a desert where
211 hotels started to be built and forming one of the most entertaining cities in the world, is driven by
212 tourism and gambling pleasure (Rowley, 2015). Between 2000 and 2010, Las Vegas remained
213 the fastest growing large city in the United States (Mackun et al., 2011). Regarding previous
214 studies conducted about and within Las Vegas, mainly in the Strip, the most popular avenue of
215 the city and with the largest supply of hotel rooms, Ro et al. (2013) discussed the affective image
216 of the major hotel's positioning, whereas the city's success as a gaming destination due to the
217 government and private institutions was proposed and analyzed by Lee (2015). Given the interest
218 triggered by Las Vegas hospitality, a large number of reviews are available, which is a
219 requirement for the proposed data-driven study. The present research started by collecting all the
220 features available on TripAdvisor's webpages from several online reviews published during
221 2015 and targeting hotels located in the Strip avenue.

222 As a result, a list of 21 different hotels was displayed, allowing to choose a hotel at a time in
223 order to extract the data from each one of them. When opening one of the chosen hotels' pages,
224 access is gained to various information regarding the hotel, such as its address, general quality
225 rating, individual reviews, photos and videos from both the hotel and the previous customers and
226 also the hotel's features. Once the hotel was selected, the procedure undertaken consisted in
227 collecting the data by extracting two reviews per month from the year of 2015, repeating this
228 process for all the 21 hotels. The uniform distribution of the reviews spanned through the

229 different months provided data for building a model that also considered the seasonality effect
 230 known of tourism (Song & Li, 2008). Starting by filtering the time of the year for the period of
 231 stay (Dec-Feb; Mar-May; Jun-Aug; Sep-Nov), the search focused on selecting the most
 232 completed reviews in order to provide all the information and variables needed until the 24
 233 reviews per year were accomplished. After choosing the reviews, all the features identified from
 234 each review, including user characteristics, were collected into a single table, including the score,
 235 as it is shown in Figure 1 where each square represents a fragment of data collected. The textual
 236 review was also collected, in case it would be needed in future research. The numbers identify
 237 the feature extracted enumerated under parenthesis in the column “origin” of Table 1.

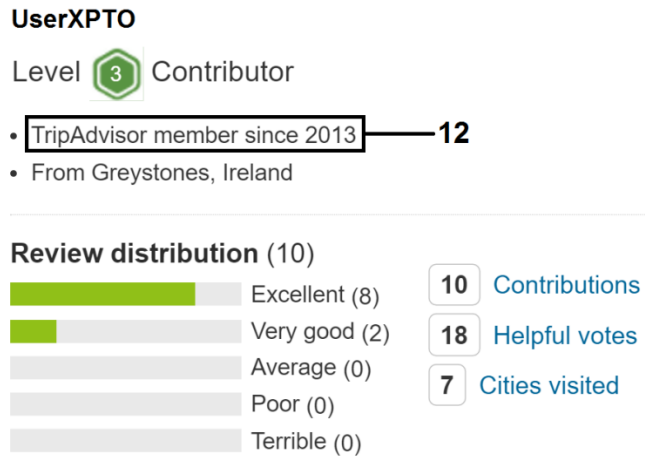


238
 239 **Figure 1** - Review and user features extracted.

240 To obtain the date the user has registered in TripAdvisor, it was enough to pass with the cursor
 241 over the username to get such additional information, displayed in Figure 2.

242 Finally, the webpage with the information supplied by TripAdvisor for each of the 21 hotels was
 243 accessed to gather relevant features from each hotel (e.g., the link for the Bellagio is:
 244 [https://www.tripadvisor.com/Hotel_Review-g45963-d91703-Reviews-Bellagio_Las_Vegas-](https://www.tripadvisor.com/Hotel_Review-g45963-d91703-Reviews-Bellagio_Las_Vegas-Las_Vegas_Nevada.html)
 245 [Las_Vegas_Nevada.html](https://www.tripadvisor.com/Hotel_Review-g45963-d91703-Reviews-Bellagio_Las_Vegas-Las_Vegas_Nevada.html)). While a large number of features are available, collecting all of them
 246 would make it difficult for an advanced data mining modeling technique to disentangle how each
 247 of them affects scores. Thus, to choose the most adequate features, an independent hotel manager
 248 aware of Las Vegas offer was asked to share his expertize on choosing the features.

249 Figure 3 shows a snap-shot of the section where the features from hotel's amenities were
 250 extracted, whereas Figure 4 shows the section from where additional relevant features such as
 251 hotel's stars and number of rooms were collected.

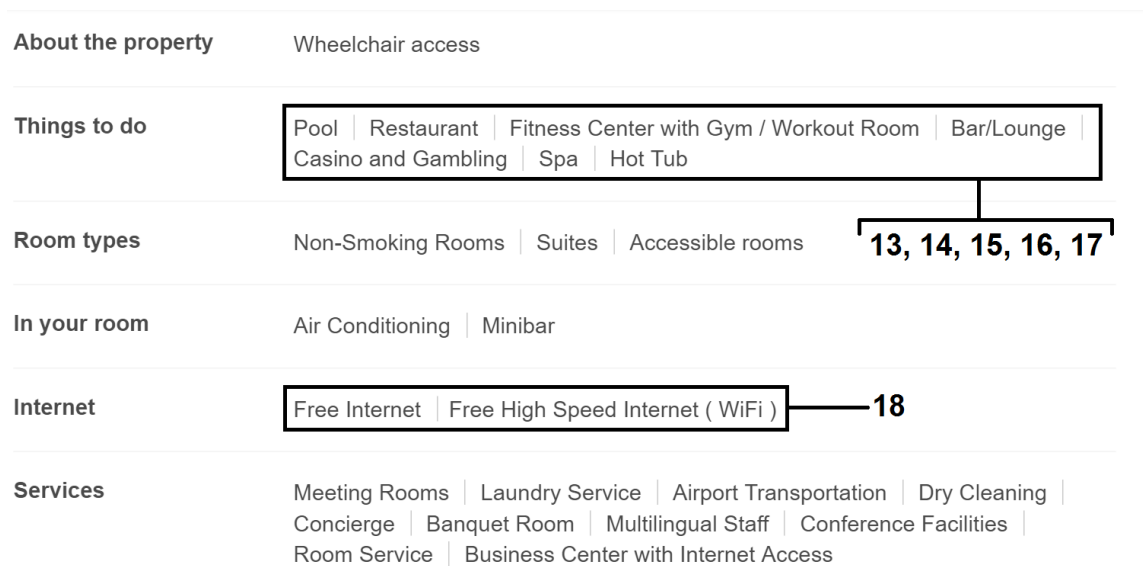


252

253

Figure 2 - Extraction of member registered date.

254



255

256

Figure 3 - Extraction of hotel's amenities features.

257

258

Additional Information about **Bellagio Las Vegas** —19

Address: 3600 Las Vegas Blvd S, Las Vegas, NV 89109-4303

Location: United States > Nevada > Las Vegas

Price Range: \$188 - \$445 (Based on Average Rates for a Standard Room)

Hotel Class: **5 star — Bellagio Las Vegas 5*** —20

Number of rooms: **3933** —21

Reservation Options:

TripAdvisor is proud to partner with Booking.com, Expedia, Hotels.com, Odigeo, Agoda, Prestigia and HotelsClick so you can book your Bellagio Las Vegas 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.

259

260

Figure 4 - Extraction of additional hotel's features.

261

262 Table 1 exhibits the features collected, identified by the “origin” equals to “extracted”, with the
 263 parenthesized numbering in the same column corresponding to the locations from where each
 264 feature was collected, as identified in Figures 1 to 4. The source type groups features into three
 265 categories, review features, user features, and hotel features, whereas the data type relates to the
 266 types of values that can be assumed by each feature, with the categorical type corresponding to a
 267 fixed number of enumerated values (e.g., the “gym” feature can assume “yes” or “no”) and the
 268 numerical type corresponding to an ordinal numbered feature. Dates are a particular type of
 269 numerical features due to its format restrictions, while “text” type corresponds to unstructured
 270 data (here reserved for the “review text”).

271

Table 1 - List of features.

