# An empirical study on the imbalance phenomenon of data from recommendation questionnaires in the tourism sector

Clara Martin-Duque, Juan José Fernández-Muñoz, Javier M. Moguerza and Aurora Ruiz-Rua

**Abstract**

**Purpose** – *Recommendation systems are a fundamental tool for hotels to adopt a differentiating competitive strategy. The main purpose of this work is to use machine learning techniques to treat imbalanced data sets, not applied until now in the tourism field. These techniques have allowed the authors to analyse the influence of imbalance data on hotel recommendation models and how this phenomenon affects client dissatisfaction.*

**Design/methodology/approach** – *An opinion survey was conducted among hotel customers of different categories in 120 different countries. A total of 135.102 surveys were collected over eleven quarters. A longitudinal design was conducted during this period. A binary logistic model was applied using the function generalized lineal model (GLM).*

**Findings** – *Through the analysis of a representative amount of data, the authors empirically demonstrate that the imbalance phenomenon is systematically present in hotel recommendation surveys. In addition, the authors show that the imbalance exists independently of the period in which the survey is done, which means that it is intrinsic to recommendation surveys on this topic. The authors demonstrate the improvement of recommendation systems highlighting the presence of imbalance data and consequences for marketing strategies.*

**Originality/value** – *The main contribution of the current work is to apply to the tourism sector the framework for imbalanced data, typically used in the machine learning, improving predictive models.*

**Keywords** *Hotel prediction models, Imbalanced data, Satisfaction attributes, Machine learning, Recommendation*

**Paper type** *Research paper*

(Information about the authors can be found at the end of this article.)

Received 7 September 2022
Revised 4 January 2023
30 March 2023
14 April 2023
28 June 2023
3 July 2023
Accepted 7 July 2023

**Subject classification codes** — 5312.90-tourism, 5308.02-consumer behavior



## Introduction

Hotels are experiencing very intense competition due, in part, to the emergence of new business models (peer to peer accommodation) that are being marketed as substitute products and, in part, due to the existence of many hotels with similar characteristics and services (Jasinskas *et al.*, 2016). Similarly, the COVID19 pandemic has led to a change in consumers' perceptions of certain health safety-related attributes (Terzić *et al.*, 2022) that have contributed to an increased competition and a demand for a different offer (Dwivedi *et al.*, 2022). In this context, recommendation systems become fundamental tools for hotels to adopt a differentiating competitive strategy (Chang *et al.*, 2016). In fact, the imbalance phenomenon is related to the lack of accurate data and model results for client dissatisfaction, affecting to the design of future hotel marketing campaigns and trademark reputation development. The importance of client dissatisfaction for hotel recommendation models is a key element in the current work, as there is a relationship between dissatisfaction and loyalty. Griffin (1997) considers the benefits of customer loyalty for a company as an increase in their willingness to try other

product lines after the company has improved its position in the market (relative to competitors) and experienced a sales increase.

Within this context, a study covering hotel stays for 120 countries and eleven trimesters was carried out to understand the existence of imbalance data and its effects on recommendation models. The methodology consists of a longitudinal design conducted during eleven trimesters, based on an anonymous questionnaire made available to hotel clients.

We have applied a generalized lineal model (GLM), a well-known binary logistic model widely used in the research context (Pampel, 2000). The main reason for using this technique is, in addition to its proven effectiveness, its high level of explainability. The weights that GLM models assign to each of their variables are easily interpretable as measures of the importance that the model assigns to each of the explanatory variables. In this way, non-experts in this type of techniques have access to an understanding of the model used (Barredo-Arrieta *et al.*, 2020).

The main contribution of the current work is to address in the tourism recommendation task the existence of imbalanced data, remarking the importance of treating the issue. Our study reviews the econometric techniques that can be used to solve these problems and to build accurate models to improve hotel marketing strategies. The hypothesis formulation is derived from the theoretical background and the existing literature, followed by the methodology and the results. The paper concludes with a discussion on the implications of our findings and conclusions.