Feature name	Origin	Source type	Data type	Description	Status
Username	Extracted (1)	User	Categorical	Username as registered in TripAdvisor	Excluded
User country	Extracted (2)	User	Categorical	User's nationality	Included
Nr. Reviews	Extracted (3)	User	Numerical	Number of reviews	Included
Nr. Hotel reviews	Extracted (4)	User	Numerical	Total hotel reviews	Included
Helpful votes	Extracted (5)	User	Numerical	Helpful votes regarding reviews's info	Included
Score	Extracted (6)	Review	Numerical	Review score {1,2,3,4,5}	Included
Review date	Extracted (7)	Review	Date	Date when the review was written	Transformed
Review text	Extracted (8)	Review	Text	Textual content of the review	Excluded
Review language	Extracted (9)	Review	Categorical	Language of the review	Excluded

Period of stay	Extracted (10)	Review	Categorical	Period of stay: {Dec-Feb, Mar-May, Jun-Aug, Sep-Nov}	Included
Traveler type	Extracted (11)	Review	Categorical	{Business, Couples, Families, Friends, Solo}	Included
Member registered year	Extracted (12)	User	Date (year)	Year the user has registered in TripAdvisor	Transformed
Pool	Extracted (13)	Hotel	Categorical	If the hotel has outside pool	Included
Gym	Extracted (14)	Hotel	Categorical	If the hotel has gym	Included
Tennis court	Extracted (15)	Hotel	Categorical	If the hotel has tennis court	Included
Spa	Extracted (16)	Hotel	Categorical	If the hotel has spa	Included
Casino	Extracted (17)	Hotel	Categorical	If the hotel has a casino inside	Included
Free internet	Extracted (18)	Hotel	Categorical	If the hotel provides free internet	Included
Hotel name	Extracted (19)	Hotel	Categorical	Hotel's name	Included
Hotel stars	Extracted (20)	Hotel	Categorical	Hotel's number of stars	Included
Nr. Rooms	Extracted (21)	Hotel	Numerical	Hotel's number of rooms	Included
User continent	Computed	User	Categorical	Continent where the user's country is located	Included
Member years	Computed	User	Numerical	Number of years the user is member of TripAdvisor	Included
Review month	Computed	Review	Categorical	Month when the review was written (from review date)	Included
Review weekday	Computed	Review	Categorical	Day of the week the review was written (from review date)	Included

272

273 After the data collection process, the dataset contained 504 records and 21 extracted features (as
274 of “origin=extracted”, from Table 1), 24 per hotel, regarding the year of 2015. However, such
275 dataset still needed to be prepared for serving as an input to the modeling stage. Since this data
276 was hand-collected and all the reviews chosen were complete, there were no missing values to be
277 dealt with. However, a closer look at the data allowed to identify a small set of features with few
278 to none value in terms of characterization of each of the reviews in the compiled dataset. These
279 features were excluded from the dataset and are marked accordingly in the column “status” in
280 Table 1. Such is the case for the review language, always in English for the collected reviews;
281 thus, the value remained the same for all the records, meaning it does not provide additional
282 information for characterizing the scores. In fact, most of the reviews found for the Strip’s hotels
283 are written in English (e.g., from the 8,878 reviews published on TripAdvisor since ever up to
284 July 31, 2016 for the “Encore at Wynn Las Vegas”, 7,951 of them are in English, almost 90% of
285 the total), an unsurprising result, given that Las Vegas is in the United States, a native English

286 country with a strong market of domestic tourism (Dawson, 2011) and also the worldwide
287 dissemination of the English language. For the case of the collected reviews, 217 of them are
288 from the United States, 72 from the UK, 65 from Canada, and 36 from Australia, in a total of 390
289 reviews from native English countries. The username was also excluded, as most of the reviews
290 were from different users (only six of the reviews were made by users from which a previous
291 review was also selected for the dataset). Finally, the textual content of the reviews was not
292 considered for modeling, since it is unstructured and additional techniques would need to be
293 employed, such as text mining. Furthermore, the focus of this research is on knowledge
294 extraction from quantitative features to overcome the limitations of textual reviews mentioned in
295 Section 2, such as the ambiguity of human language.

296 Another procedure that usually takes place in data mining is feature engineering, which is
297 considered a key step by Domingos (2012). Therefore, a few of the features were transformed
298 (Table 1, “status=transformed”) into new ones, which were computed (Table 1,
299 “origin=computed”). For example, the year when the user registered as a TripAdvisor member is
300 just an occurrence in time, whereas the number of years of membership represents how long the
301 user is active in TripAdvisor. Thus, the “member registered year” was transformed in “member
302 years”. The same happened for “review date”, from where “review month” and “review
303 weekday” were computed. Also, the country from where the reviewer is native was used to
304 obtain the corresponding continent, although in this case the “country” feature was kept, since it
305 may conceal meaningful value through user country’s characterization of the review score.

306 The result of these data collection and preparation procedures is a dataset with a total of 19 input
307 features plus the outcome to predict, the score given by users (Table 1 features with
308 status=“included”).

309

310 **3.2. Data mining**

311 According to Turban et al. (2008, p. 305), data mining is “the process that uses statistical,
312 mathematical, artificial intelligence and machine-learning techniques to extract and identify
313 useful information and subsequently gain knowledge from large databases”. Data mining usage
314 virtually spreads across any field of research from where data analysis is in demand. For

315 example, it is mostly used for companies in order to analyze customer data within the customer
316 relationship management (CRM) structure (Ngai et al., 2009). Due to its nature originated in
317 both statistical and machine learning fields, data mining focuses on the machine-driven model
318 building instead of hypothesis testing supervised by a specialized researcher (Magnini et al.,
319 2003). Furthermore, it was discussed by the same researchers that data mining techniques
320 discover patterns that can be used in order to strengthen the relationship between the hotel and
321 the frequent consumers, predicting the potential value of each customer and avoiding the cost of
322 attracting new ones. Also in hospitality, by clustering the customers (e.g., through traveler type)
323 it is possible for the company to know its target and therefore to be more efficient in satisfying
324 customer needs. It is also an important tool for the marketing department, since with this
325 information it is possible to previously create personalized advertisements or create direct-mail
326 campaigns (Magnini et al., 2003).

327 A data mining project usually consists in cycles of relevant consecutive stages such as data
328 understanding, preparation, modeling and evaluation (Moro et al., 2014). A few methodologies
329 have emerged for the definition of guidelines to conduct a data mining project, such as the
330 CRISP-DM (Moro et al., 2011). One of the most critical steps in data mining is data preparation
331 for modeling, which includes feature selection and feature engineering, i.e., choosing the
332 variables that best characterize the problem and, if needed, compute or obtain additional features
333 (Domingos, 2012; Moro et al., 2016a).

334 Although text mining is one of the most common techniques when analyzing online reviews, as
335 it establishes patterns that determine trends through textual comments (Lau et al., 2005), this
336 study focused on assessing the patterns hidden in the quantitative fields from TripAdvisor,
337 instead of the textual review itself. Thus, as the problem is to model the score (the outcome to
338 predict) granted by users through the remaining features (the inputs), it becomes a supervised
339 learning problem. Therefore, for modeling, the support vector machine was chosen, as it is one
340 of the most advanced supervised learning techniques, by transforming inputs into a high m-
341 dimensional feature space, using a nonlinear mapping. Consequently, the algorithm fits its way
342 to the best linear separating hyper plane, connected through the distributed set of support vector
343 points, which determines the support vector in the feature space, thus providing an accurate
344 performance (Moro et al., 2016b).

345 While the high level of accuracy of support vector machines makes of them attractive to use, the
346 inherent complexity makes them unreadable by a human user, as opposed to regression or
347 decision tree models (Cortez & Embrechts, 2013). For opening such types of “black-box”
348 models, from which neural networks are also an example, a few techniques can be used. Hence,
349 knowledge extraction from complex models can be achieved through rule extraction or
350 sensitivity analysis (Moro et al., 2014). The latter applies changes in the inputs through their
351 range of possible values and evaluates how it affects the predicted output value (Palmer et al.,
352 2006). Cortez and Embrechts (2013) further developed the sensitivity analysis method by
353 proposing a data-based sensitivity analysis (DSA) that takes advantage of the data used for
354 training the model to assess multiple variations of the input features, thus evaluating the
355 influence each feature exerts on the remaining ones, besides the impact on the outcome feature.
356 The DSA has been adopted with success for extracting knowledge from models in a wide variety
357 of studies such as wine modeling (Cortez et al., 2009), jet grouting (Tinoco et al., 2011) and bank
358 telemarketing (Moro et al., 2014), and it was therefore also chosen for the present study.

359 Considering the score available for users to rate hotels in TripAdvisor is an integer value
360 between 1 and 5, with 1 representing the lowest and 5 the highest scores respectively, the
361 problem becomes a regression problem (Sharda et al., 2017), where the model needs to fit the
362 input data for modeling the numerical outcome. Accordingly, two metrics were adopted for
363 computing model accuracy: the Mean Absolute Error (MAE) and the Mean Absolute Percentage
364 Error (MAPE). The MAE is the mean of all absolute differences between the real value and the
365 one predicted by the model, thus measuring how far the estimates are from actual values. The
366 MAPE metric is the mean of all absolute differences between the real value and the one
367 predicted by the model divided by the real score, in order to extract a percentage regarding each
368 deviation. Both metrics are described in detail by Hyndman and Koehler (2006). One of the
369 disadvantages of MAPE is that it becomes undetermined for outcome values near zero.
370 Nevertheless, such issue does not apply to the present study, since the outcome varies from 1 to
371 5.

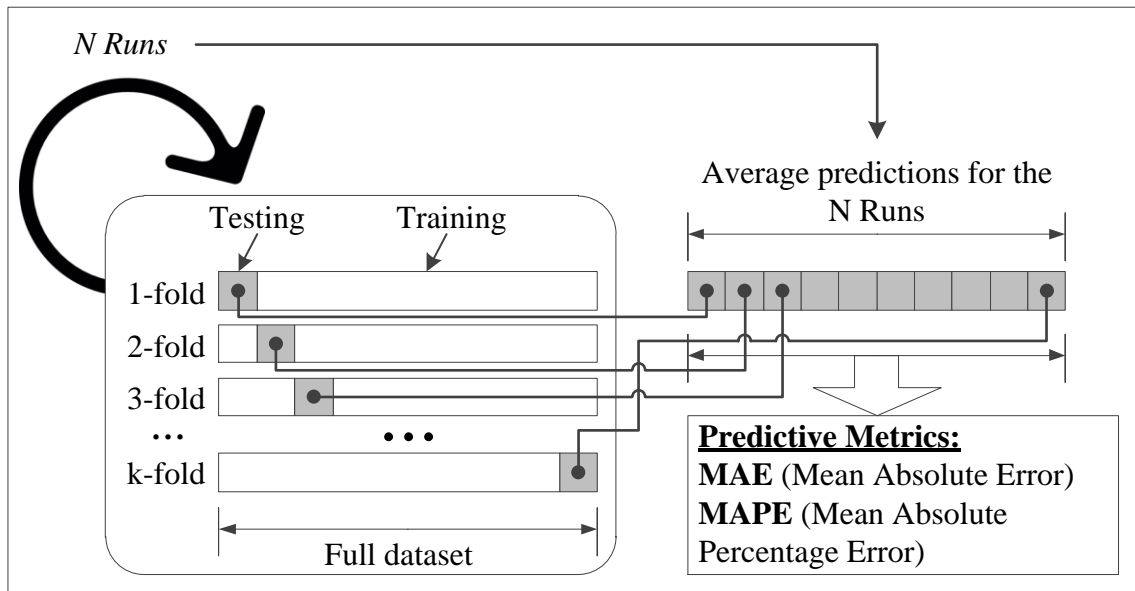
372

373

374

375 **3.3. Modeling and knowledge extraction**

376 With the dataset ready for modeling, a procedure took place to assess the robustness of the model
377 built on the data. Figure 5 shows a visual picture of such procedure. The evaluation of the model
378 was executed through a k-fold cross-validation technique where the whole dataset is divided into
379 k folds or sections grouping consecutive reviews from the dataset (Bengio & Grandvalet, 2004).
380 The k value was set to 10 (a value recommended by Refaeilzadeh et al., 2009), implying that
381 90% (454 reviews) of the data was used for training the model while the remaining 10% (50
382 reviews) for testing it, thus assuring independence of the split between training and test data. The
383 train-test execution was run 10 times, by varying the fold of data for testing model accuracy,
384 hence computing the predicted score once per record. Since the support vector machine
385 implements a non-linear complex model, to further assure model evaluation, the 10-fold cross-
386 validation was conducted 20 times, with the final score being computed by the average of the 20
387 executions. Performance modeling was then assessed by computing both MAE and MAPE
388 metrics for these averaged predicted results for each of the reviews in the dataset.

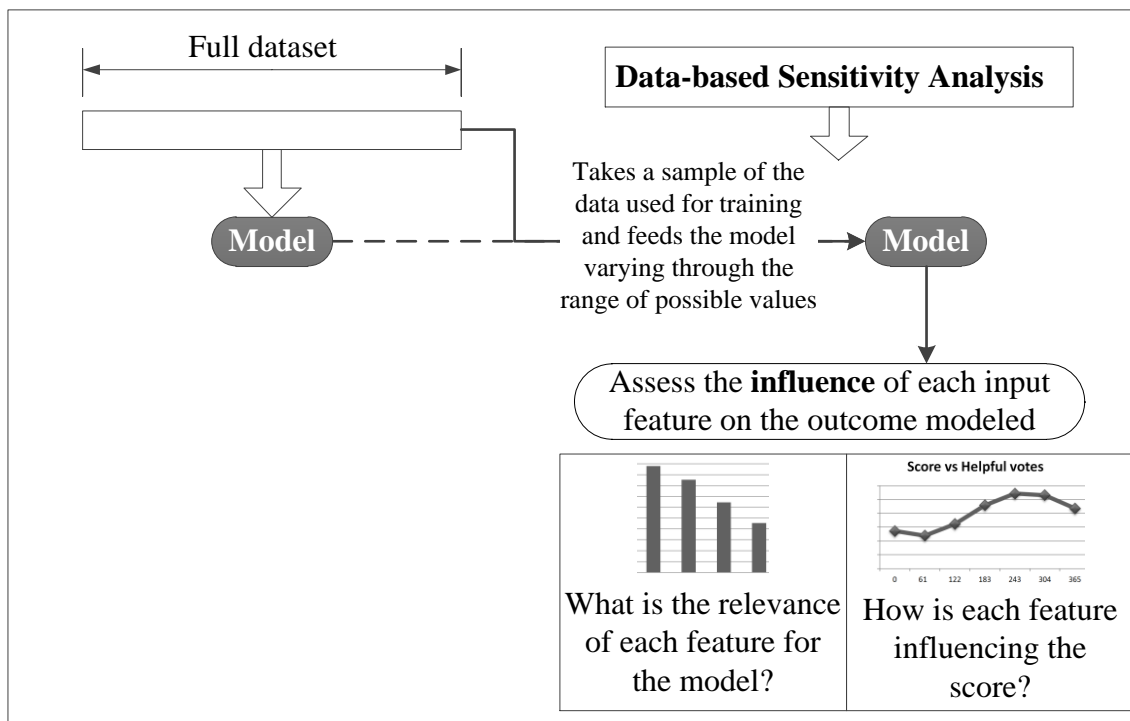


389

390 **Figure 5 - Modeling performance assessment.**

391 Assuming the input dataset prepared conceals relations between the input features and the score,
392 and that the chosen modeling technique (i.e., support vector machine) is able to unveil such
393 relations, the resulting computed predictive metrics would then comprehend satisfactory results
394 in terms of accuracy. Hence, a model built on the whole dataset using the same modeling

395 technique will also conceal such knowledge, enabling to extract it through the DSA. Figure 6
 396 shows the procedure undertaken for such knowledge extraction. First, a model is built on the
 397 whole dataset. Then, the model is used for exposing through DSA which are the features that
 398 influence most the score, translating such knowledge in terms of percentage relevance to which
 399 each feature contributes for modeling the score. Finally, using also DSA it is possible to observe
 400 how each of the most relevant features manages to influence the score.



401

402

Figure 6 - Knowledge extraction through sensitivity analysis.

403 To conduct all experiments, the R statistical tool was adopted (see: <https://cran.r-project.org/>). It
 404 provides a free and open source framework with multiple methods and functions to perform data
 405 analysis (James et al., 2013). Moreover, it has generated a worldwide enthusiasm translated in a
 406 vast community of contributors of a myriad of packages that can be freely downloaded and used
 407 for diverse purposes (Cortez, 2014). Specifically designed for data mining, by providing a simple
 408 and coherent set of functions, the “rminer” package was chosen (Cortez, 2010). Furthermore, this
 409 package also implements functions for extracting knowledge from models through sensitivity
 410 analysis, including the DSA.

411

412 **4. Results and discussion**

413 As described in Section 3 and illustrated in Figure 5, modeling performance was first assessed
 414 using an evaluation scheme including a realistic 10-fold cross-validation procedure to test the
 415 model with unforeseen data, which was ran twenty times. Table 2 shows the predictions for three
 416 randomly selected reviews with the data used as an input to the model (data is displayed
 417 vertically for space optimization purpose only). The predicted score is an average of the 20
 418 executions of the procedure, as described earlier in Section 3. The absolute deviation is the
 419 difference between the real and the predicted scores, with the MAE metric resulting from the
 420 average of all deviations for the 504 reviews. The percentage deviation corresponds to the
 421 relation between the absolute deviation and real score, with the MAPE metric being the
 422 computed average of all percentage deviations.

423 **Table 2** - Prediction results for three reviews.

Reviews	#1	#2	#3
User country	USA	USA	Ireland
User continent	America	America	Europe
Member years	2	1	3
Review month	February	October	April
Review weekday	Saturday	Friday	Friday
Nr. Reviews	36	23	19
Nr. Hotel reviews	9	17	9
Helpful votes	25	11	28
Traveler type	Families	Families	Couples
Period of stay	Mar-May	Sep-Nov	Mar-May
Hotel name	Circus Circus Hotel & Casino Las Vegas	Monte Carlo Resort&Casino	Tropicana Las Vegas - A Double Tree by Hilton Hotel
Hotel stars	3	4	4
Nr. Rooms	3,773	3,003	1,467
Free internet	YES	NO	YES
Pool	NO	YES	YES
Gym	YES	YES	YES
Tennis court	NO	NO	YES
Spa	NO	YES	YES
Casino	YES	YES	YES
Real score	5	3	5

Predicted score	3.9	3.6	4.6
Absolute deviation	1.1	0.6	0.4
% deviation	22.0%	20.0%	8.0%

424

425 The results for both metrics adopted, MAE and MAPE, can be seen on Table 3. In the scale from
 426 1 to 5 used for the score on TripAdvisor, the support vector machine achieved an average
 427 absolute deviation of 0.745, an indicator that it presents a predicted value close to the real score,
 428 by less than one. MAPE translates such deviation into a percentage: the average predicted score
 429 deviates by 27.32% from the real score. While such results show the model is not totally accurate
 430 for every review (as it can be seen from the three cases illustrated in Table 2), these also provide
 431 proof that the model constitutes a valid approximation for modeling TripAdvisor score.
 432 Furthermore, other studies have discovered valid insightful knowledge from a model with a
 433 MAPE of around 27% (e.g., Moro et al., 2016b).

434

Table 3 - Modeling performance assessment metrics.