## Theoretical context

Hotels attributes are key to customer positive experience as they directly affect customer satisfaction and consequently influence loyalty (Gallarza *et al.*, 2016; Baniya and Thapa, 2017; Bergel *et al.*, 2019; Kunja *et al.*, 2021; Luong *et al.*, 2020; Nguyen *et al.*, 2022) and future intentions to purchase (Akinci and Aksoy, 2019; Bergel *et al.*, 2019). After the current situation caused by the COVID-19 pandemic, technology innovation is likely to play a key role in the hotel industry recovery, primarily focussing on guest interaction with employee's reduction and enhanced cleaning (Heinonen and Strandvik, 2020; Shin and Kang, 2020).

The relationship between satisfaction and recommendation has been extensively studied in the specialized tourism sector literature (Bowen and Chen, 2001; Kandampully and Suhartanto, 2000; Oh, 1999; Raza *et al.*, 2012; Oklevik *et al.*, 2018; Sukhu *et al.*, 2019; Luong *et al.*, 2020; Serra-Cantallops *et al.*, 2018). Most research focuses on how satisfaction affects recommendation (Serra-Cantallops *et al.*, 2018; Luong *et al.*, 2020; Nguyen *et al.*, 2022) rather than focussing on the importance of imbalanced data assessment in the hotel recommendation models and its impact on the decision-making process.

The existing literature makes use of raw data, without considering the imbalance between the larger number of satisfied customers and the smaller number of unsatisfied customers as a drawback. This imbalance of data has been proven to lead to flaws in the prediction algorithms used in machine learning (Fernández Hilario *et al.*, 2018). Therefore, it is not known how the imbalance data affects the prediction models and this generates inaccuracies in the results in the accommodation field. While it is true that there are some studies that identify the existence of imbalanced data in hotels (Chawla *et al.*, 2002, 2004; Li and Sun, 2012; Fernández-Muñoz *et al.*, 2019) none of them go as far as applying it to predictive models.

### *Customer satisfaction and recommendation systems in tourism*

According to some authors, the satisfaction of tourists is directly related to their attempt to revisit or recommend the establishment (Zeithaml *et al.*, 1996; Yoon and Uysal, 2005; González *et al.*, 2007; Moliner *et al.*, 2015; Luong *et al.*, 2020). It should be borne in mind that a satisfied customer does not have to be loyal and customer loyalty does imply satisfaction (Oliver, 1999; Gallarza *et al.*, 2016;

Baniya and Thapa, 2017; Serra-Cantallops *et al.*, 2018). A typical approach to measure satisfaction and loyalty in the industry is the Net Promoter Score (Reichheld, 2003).

Many studies refer to this behaviour of recommending to others as Word Of Mouth (WOM) intention (Dabholkar *et al.*, 1995; Luong *et al.*, 2020) and electronic Word of Mouth (eWOM) recommendation (Ladhari and Michaud, 2015). Marketing studies point to the strength of WOM and eWOM as possessing greater effectiveness than traditional advertising (Cantallops and Salvi, 2014; Kotler *et al.*, 2010). The WOM and eWOM are generated according to the level of satisfaction or quality perceived during the service. Thus, satisfied customers are those who have had a positive experience and want to share it. By contrast, unsatisfied customers are those who have had a bad experience and want to share their discomfort for the service received (Caruana and Schembri, 2016). Customer instability is an aspect that relates to loyalty. When a client has had an unsatisfactory experience, he can respond either by not returning to the establishment, complaining or spreading a negative WOM/eWOM (Aguilar-Rojas *et al.*, 2015).

Customer loyalty has become a very powerful marketing tool for companies (Abdul-Rahman and Kamarulzaman, 2012; Jasinskas *et al.*, 2016; Baniya and Thapa, 2017). Some authors state that, from an economical viewpoint, it is more beneficial to retain existing customers than to attract new ones (Oliver, 1999; Jasinskas *et al.*, 2016). Griffin (1997) sums up the benefits of customer loyalty for a company as follows: sales increase, market positioning improves relative to competitors, marketing costs decrease, customers are less price sensitive and will be more likely to try other product lines.

As far as customer dissatisfaction is concerned, it has been approached from a marketing point of view (Berezina *et al.*, 2016). This perspective points out that while satisfaction relates to positive attitudes towards the brand and the intention to buy back, dissatisfaction relates to negative attitudes (few buyback intentions and negative opinions) (Harrison-Walker, 2001).