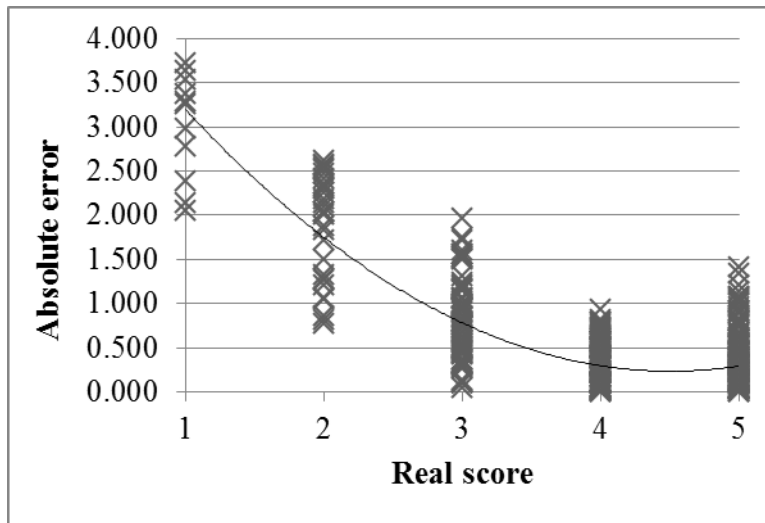
Metric	Result
MAE	0.745
MAPE	27.32%

435

436 The knowledge discovery phase aims to provide the major contribution of this research, as it
 437 lends insights on the characterization of review scores of such a renowned location as it is the
 438 case of Las Vegas Strip, while keeping in mind the relevance widely discussed in the literature of
 439 online customers' feedback to the hospitality industry (e.g., Ye et al., 2009a). Thus,
 440 understanding what drives users to publish a given score can ultimately leverage managerial
 441 decision support in hospitality. Therefore, the understanding of the factors that influence why a
 442 given hotel is being rated with a certain score can be valuable for managers to act on parameters
 443 they control (e.g., hotel related features) and to preventively manage their units according to the
 444 expected tourists' demands (e.g., by knowing the more demanding tourists).

445 Figure 7 displays the relation between the absolute error and the real score. The model performs
 446 better when predicting higher scores, while lower scores, since are less represented, tend to result

447 in higher errors. However, such a poor prediction performance points out to a limitation as bias
448 occurs in the model, resulting in underpredicting low ratings and overpredicting high ratings.

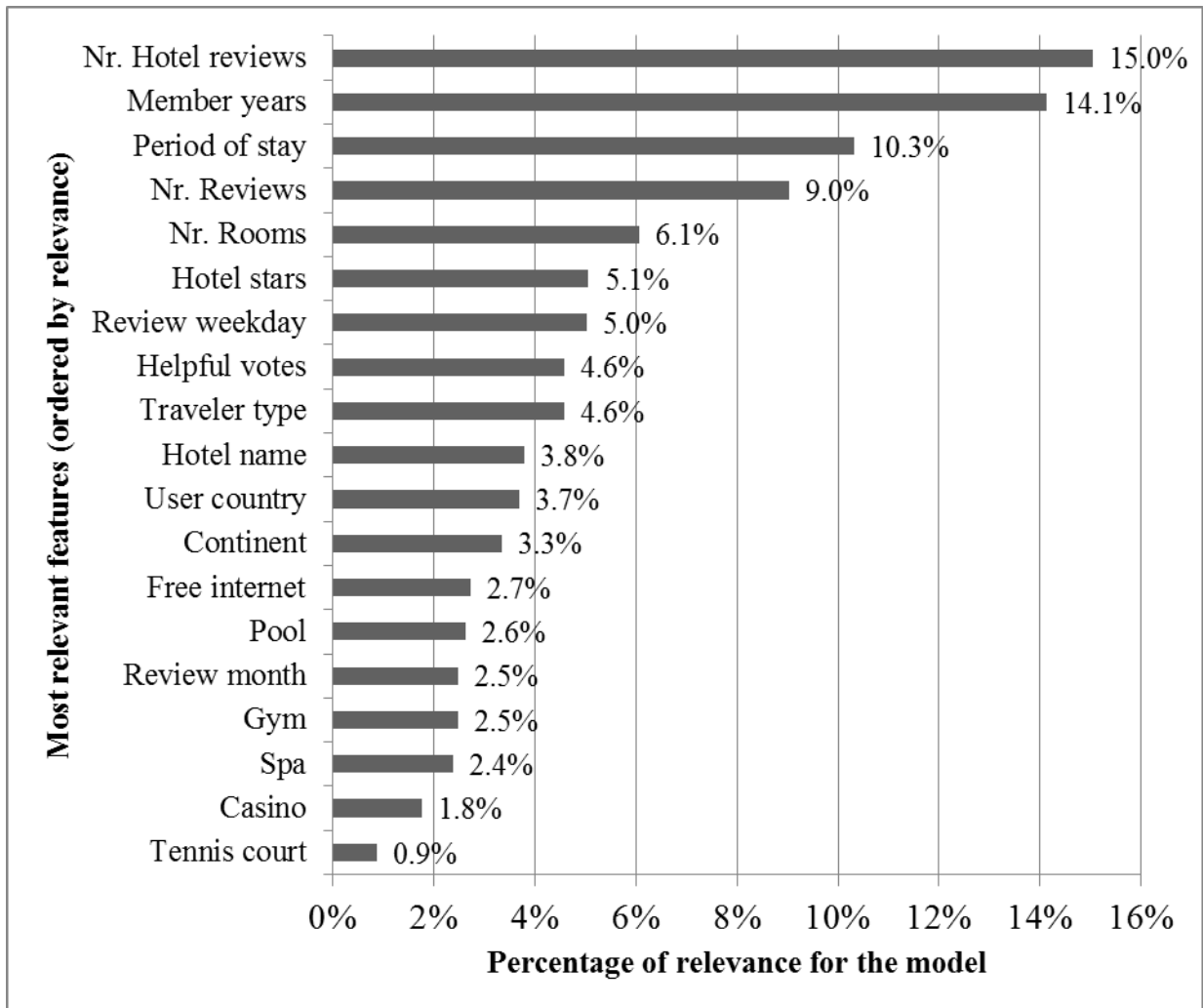


449

Figure 7 - Scatterplot of real scores versus absolute error

450

451 As stated previously, the method chosen for knowledge extraction was the DSA. It provides
452 means of presenting for each feature the percentage of relevance that the feature has on the
453 model by analyzing outcome fluctuation to input features' variation. Sensitivity analysis requires
454 a single model, which was built using the whole dataset, as shown in Figure 6. Figure 8 exhibits
455 the percentage relevance computed through DSA for all the features. Considering DSA's
456 computation is based on a random sample selection, the procedure encompassed twenty
457 executions, and the relevance computation of each individual feature showed is the resulting
458 average of the executions, hence strengthening confidence in the achieved results. The seven
459 most relevant, with an individual relevance above 5% each, conceal around 65% of relevance of
460 the model, and will be analyzed further ahead.



461

462

Figure 8 - Most relevant features according to their relevance.

463

464

465

466

467

468

469

470

471

472

473

The two most relevant features are both related to the user. The number of reviews of hotels that the user has made contributes with an influence to the final score greater than any of the remaining features, with 15% of relevance. A similar result occurs for the membership years that the user has since first registered in TripAdvisor, with a relevance of 14.1%. In fact, the fourth most relevant feature is the number of reviews, which is closely related to the most relevant feature (“nr. hotel reviews”), as it includes all the reviews, together with the restaurant and attraction units summing up to hotels’ reviews. These three features hold almost 40% of model relevance when modeling the score. This is an interesting discovery, suggesting the score is clearly biased by the users’ experience acquired over time, influencing self-awareness of what is a fair rate. Hence, managers should have this into account when considering the score their units are having on TripAdvisor. Namely, they can optimize answering reviews by framing template

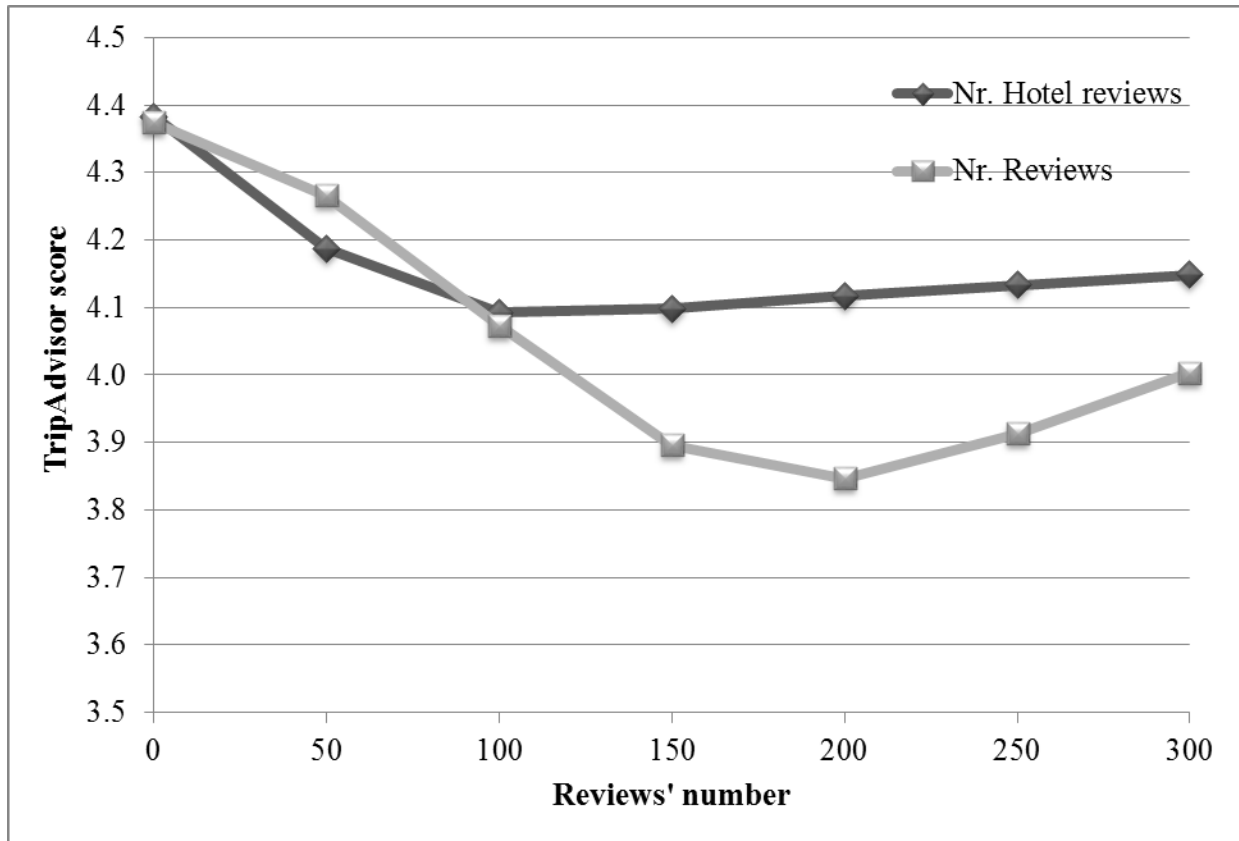
474 responses according to users' features. This is an important contribution, as online reviews
475 usually accumulate without managers being able to deal with such high volumes of reviews.

476 The period of stay is the third most relevant feature, with 10.3% of influence when compared to
477 the remaining features. Such result was expected, given the seasonality effect known of tourism
478 and hospitality (Song & Li, 2008). Surprisingly, the most relevant hotel features only appear in
479 fifth and sixth places, the number of rooms and stars, respectively. Moreover, previous studies
480 concluded that the number of stars affects online booking (e.g., Ye et al., 2011). Also worth of
481 note is the fact that the weekday the user has published the review plays 5% of the role when it
482 comes to modeling TripAdvisor score. The remaining features are all below 5% in terms of
483 relevance, including hotel name and user country. It was expected that the brand name and image
484 behind the hotel contributed more to user rating, as it is suggested by previous research on hotel
485 brand influence (e.g., Sparks & Browning, 2011). Also worth of noticing is the fact that the
486 features that can be entirely controlled by the hotel, such as the amenities (e.g., free internet,
487 pool, gym, spa, casino and tennis court) are influencing less than 3% each.

488 Considering the location-based nature of this empirical research, the results hereby presented
489 must be discussed in the light of Las Vegas importance in hospitality and tourism. Las Vegas is a
490 top tourism destination in the United States, which reflects into the high number of reviews in
491 TripAdvisor. As an example, O'Mahony and Smyth (2010) found 146,409 published reviews by
492 32,002 users prior to April 2009 for Las Vegas, whereas the same study found around half of
493 reviews for Chicago in the same period, a much larger city. These figures reveal that Las Vegas
494 is a very mature tourism market, with its tourists being fully aware of online reviews, whether by
495 publishing new reviews or for obtaining feedback. The more recent study by Rosman and
496 Stuhura (2013) emphasizes the immediacy of online feedback in Las Vegas. In addition, it is
497 known the effect of self-congruity on tourism destinations and, particularly, on Las Vegas
498 tourists (Usakli & Baloglu, 2011). Therefore, experienced tourists translated in a higher degree
499 of TripAdvisor membership may unconsciously be influenced by such experience when
500 providing feedback in such a mature market as Las Vegas. Furthermore, the Las Vegas brand
501 itself is able to generate controversial feelings capable of affecting tourists' perception
502 (Griskevicius et al., 2009). All these characteristics are aligned with the model built on

503 TripAdvisor's review features, with experience counting as the top influencing factor, while
504 hotel brand having a significant lower relevance.

505 After analyzing the relevance of features on TripAdvisor score, it is interesting to dive deeper
506 into each of the most relevant ones (with relevance above 3.5%, as identified in Figure 8) in an
507 attempt to understand how these features affect the score. Both the most relevant ("nr. Hotel
508 reviews") and the fourth most relevant ("nr. Reviews") features overlap in the sense that the
509 latter includes the former, plus the reviews the user has made on attraction units and restaurants.
510 Therefore, these two features are analyzed together. Figure 9 shows how each influence the
511 score. As expected (Magnini et al., 2003), the experience momentum after the initial first reviews
512 tend to turn the customer more demanding when publishing online score. Nevertheless, such
513 effect is more profound for the global counter of reviews, including attraction units and
514 restaurants. This finding is aligned with previous study by McCartney (2008), which stated that
515 gaming and casino attractions leverage tourists' requirements in terms of hospitality. Hence,
516 global reviews may have the effect of plunging scores to values below 3.9.

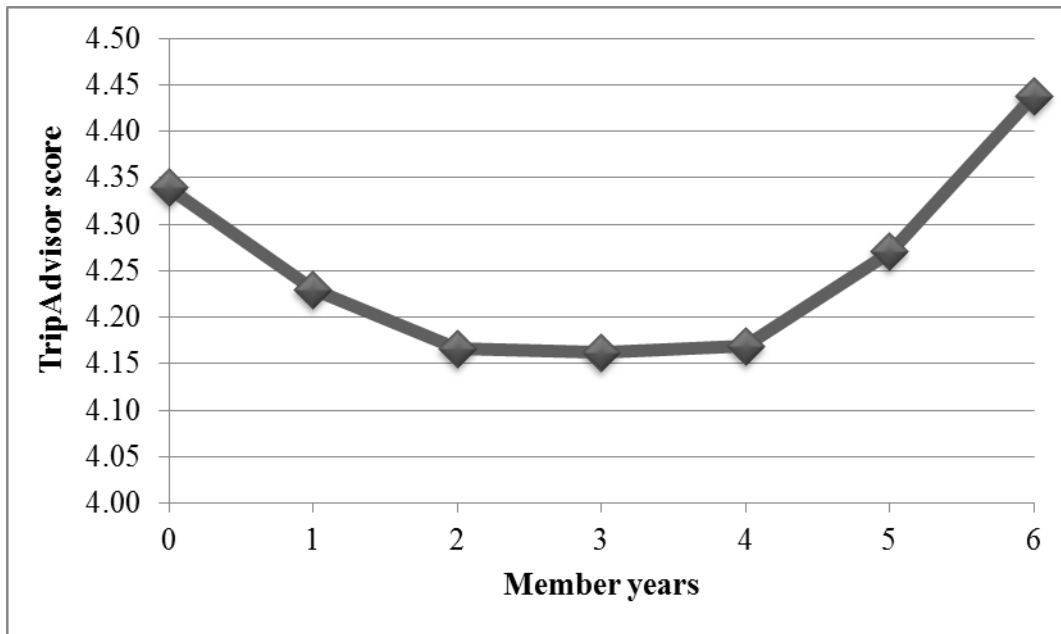


517

518 **Figure 9** - Influence of "Nr. Hotel reviews" and "Nr. Reviews" on TripAdvisor score.

519

520 Figure 10 displays the effect of the number of years as a TripAdvisor member on the given score.
 521 Up to four years of membership, the conclusions are similar to the number of reviews made;
 522 however, users registered five years ago or more tend to be more positive by granting better
 523 review scores. While for the number of reviews, it can also be observed on Figure 9 a slight
 524 increase on the score after a certain threshold (this is particularly visible on the "nr. Reviews"
 525 feature), the results for "member years" clearly amplify such tendency, with older members
 526 giving scores above new members. Some hypotheses can be raised based on this result. One of
 527 the most plausible is that tourists with more experience have better knowledge on the destination
 528 and units available, thus they will choose the hotels that please them the most, resulting in higher
 529 scores. Also, experienced TripAdvisor members are probably keener to read other members'
 530 reviews and so be better informed to make judged decisions on their own stays (Liu et al., 2015).
 531 Nevertheless, more data would be needed to confirm or reject such hypotheses.

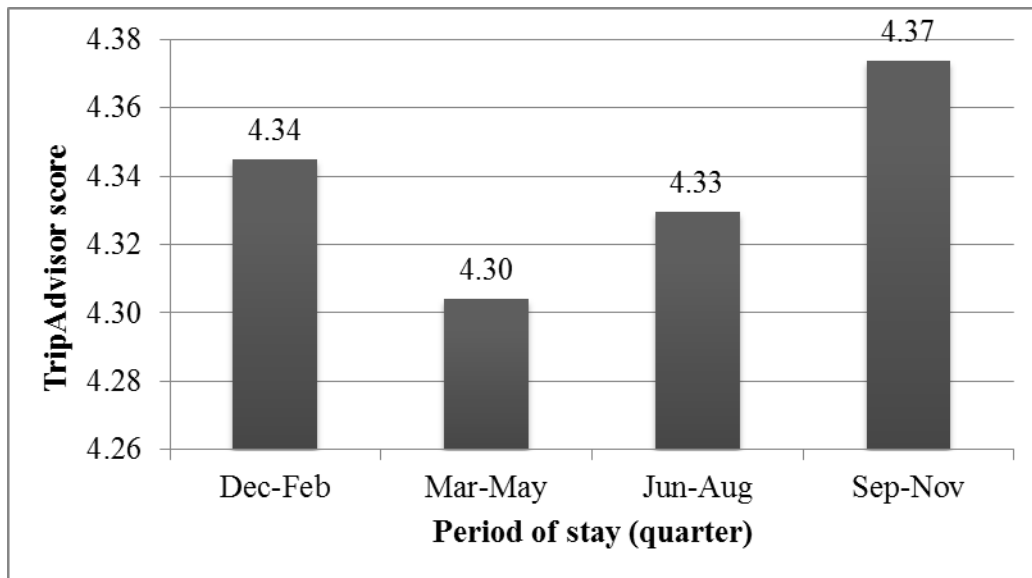


532

Figure 10 - Influence of "Member years" on TripAdvisor score.