Most of these investigations are based on cases where the relationship between client and service provider is satisfactory (Bozzo, 2008). However, customer dissatisfaction can play an important role as it helps us to identify those areas of the hotel that require future improvement (Berezina *et al.*, 2016). In other words, to ensure customer satisfaction it is essential to identify their dissatisfaction (Gazzola *et al.*, 2019; Dinçer and Alrawadieh, 2017).

In the same line of thought and based on the asymmetric relationships approach, Mittal *et al.*, 2001 stated that the product/service attributes' negative performance may have more influence on satisfaction than a positive performance. To measure the asymmetric impacts of the hotel service dimensions on customer satisfaction Davras and Caber (2019) propose the penalty-reward-contrast analysis method. The results show that the Entertainment Services are the only ones that generate such dissatisfaction if they are missing. Bi *et al.* (2020) use the same method, but their study is aimed at evaluating the asymmetric effects of attribute performance (AP). The results demonstrate the existence of asymmetry in the customer satisfaction regarding the market segments.

Regarding the relationship of customer dissatisfaction and recommendations, it should be mentioned that most of the research in hotels has been done from a qualitative perspective. Most of the studies focus on conducting content analysis in which the discourse of customers is analysed. Among them, the study by Berezina *et al.* (2016) analyses and compares the reviews of satisfied and dissatisfied clients through a text-mining approach. The results reveal that satisfied customers tend to cite the most intangible aspects of the service while dissatisfied customers refer to the tangible aspects (attributes). Furthermore, there is an important research work in customer dissatisfaction based on content analysis from TripAdvisor recommendations and other 2.0 platforms. The said work suggests that the existence of asymmetry in hotel Ratings (Fong *et al.*, 2016), in the form of dual-valence reviews (an extreme rating, e.g. "excellent" or "terrible"), may also be associated with reviews featuring both positive and negative comments.

The abovementioned studies (Mittal *et al.*, 2001; Davras and Caber, 2019; Bi *et al.*, 2020) analyse the attribute asymmetry and the results in client satisfaction measurement but they do not consider the existence of imbalance data. They categorize the importance of attributes for customer satisfaction based on how necessary these attributes are for the customer (Mittal *et al.*, 2001). However, the method to differentiate between these attributes or how the reviews are assigned to these attributes (Fong *et al.*, 2016) is purely qualitative.

### *The imbalance phenomenon in tourism data*

Big data allows us to make this distinction without the intervention and interpretation of the researcher (Bagherzadeh *et al.*, 2021; Samara *et al.*, 2020). The existence of imbalance data is related to a quantitative analysis of the sample or data collected, and its treatment or not will produce competitive advantage differences among the companies and their marketing campaigns. Quality hotels studies assume there is an equilibrium between the number of satisfied customers (who would recommend the hotel) and unsatisfied customers (who would not recommend the hotel), but this is usually far from true (Fernández-Muñoz *et al.*, 2019). The existence of imbalanced samples in which the number of satisfied customers is larger than the number of unsatisfied customers directly affects the predictive capacity of the models and has implications for hotel management, development of marketing campaigns, construction of a brand image and the development of competitive advantage. As a result of this we consider Hypothesis 1.

*H1.* There is an imbalance between the number of satisfied and dissatisfied customers.

Over the years, many authors have analysed the relationship between satisfaction and recommendation. Customer satisfaction has been measured through a number of attributes that have varied in the different models proposed. Beyond these attributes, in this section we analyse the different methodologies that have been applied to measure customer satisfaction and their willingness to recommend. Dörtyol *et al.* (2014) analysed the most important dimensions when recommending and determining the attributes that most influence hotel recommendation in Turkey. Based on their findings, the attributes are "Hotel employees and problem solving", "transportation", "food quality and reliability", "climate and hygiene" and "level of price". Using a similar methodology based on regression analysis, Baniya and Thapa (2017) evaluated the relationship between satisfaction, intention to revisit and recommendation for international hotels. They measured customer satisfaction through a series of attributes. On one hand, they evaluated the quality of the service measured through business facility and value; on the other, they evaluated the service offered at Room and Front Desk through Food and Recreation service and Security. Then, they used the same methodology to relate customer satisfaction to loyalty. The results obtained indicated that the tourist satisfaction of the hotels predicts client intention of recommending and loyalty.

Table 1 shows a summary of the most relevant tourism studies on imbalanced data.

Data pre-processing techniques and machine learning models to address the imbalance data in general not only in tourism.

The first general reference on a solution to the imbalance phenomenon is the work by Chawla *et al.* (2002), where a methodology based on the creation of synthetic samples is used to balance the dataset at hand. Regarding the literature on the treatment of imbalanced data in the tourism sector using machine learning techniques, the review paper by Guerra-Montenegro *et al.* (2021) shows that, up to that date, only a remarkable paper was found: Li and Sun (2012) focused on firm failure prediction. In such work, nearest neighbour techniques jointly with a Support Vector Machine (SVM) approach are used to generate additional samples of the minority class.

In 2014, Xu *et al.* (2014), based on the work by Chawla *et al.* (2002), use a synthetic minority over-sampling technique for the prediction of financial distress of Chinese tourism and hospitality firms.

**Table 1** References in tourism studies on imbalanced data

| *Year* | *Author* | *Topic* | *Technique used* | *Period* | *Country* |
|---|---|---|---|---|---|
| 2002 | Chawla *et al* | General paper on imbalanced data | Creation of synthetic samples | – | – |
| 2012 | Li and Sun | Firm failure prediction | K Nearest neighbour | 1998–2010 | China |
| 2014 | Xu *et al* | Financial distress of Chinese tourism and hospitality firms | Synthetic minority over-sampling | 1999–2013 | China |
| 2018 | Fernández *et al* | Recommendation of hotel stays | Logit regression model | 2014 | Various |
| 2021 | Guerra-Montenegro *et al* | Computational Intelligence in the hospitality industry | Systematic review | 1998–2018 | Various |
| 2022 | Ma | Planification strategies within a tourism management system | Pre-processing of the minority class | Unspecified | Unspecified |
| 2022 | Hoffman *et al* | Accommodations' sustainability | Comparison of the machine learning techniques | 2020 | 37 European countries |

**Source(s):** Prepared by the authors

In 2018, Fernández Hilario *et al.* (2018) performed a study showing that the imbalance in the data from recommendation questionnaires affects the prediction accuracy of the models used, especially the prediction provided by unsatisfied clients, tending to consider them as satisfied customers. In 2022, Ma (2022) used another variation of the Chawla *et al.* (2002) methodology by pre-processing the minority class using a k-means clustering technique. In this case, the aim was to predict planification strategies within a tourism management system. And finally, in Hoffmann *et al.* (2022) different machine learning algorithms are compared to measure their accuracy on the prediction of accommodations' sustainability.

Although the previous studies that identify the existence of imbalanced data in the tourism sector, only the work by Fernández-Muñoz *et al.* (2019) is focused on the prediction of recommendations from quality attributes of the hotels. Therefore, there are strong reasons to consider the validation of Hypothesis 2.

*H2.* The imbalance between the number of satisfied and dissatisfied customers affects the accuracy of statistical techniques typically used to predict client recommendation preferences.

In summary, the main objective of this paper is to show systematically how the imbalance phenomenon affects to the prediction accuracy of machine learning techniques. The studies listed in Table 1 are focused on techniques and not on data. As a novelty, in this work, we concentrate on the data perspective and the systemic imbalanced structure of data obtained from recommendation questionnaires in the tourism sector.

## Methodology

### *Data collection procedure*

The sample collects the opinions of clients on 5 attributes at holiday and urban hotels of different categories in various countries. These opinions have been collected by means of a survey that is different from websites such as Tripadvisor where opinions are collected only from those customers who intend to write a review, prioritizing those who are either very satisfied or very dissatisfied. In this case, a greater diversity of opinions is collected.