533

534 The third most relevant feature for modeling score was the period of stay, in quarter fractions of
 535 a year. Figure 11 shows the seasonality effect on TripAdvisor score. Several previous studies are
 536 found concluding that Las Vegas holds a seasonality effect on its tourism (e.g., Yang & Gu,
 537 2012; Day et al., 2013). Considering Las Vegas is located in a hot desert, the colder months of
 538 autumn and winter tend to attract more tourists. Although the visible effect on the bar plot is very
 539 small, with Sep-Nov reaching the peak of 4.37 of score, while Mar-May bottoms at 4.30, by
 540 holding relevance above 10% for the model implicates its variation although small does affect
 541 TripAdvisor score and probably such influence gets confounded in aggregation with the
 542 remaining features.



543

544

Figure 11 - Influence of "Period of stay" on TripAdvisor score.

545

546

547

548

549

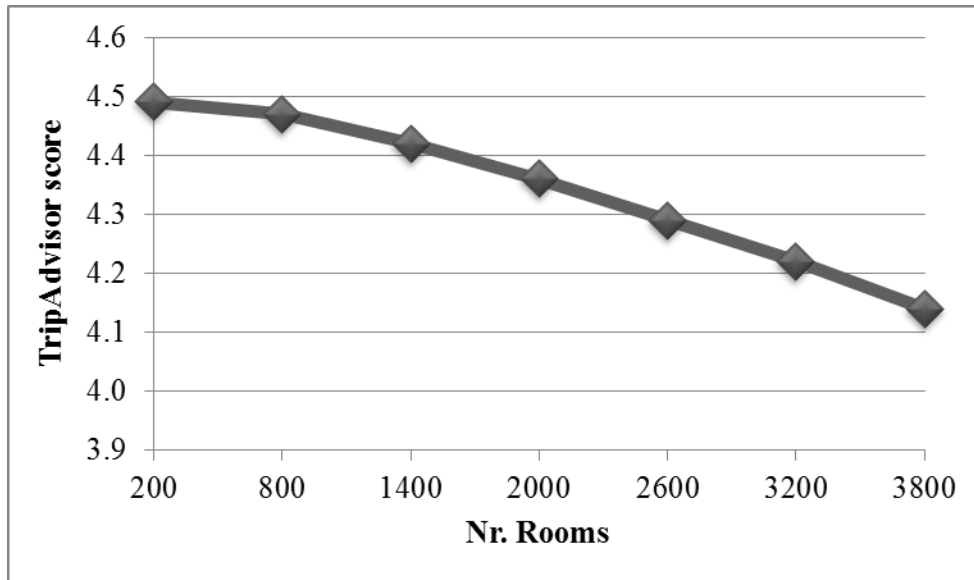
550

551

552

553

The number of rooms the hotel unit has is the fifth most relevant feature, although with a contribution of just 6.1% pales in comparison with the top four, all above 9% of relevance. Still, it is the most relevant feature in respect to hotel specifications. Figure 12 shows that smaller units tend to have better review scores. This effect is significant, with the average difference score between an hotel with 200 rooms and another with 3,800 reaching 0.4 points. The recent study by Jiménez et al. (2016) based on Spain and Portugal hotel units also found a similar relation: as the number of rooms increases, the TripAdvisor score decreases. Hotels smaller tend to offer a friendlier and non-crowd environment which may be promoted as an advantage against large resorts, suiting better tourists enjoying quiet stays inside the unit (Chambers, 2010).

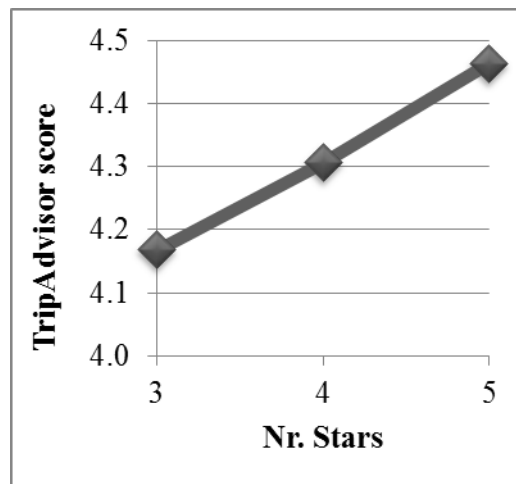


554

Figure 12 - Influence of "Nr. Rooms" on TripAdvisor score.

555

556 Figure 13 displays the effect of the number of stars of the hotel on TripAdvisor score. The result
 557 is expected: the higher the number of stars, the higher the score. Las Vegas Strip hotels' range
 558 from three to five stars. Hu and Chen's (2016) study is aligned with the findings unveiled from
 559 Figure 13 in that hotel stars influence positively reviews' ratings.



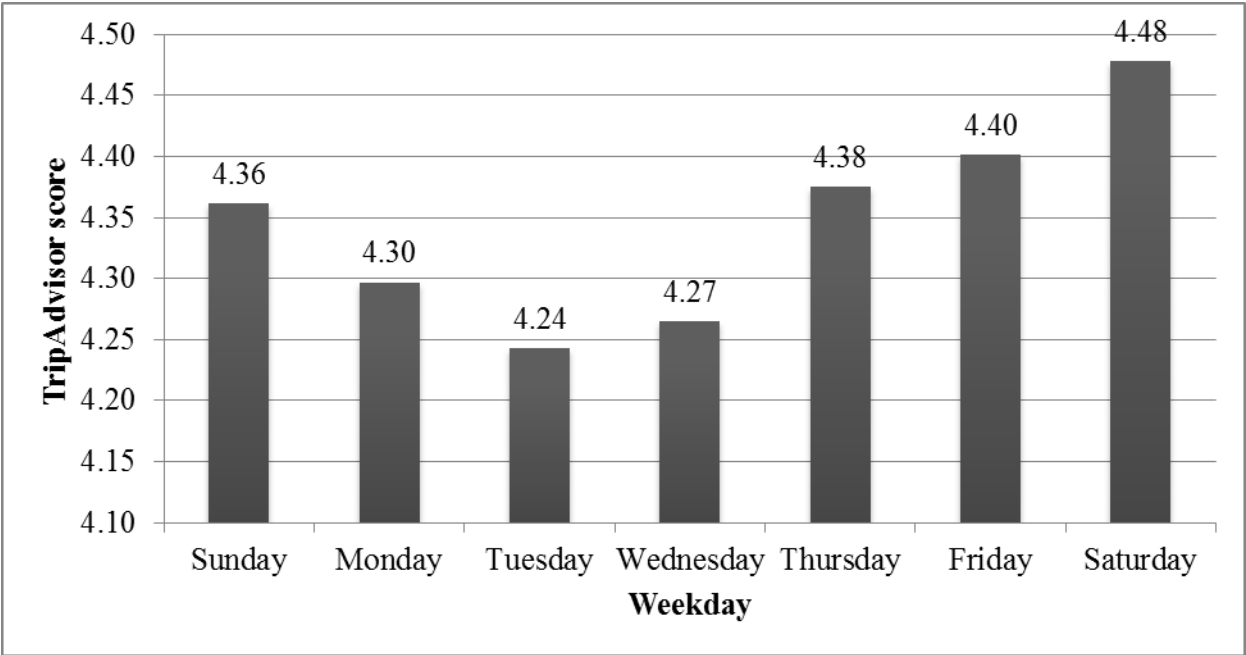
560

Figure 13 - Influence of "Nr. Stars" on TripAdvisor score.

561

562 The seventh most relevant feature is a surprise: the weekday when the review was published
 563 achieved a relevance of 5% (Figure 8). From Figure 14 it is possible to observe that the weekday
 564 influences directly TripAdvisor score in a range of 0.24 points (from 4.24 on Tuesday to 4.48 on
 565 Saturday). The effect of seasonality is known in tourism, but the finding related to the influence

566 of the weekday's of publication has no precedent in tourism. Furthermore, user feedback may
567 vary a lot in terms of lag related to the period of stay, as some tourists provide feedback directly
568 on sight, while others wait some days before writing the review. Nevertheless, other studies on
569 social media have also found an influence of the weekday of publication on the impact of
570 publishing contents, such as the finding by Moro et al. (2016b) on a company's Facebook posts.
571 Seemingly reviews published near the weekend tend to receive better scores, as shown in Figure
572 14. The ending of a week, with a restful weekend nearby and, particularly, Saturday, the first
573 weekend day, are known to have a positive psychologically effect on people, and are also playing
574 a role in granting scores on TripAdvisor (Ryan et al., 2010).



575

576 **Figure 14** - Influence of "Weekday" on TripAdvisor score.

577

578 Other features contributing with a relevance below 5% including "helpful votes", "traveler type",
579 "hotel name" and "user country" are not scrutinized in this paper. Nevertheless, each of them
580 plays a role on the built model, although with a less relevant role in comparison with the top
581 influencing features.

582

583

584 **5. Conclusions**

585 It is currently unquestionable that online feedback reviews in tourism have the power to
586 influence to a certain degree forthcoming tourists. Hence, hospitality unit managers have recently
587 included such source of information in their decision making processes. TripAdvisor is the
588 largest online platform for providing feedback on tourism and hospitality and one of the main
589 sources for managers to control customer feedback.

590 A TripAdvisor member has mainly two means for providing feedback: a free text area for input
591 of textual comments; and a quantitative score between 1 and 5. The textual comments, by
592 concealing interesting user sentiments, have been widely studied in the literature. However,
593 knowledge extraction based on such comments is usually harder to achieve when compared to
594 the quantitative score. Furthermore, the inherent subjectivity associated with human language
595 poses difficult challenges to overcome. On the opposite side, the quantitative score is an
596 objective measure, easier to model. Still, research on the score is rather scarce in comparison to
597 research on textual reviews. Hence, the knowledge extraction procedure presented in this paper
598 is based on modeling TripAdvisor score. The present study aimed at: (1) unveiling how each of
599 the features used to feed the model affects the score granted, and (2) understanding the specific
600 effect of the individual features on the score.

601 The empirical research presented in this paper focused in the mature Las Vegas Strip hospitality
602 market linked to gaming and pleasure industries, translated in a high number of reviews on
603 TripAdvisor for each of its 21 hotel units. This location-based study benefits from a controlled
604 environment as external factors that may subtly affect customer satisfaction (such as location,
605 local tourist attractions) are identical or very similar (and hence practically controlled for). Such
606 advantage ends up providing a clearer picture about the remaining dimensions encompassed in
607 the built model, namely: (1) user membership in TripAdvisor; (2) hotel characteristics; (2) and
608 reviews details.

609 Several contributions rise from this study. First, a TripAdvisor score model was built with an
610 acceptable MAE of 0.745 and a MAPE of 27%, assuring the deviation from the score predicted
611 and the real value constituted an interesting approximation as a predictive model. Such
612 achievement was possible by using an advanced data mining technique, support vector machine,

613 fed through 19 features encompassing three variable dimensions, user membership, hotel and
614 review features, while keeping the location fixed. This is an interesting finding, as it differs from
615 current literature offering correlation analysis between pairs or small sets of features, instead of
616 the proposed single model built on a larger number of features. Such model can then be used as a
617 baseline for extracting knowledge through the data-based sensitivity analysis translated into
618 individual relevance of features, i.e., on how each of them contributes to explain the scores
619 granted on TripAdvisor.

620 The second set of contributions is unveiled through extracting knowledge from the model and
621 implies managerial considerations when encompassing TripAdvisor data in hospitality analysis.
622 The major findings include (1) the magnitude of the effect of the personal characteristics of the
623 reviewers, (2) the nonlinear relationship between the reviewer's activity on TripAdvisor (which
624 may be regarded as a proxy for travel experience) and the valence of the reviewer's rating scores,
625 and (3) the seasonal and day of the week effect observed. The remaining results obtained are
626 consistent with the findings of previous related studies. The relevance discovered related to
627 TripAdvisor membership experience may lead to managerial guidelines for supporting the
628 process of answering online reviews. Two types of application of such knowledge are possible. If
629 the hotel holds a small team to answer reviews piling in comparison to a vast number of reviews
630 in TripAdvisor, then the hotel may implement a selection procedure for choosing the most
631 suitable user profiles to direct efforts in answering those, aligned with the hotel strategy.
632 Moreover, hotel managers can optimize answering reviews by framing template responses
633 according to users' profiles, leading to an efficiency improvement by directing efforts of team
634 members. In alignment with the same recommendation, efforts in answering online reviews may
635 be redirected to answering the more negative reviews during the middle of the week, considering
636 the observed influence of such feature. However, additional studies would need to be conducted
637 in order to adjust such proposed reviews' answering strategies.

638 It should be noted that, by being a location-based study, users' awareness of Las Vegas brand
639 itself must be an accountable factor on influencing score. Furthermore, such renowned brand is
640 able to generate controversial feelings capable of affecting tourists' perception. This fact may
641 also play a role on the lower ranked hotel features in terms of relevance when compared to user
642 characteristics. As Magnini et al. (2003) discussed, customer satisfaction may bias a data mining

643 approach in tourism due to the relative importance each user attributes to certain characteristics.
644 The present study sheds additional light by concluding that experience as a TripAdvisor member
645 does affect the score rank given by users. However, the present study is focused solely on
646 reviews for hotels in Las Vegas Strip, thus its conclusions have to remain location-based.
647 Furthermore, the relative importance of user versus hotel features can be affected by the specific
648 Las Vegas context, as it is known from previous studies that hotel location influences scores
649 granted. Thus, additional research is in demand to confirm or refute the possible generalization
650 of TripAdvisor experience influence on score. Furthermore, future research may include
651 studying different locations, with different characteristics. Also, more features from other
652 sources may be included in the model, considering the capability of support vector machines for
653 disentangling relationships between a wide number of different features. Additionally, future
654 research should focus on reducing model bias, aiming at tuning the model for improving
655 prediction performance.

656

657 **References**

- 658 Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. New York: Springer Science & Business
659 Media.
- 660 Ampofo, L., Collister, S., O'Loughlin, B., & Chadwick, A. (2015). Text mining and social
661 media: When quantitative meets qualitative and software meets people. In P. Halfpenny & R.
662 Procter (Eds.), *Innovations in Digital Research Methods* (pp. 161-192). Los Angeles: SAGE.
- 663 Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-
664 validation. *Journal of Machine Learning Research*, 5, 1089-1105.
- 665 Buccafurri, F., Lax, G., Nicolazzo, S., & Nocera, A. (2014). Fortifying TripAdvisor against
666 reputation-system attacks. In C. A. Shoniregun & G. A. Akmayeva (Eds.), *Proceedings of World
667 Congress on Internet Security (WorldCIS-2014)*. Paper presented at the 2014 World Congress on
668 Internet Security, London, UK (pp. 20-21). New York: IEEE.
- 669 Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment Classification of Consumer Generated
670 Online Reviews Using Topic Modeling. *Journal of Hospitality Marketing & Management*.
671 Advance online publication. doi:10.1080/19368623.2017.1310075.
- 672 Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the “helpfulness” of
673 online user reviews: A text mining approach. *Decision Support Systems*, 50(2), 511-521.
- 674 Casalo, L. V., Flavian, C., Guinaliu, M., & Ekinci, Y. (2015). Do online hotel rating schemes
675 influence booking behaviors?. *International Journal of Hospitality Management*, 49, 28-36.
- 676 Chambers, L. (2010). *Destination competitiveness: An Analysis of the characteristics to
677 differentiate all-inclusive hotels & island destinations in the Caribbean* (Thesis, Rochester
678 Institute of Technology, Rochester, USA). Retrieved from
679 <http://scholarworks.rit.edu/theses/471/>.
- 680 Corbitt, B. J., Thanasankit, T., & Yi, H. (2003). Trust and e-commerce: a study of consumer
681 perceptions. *Electronic Commerce Research and Applications*, 2(3), 203-215.
- 682 Cortez, P. (2010). Data mining with neural networks and support vector machines using the
683 R/rminer tool. In P. Perner (Ed.), *Advances in Data Mining - Applications and Theoretical*

684 *Aspects*. Paper presented at the 2010 Industrial Conference on Data Mining, Lecture Notes in
685 Artificial Intelligence 6171, Berlin, Germany (pp. 572-583). Berlin: Springer Berlin Heidelberg.

686 Cortez, P. (2014). *Modern optimization with R*. New York: Springer.

687 Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences
688 by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547-553.

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

691 Dawson, M. (2011). 'Travel Strengthens America'? Tourism promotion in the United States
692 during the Second World War. *Journal of Tourism History*, 3(3), 217-236.

693 Day, J., Chin, N., Sydnor, S., & Cherkauer, K. (2013). Weather, climate, and tourism
694 performance: A quantitative analysis. *Tourism Management Perspectives*, 5, 51-56.

695 Domingos, P. (2012). A few useful things to know about machine learning. *Communications of*
696 *the ACM*, 55(10), 78-87.

697 Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online
698 tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management*, 52,
699 498-506.

700 Filieri, R., Algezau, S., & McLeay, F. (2015). Why do travelers trust TripAdvisor?
701 Antecedents of trust towards consumer-generated media and its influence on recommendation
702 adoption and word of mouth. *Tourism Management*, 51, 174-185.

703 Griskevicius, V., Goldstein, N. J., Mortensen, C. R., Sundie, J. M., Cialdini, R. B., & Kenrick, D.
704 T. (2009). Fear and loving in Las Vegas: Evolution, emotion, and persuasion. *Journal of*
705 *Marketing Research*, 46(3), 384-395.

706 Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews:
707 Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, 59, 467-483.

708 He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case
709 study in the pizza industry. *International Journal of Information Management*, 33(3), 464-472.

710 Henning-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic
711 word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate
712 themselves on the Internet? *Journal of Interactive Marketing*, 18(1), 38-52.

713 Hu, Y. H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review
714 visibility, and interaction between hotel stars and review ratings. *International Journal of*
715 *Information Management*, 36(6), 929-944.

716 Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy.
717 *International Journal of Forecasting*, 22(4), 679-688.

718 James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*
719 *(Vol. 6)*. New York: Springer.

720 Jeong, M., & Jeon, M. M. (2008). Customer reviews of hotel experiences through consumer
721 generated media (CGM). *Journal of Hospitality & Leisure Marketing*, 17(1-2), 121-138.

722 Jiménez, S. M., Morales, A. F., de Sandoval, J. L. X., & Stefaniak, A. C. (2016). Hotel
723 assessment through social media–TripAdvisor as a case study. *Tourism & Management Studies*,
724 12(1), 15-24.