The questionnaire is focused basically on recollecting customers' evaluations about their quality perceptions. Five main general aspects of the stay: costumer service (Nunkoo *et al.*, 2019); price

(Han and Hyun, 2015); cleanliness (Malik *et al.*, 2020; Nunkoo *et al.*, 2019); facilities; recommendation (Bowen and Chen, 2001; Kandampully and Suhartanto, 2000; Oh, 1999; Raza *et al.*, 2012; Oklevik *et al.*, 2018; Sukhu *et al.*, 2019; Luong *et al.*, 2020; Serra-Cantallops *et al.*, 2018).

The customer's perceptions were recollected together with the following data about the main features of the hotels: stars; typology, urban or vacation; country.

The data provide opinions obtained at hotels from around the world (America, Africa, Asia, Europe and Australia), with a total of 135,102 opinions from customers during eleven consecutive trimesters. In contrast to other research that focuses on a single area of study (González *et al.*, 2007), this article addresses the issue with field work covering 120 countries from all continents. The average number of stays per trimester at 104 hotels is 12,282.

A longitudinal design was conducted during eleven trimesters. The study was based on the answers to an open questionnaire made available to hotel clients, either personally or by e-mail, immediately at the end of their stay. The survey was anonymous and voluntary, with no ethical approvals needed and the participants did not have any reward. Due to the large amount of completed questionnaires, those with missing data were removed. The data was analysed using the statistical software R (R Development Core Team, 2008) and stored and administered by company HOTELS QUALITY (www.hotels-quality.com). It is also important to remark that more recent data show similar patterns, but due to confidentiality reasons, only the data corresponding to the period 2015–2017 were available. For the same reason, the data have been masked preserving the relevant information to carry out the current analysis.

Tables 2–5 describe the sample from different perspectives. Tables 2 and 3 show the absolute frequency distribution of the hotels involved in this study, whereas Tables 4 and 5 present the absolute frequency distribution of the customers stays. Finally, Table 6 includes the positive and negative proportion of recommendations given by the clients.

A systematic high imbalance of data can be observed in Table 6. In particular, the proportion of positive answers ranges from 0.95 to 0.97 whereas the proportion of negative answers goes from 0.3 to 0.5. Although there are many studies in the literature using predictive methods for recommendation data, none reports this systematic imbalance. Li and Sun (2012) analyse cases in which the data set only reports negative cases in a proportion of 2%. Hirokawa and Hashimoto (2018) consider a prediction model with very bad results because they have only 10% of negative cases.

In this regard, some previous studies present isolated solutions to solve this, but it is not assumed as an endemic situation within data coming from recommendation surveys (Fernández-Muñoz *et al.*, 2019; Li and Sun, 2012).

**Table 2** Absolute frequency distribution of the hotels by typology

| *Year* | *Trimester* | *Total* | *Vacation* | *Urban* | *Other* |
|---|---|---|---|---|---|
| 2015 | T2 | 109 | 11 | 93 | 5 |
| | T3 | 102 | 15 | 84 | 3 |
| | T4 | 217 | 31 | 127 | 59 |
| 2016 | T1 | 95 | 14 | 70 | 11 |
| | T2 | 99 | 16 | 70 | 13 |
| | T3 | 97 | 15 | 70 | 12 |
| | T4 | 91 | 15 | 65 | 11 |
| 2017 | T1 | 87 | 13 | 64 | 10 |
| | T2 | 85 | 13 | 63 | 9 |
| | T3 | 85 | 17 | 58 | 10 |
| | T4 | 80 | 15 | 56 | 9 |

**Source(s):** Prepared by the authors

**Table 3** Absolute frequency distribution of the hotels by stars

| *Year* | *Trimester* | *Total* | *5 stars* | *4 stars* | *3 stars* | *2 stars* | *Other* |
|---|---|---|---|---|---|---|---|
| 2015 | T2 | 109 | 20 | 36 | 34 | 3 | 14 |
| | T3 | 102 | 18 | 35 | 28 | 3 | 18 |
| | T4 | 217 | 40 | 86 | 60 | 9 | 22 |
| 2016 | T1 | 95 | 15 | 30 | 28 | 4 | 18 |
| | T2 | 99 | 17 | 31 | 26 | 3 | 22 |
| | T3 | 97 | 15 | 70 | 10 | 2 | 0 |
| | T4 | 91 | 18 | 30 | 22 | 1 | 20 |
| 2017 | T1 | 87 | 17 | 28 | 20 | 2 | 20 |
| | T2 | 85 | 17 | 25 | 20 | 2 | 21 |
| | T3 | 85 | 16 | 28 | 19 | 2 | 20 |
| | T4 | 80 | 17 | 22 | 18 | 2 | 21 |