725 Kim, W. G., Kim, W. G., Park, S. A., & Park, S. A. (2017). Social media review rating versus
726 traditional customer satisfaction: Which one has more incremental predictive power in
727 explaining hotel performance?. *International Journal of Contemporary Hospitality Management*,
728 29(2), 784-802.

729 Kwok, L., Xie, K., & Tori, R. (2017). Thematic framework of online review research: A
730 systematic analysis of contemporary literature on seven major hospitality and tourism journals.
731 *International Journal of Contemporary Hospitality Management*, 29(1), 307-354.

732 Lau, K. N., Lee, K. H., & Ho, Y. (2005). Text mining for the hotel industry. *Cornell Hotel and*
733 *Restaurant Administration Quarterly*, 46(3), 344-362.

734 Law, R., Buhalis, D., & Cobanoglu, C. (2014). Progress on information and communication
735 technologies in hospitality and tourism. *International Journal of Contemporary Hospitality*
736 *Management*, 26(5), 727-750.

737 Lee, K. (2015). *Transforming for the Future: The New Economic Driver for the Las Vegas*
738 *Tourism Industry* (Thesis, University of Nevada, Las Vegas, United States). Retrieved from
739 <http://digitalscholarship.unlv.edu/thesesdissertations/2611/>.

740 Liburd, J. J. (2012). Tourism research 2.0. *Annals of Tourism Research*, 39(2), 883-907.

741 Liu, Z., Le Calvé, A., Cretton, F., Balet, N. G., Sokhn, M., & Délétroz, N. (2015). Linked Data
742 Based Framework for Tourism Decision Support System: Case Study of Chinese Tourists in
743 Switzerland. *Journal of Computer and Communications*, 3(05), 118-126.

744 Mackun, P. J., Wilson, S., Fischetti, T. R., & Goworowska, J. (2011). *Population distribution*
745 *and change: 2000 to 2010*. US Department of Commerce, Economics and Statistics
746 Administration, US Census Bureau. Retrieved from
747 <https://www.census.gov/prod/cen2010/briefs/c2010br-01.pdf>.

748 Magnini, V. P., Honeycutt Jr, E. D., & Hodge, S. K. (2003). Data mining for hotel firms: Use
749 and limitations. *Cornell Hospitality Quarterly*, 44(2), 94-105.

750 Mauri, A. G., & Minazzi, R. (2013). Web reviews influence on expectations and purchasing
751 intentions of hotel potential customers. *International Journal of Hospitality Management*, 34, 99-
752 107.

753 McCartney, G. (2008). The CAT (casino tourism) and the MICE (meetings, incentives,
754 conventions, exhibitions): Key development considerations for the convention and exhibition
755 industry in Macao. *Journal of Convention & Event Tourism*, 9(4), 293-308.

756 Min, H., Min, H., & Emam, A. (2002). A data mining approach to developing the profiles of
757 hotel customers. *International Journal of Contemporary Hospitality Management*, 14(6), 274-
758 285.

759 Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank
760 telemarketing. *Decision Support Systems*, 62, 22-31.

761 Moro, S., Cortez, P., & Rita, P. (2016a). A framework for increasing the value of predictive data-
762 driven models by enriching problem domain characterization with novel features. *Neural*
763 *Computing and Applications*. Advance online publication. doi:10.1007/s00521-015-2157-8.

764 Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An
765 application of the crisp-dm methodology. In P. Novais et al. (Eds.), *Proceedings of European*
766 *Simulation and Modelling Conference (ESM'2011)*. Paper presented at the 2011 European
767 Simulation and Modelling Conference, Guimarães, Portugal (pp. 117-121). Ostend: Eurosis.

768 Moro, S., Rita, P., & Vala, B. (2016b). Predicting social media performance metrics and
769 evaluation of the impact on brand building: A data mining approach. *Journal of Business*
770 *Research*, 69(9), 3341-3351.

771 Moro, S., & Rita, P. (2016). Forecasting tomorrow's tourist. *Worldwide Hospitality and*
772 *Tourism Themes*, 8(6), 643-653.

773 Neirotti, P., Raguseo, E., & Paolucci, E. (2016). Are customers' reviews creating value in the
774 hospitality industry? Exploring the moderating effects of market positioning. *International*
775 *Journal of Information Management*, 36(6), 1133-1143.

776 Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer
777 relationship management: A literature review and classification. *Expert Systems with*
778 *Applications*, 36(2), 2592-2602.

779 Nguyen, K. A., & Coudounaris, D. N. (2015). The mechanism of online review management: A
780 qualitative study. *Tourism Management Perspectives*, 16, 163-175.

781 O'Connor, P. (2010). Managing a hotel's image on TripAdvisor. *Journal of Hospitality*
782 *Marketing & Management*, 19(7), 754-772.

783 O'Mahony, M. P., & Smyth, B. (2010). A classification-based review recommender. *Knowledge-*
784 *Based Systems*, 23(4), 323-329.

785 O'Reilly, T., & Battelle, J. (2009). *Web squared: Web 2.0 five years on*. O'Reilly Media, Inc.

786 Qazi, A., Syed, K. B. S., Raj, R. G., Cambria, E., Tahir, M., & Alghazzawi, D. (2016). A
787 concept-level approach to the analysis of online review helpfulness. *Computers in Human*
788 *Behavior*, 58, 75-81.

- 789 Palmer, A., Montaña, J. J., & Sesé, A. (2006). Designing an artificial neural network for
790 forecasting tourism time series. *Tourism Management*, 27(5), 781-790.
- 791 Papathanassis, A., & Knolle, F. (2011). Exploring the adoption and processing of online holiday
792 reviews: A grounded theory approach. *Tourism Management*, 32(2), 215-224.
- 793 Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. *Annals of*
794 *Tourism Research*, 50, 67-83.
- 795 Phillips, P., Zigan, K., Silva, M. M. S., & Schegg, R. (2015). The interactive effects of online
796 reviews on the determinants of Swiss hotel performance: A neural network analysis. *Tourism*
797 *Management*, 50, 130-141.
- 798 Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In L. Liu & M. T. Özsu (Eds.)
799 *Encyclopedia of database systems* (pp. 532-538). USA: Springer.
- 800 Ro, H., Lee, S., & Mattila, A. S. (2013). An affective image positioning of Las Vegas hotels.
801 *Journal of Quality Assurance in Hospitality & Tourism*, 14(3), 201-217.
- 802 Rosman, R., & Stuhura, K. (2013). The implications of social media on customer relationship
803 management and the hospitality industry. *Journal of Management Policy and Practice*, 14(3),
804 18-26.
- 805 Rowley, R. J. (2015). Multidimensional community and the Las Vegas experience. *GeoJournal*,
806 80(3), 393-410.
- 807 Ryan, R. M., Bernstein, J. H., & Brown, K. W. (2010). Weekends, work, and well-being:
808 Psychological need satisfactions and day of the week effects on mood, vitality, and physical
809 symptoms. *Journal of Social and Clinical Psychology*, 29(1), 95-122.
- 810 Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends
811 and future directions. *Journal of Travel & Tourism Marketing*, 32(5), 608-621.
- 812 Sharda, R., Delen, D. & Turban, E. (2017). *Business Intelligence, Analytics and Data Science: A*
813 *Managerial Perspective (4th edition)*. Pearson Education.

814 Song, H., & Li, G. (2008). Tourism demand modelling and forecasting – A review of recent
815 research. *Tourism Management*, 29(2), 203-220.

816 Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions
817 and perception of trust. *Tourism Management*, 32(6), 1310-1323.

818 Stringam, B. B., Gerdes Jr, J., & Vanleeuwen, D. M. (2010). Assessing the importance and
819 relationships of ratings on user-generated traveler reviews. *Journal of Quality Assurance in*
820 *Hospitality & Tourism*, 11(2), 73-92.

821 Tinoco, J., Correia, A. G., & Cortez, P. (2011). Application of data mining techniques in the
822 estimation of the uniaxial compressive strength of jet grouting columns over time. *Construction*
823 *and Building Materials*, 25(3), 1257-1262.

824 Turban, E., Aronson, J. E., Liang, T. P. & Sharda, R. (2008). *Decision Support and Business*
825 *Intelligence Systems (8th edition)*. Pearson Education.

826 Usakli, A., & Baloglu, S. (2011). Brand personality of tourist destinations: An application of
827 self-congruity theory. *Tourism Management*, 32(1), 114-127.

828 Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on
829 consumer consideration. *Tourism Management*, 30(1), 123-127.

830 Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and*
831 *techniques*. Morgan Kaufmann.

832 Yang, L. T., & Gu, Z. (2012). Capacity optimization analysis for the MICE industry in Las
833 Vegas. *International Journal of Contemporary Hospitality Management*, 24(2), 335-349.

834 Ye, Q., Law, R., & Gu, B. (2009a). The impact of online user reviews on hotel room sales.
835 *International Journal of Hospitality Management*, 28(1), 180-182.

836 Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler
837 behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings.
838 *Computers in Human Behavior*, 27(2), 634-639.

- 839 Ye, Q., Zhang, Z., & Law, R. (2009b). Sentiment classification of online reviews to travel
840 destinations by supervised machine learning approaches. *Expert Systems with Applications*,
841 36(3), 6527-6535.
- 842 Zeng, B., & Gerritsen, R. (2014). What do we know about social media in tourism? A review.
843 *Tourism Management Perspectives*, 10, 27-36.