**Source(s):** Prepared by the authors

**Table 4** Absolute frequency distribution of the stays by typology

| *Year* | *Trimester* | *Total* | *Vacation* | *Urban* | *Other* |
|---|---|---|---|---|---|
| 2015 | T2 | 16,837 | 4,919 | 10,351 | 1,567 |
| | T3 | 13,950 | 3,724 | 9,415 | 811 |
| | T4 | 12,552 | 1,208 | 10,207 | 1,137 |
| 2016 | T1 | 12,728 | 1,088 | 10,296 | 1,344 |
| | T2 | 12,747 | 1,069 | 10,230 | 1,448 |
| | T3 | 12,791 | 1,444 | 10,374 | 973 |
| | T4 | 11,987 | 973 | 10,330 | 684 |
| 2017 | T1 | 10,516 | 569 | 9,254 | 693 |
| | T2 | 10,833 | 505 | 9,471 | 858 |
| | T3 | 10,587 | 747 | 8,674 | 1,166 |
| | T4 | 9,574 | 536 | 8,366 | 672 |

**Source(s):** Prepared by the authors

**Table 5** Absolute frequency distribution of the stays by typology

| *Year* | *Trimester* | *Total* | *5 stars* | *4 stars* | *3 stars* | *2 stars* | *Other* |
|---|---|---|---|---|---|---|---|
| 2015 | T2 | 16,837 | 3,149 | 5,407 | 4,423 | 45 | 3,813 |
| | T3 | 13,950 | 2,499 | 4,136 | 3,498 | 75 | 3,742 |
| | T4 | 12,552 | 2,740 | 3,606 | 3,301 | 74 | 2,831 |
| 2016 | T1 | 12,728 | 2,644 | 2,895 | 3,267 | 84 | 3,838 |
| | T2 | 12,747 | 2,932 | 3,155 | 2,870 | 77 | 3,713 |
| | T3 | 12,791 | 2,804 | 3,184 | 2,836 | 67 | 3,900 |
| | T4 | 11,987 | 2,240 | 2,698 | 3,116 | 35 | 3,898 |
| 2017 | T1 | 10,516 | 2,029 | 3,000 | 2,353 | 18 | 3,116 |
| | T2 | 10,833 | 2,306 | 2,749 | 2,618 | 43 | 3,117 |
| | T3 | 10,587 | 2,223 | 2,649 | 2,554 | 47 | 3,114 |
| | T4 | 9,574 | 1,926 | 1,981 | 2,485 | 65 | 3,117 |

**Source(s):** Prepared by the authors

### *Instrument and measurement*

The questionnaire is organized in two sections. Firstly, descriptive statistics such as the typology (urban or vacational) and the number of hotels stars (from two to five) were collected. Secondly, the clients answer to questions about four general attributes related to their stay, namely: customer services, general cleanliness, facilities and quality–price ratio. Likert's scale was used to measure

**Table 6** Proportions of positive and negative recommendations

| *Year* | *Trimester* | *Total* | *Proportion Positive* | *Proportion Negative* |
|---|---|---|---|---|
| 2015 | T2 | 16,837 | 0.96 | 0.04 |
| | T3 | 13,950 | 0.95 | 0.05 |
| | T4 | 12,552 | 0.97 | 0.03 |
| 2016 | T1 | 12,728 | 0.96 | 0.04 |
| | T2 | 12,747 | 0.96 | 0.04 |
| | T3 | 12,791 | 0.96 | 0.04 |
| | T4 | 11,987 | 0.96 | 0.04 |
| 2017 | T1 | 10,516 | 0.96 | 0.04 |
| | T2 | 10,833 | 0.96 | 0.04 |
| | T3 | 10,587 | 0.96 | 0.04 |
| | T4 | 9,574 | 0.96 | 0.04 |

**Source(s):** Prepared by the authors

the clients' attributes perception ranging from 1 to 6 (1 = very poor, 2 = poor, 3 = fair, 4 = good, 5 = very good and 6 = excellent).

The election of attributes is based on the consumer decision-making process (CDP) model stated that customers get influenced by some special elements in products and thus end up purchasing in corresponding to their needs and preferences.

Chang and Wong (2005) stated that hotel attributes can be either intangible or tangible characteristics or physical attributes (elements that can be seen) such as price, facilities, location, the existence of choices, WOM, advertising, a familiar name and past experience, etc. Intangible elements for a hotel can be characteristics such as security, dependability, service quality, reputation and staff behaviour. This inherent feature of the service makes customer recommendations a very important tool when deciding to buy a product (Gellerstedt and Arvemo, 2019). In this sense Kim *et al.* (2019), in their study about hotel industry in Korea, revealed in the luxury and upscale hotels a significant gap between the importance of the attributes and their satisfaction. This gap is even larger when confronting intangible attributes against tangible attributes, generating differences in the positive WOM and revisit intention. Bodet *et al.* (2017) highlight the importance of cleanliness and facilities as basic attributes in their cross-country and cross-hotel study for a Tetraclasse model.

All the studies based on hotel attributes show that customers prefer various attributes (instead of a single one) and similar features of accommodation facility.

Baniya and Thapa (2017) highlight the importance of service quality, room and front desk as the most important hotel attributes that lead to international tourist satisfaction in Nepal. They stated the relationship between satisfaction and loyalty and the fact that with superior performance of the two attributes the tourists are likely to revisit the Nepal (Pokhara) hotels and generate positive WOM.

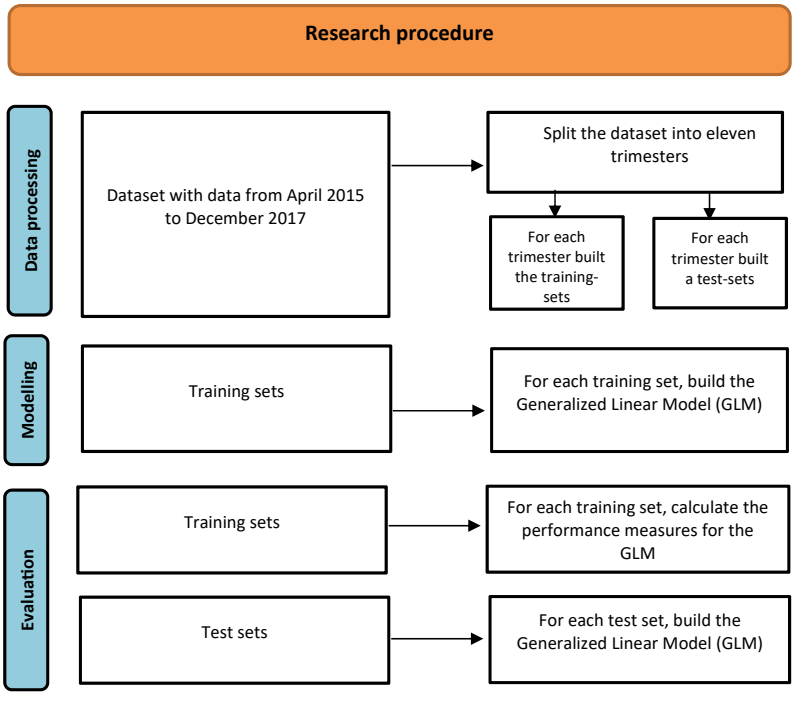
In order to analyse customer recommendation, a dichotomous recommendation variable was included in this second part of the questionnaire (0 = no recommendation, 1 = recommendation). In our analysis, the previous four general attributes constitute the independent variables while the dichotomous one constitutes the dependent variable.

The data have been divided divide into a training and testing dataset: 70% of the data were used to train the model and 30% of the data to test the model, obtaining similar results for both cases. For the sake of space, we only show the results for the training data set. Figure 1 shows the flowchart describing the investigation procedure.

### *Analysis*

For each of the eleven trimesters a binary logistic model (Hosmer and Lemeshow, 2000; Kleinbaum and Kleim, 2002; Long, 1997; Pampel, 2000) has been executed using the "GLM"

**Figure 1** Flow chart describing the investigation procedure




**Source(s):** Prepared by the authors

function included in the R software, which has been parametrized to reflect the dichotomous characteristic of the dependent variable. Given that this work is focused on the systemic structure of data, we choose the GLM model due to its well-known simplicity and effectiveness and no pre-processing data techniques are used. We use this model to determine whether the recommendation variable can be explained in terms of the set of independent attributes. In such model, the probability of recommendation is given by P(Y = 1), whereas the non-recommendation probability P(Y = 0) equals 1-P(Y = 1). In the case at hand, the hypothesis is that $X_1$, $X_2$, $X_3$ and $X_4$ (respectively, customer services, general cleanliness, facilities and quality–price ratio) are four influential factors on response attribute Y (the recommendation variable). The logistic regression model represents the ratio of recommendation and non-recommendation probabilities through its natural logarithm:

$$\text{Logit P} = \alpha + \beta 1\text{X1} + \beta 2\text{X2} + \beta 3\text{X3} + \beta 4\text{X4} \quad (1)$$

## *Results*

In this section, we analyse the performance of the predictive method using a battery of well-known measures. A detail analysis and description of such measures can be consulted in De Diego *et al.* (2022). Next, we describe such measures. For the sake of simplicity, we consider a binary classification problem with positive (+1) and negative (−1) labels:

1. True Positive (TP): a true positive takes place when a datum is correctly classified into class +1 while belonging to class +1.
2. True Negative (TN): a true negative takes place when a datum is correctly classified into class −1 while belonging to class −1.

3. False Positive (FP): a false positive takes place when a datum is wrongly classified into class $+1$ while belonging to class $-1$.
4. False Negative (FN): a false negative takes place when a datum is wrongly classified into class $-1$ while belonging to class $+1$.
5. Accuracy (ACC): proportion of correctly classified data, that is, $\text{ACC} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$.
6. Positive Predictive Value (PPV): also known as precision, proportion of true positives out of all data predicted as positive, that is, $\text{PPV} = \frac{\text{TP}}{\text{TP}+\text{FP}}$.
7. True Positive Rate (TPR): also called sensitivity or recall, proportion of true positives out of all positive data, that is, $\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}}$.
8. Negative Predictive Value (NPV): proportion of true negatives out of all data predicted as negative, that is, $\text{NPV} = \frac{\text{TN}}{\text{TN}+\text{FN}}$.
9. True Negative Rate (TNR): also known as specificity, proportion of true negatives out of all negative data, that is, $\text{TNR} = \frac{\text{TN}}{\text{TN}+\text{FP}}$.
10. Positive F1 score ($F_1^+$): joint measure of positive accuracy defined as the harmonic mean of PPV and TPR, that is, $F_1^+ = 2\,\frac{\text{PPV}\times\text{TPR}}{\text{PPV}+\text{TPR}}$.
11. Negative F1 score ($F_1^-$): joint measure of negative accuracy defined as harmonic mean of NPV and TNR, that is, $F_1^- = 2\,\frac{\text{NPV}\times\text{TNR}}{\text{NPV}+\text{TNR}}$.

Table 7 shows the evaluation measures for each trimester since April 2015 to December 2017. It can be observed that systematically every trimester the Positive Predictive Value and True Positive Rate measures take very high values close to 1, whereas the Negative Predictive Value and True Negative Rate clearly reach lower values (ranging TPR from 0.64 to 0.72 and TNR from 0.49 to 0.61). This performance is also reflected in the Positive F1 score (a summary of PPV and TPR) and Negative F1 score (a summary of NPV and TNR). Therefore, $F_1^+$ and $F_1^-$ describe the performance of the predictive method for each class separately.

Nevertheless, the measurement of Accuracy is clearly affected by the imbalance of the data. The fact that the number of positive instances is much larger than the number of negative samples implies that the weight of the True Positive instances masks the bad performance of the predictive method for the negative class in the Accuracy measure. These results are coherent with the descriptive statistics in Table 6, where systematic high imbalance of data was reported.

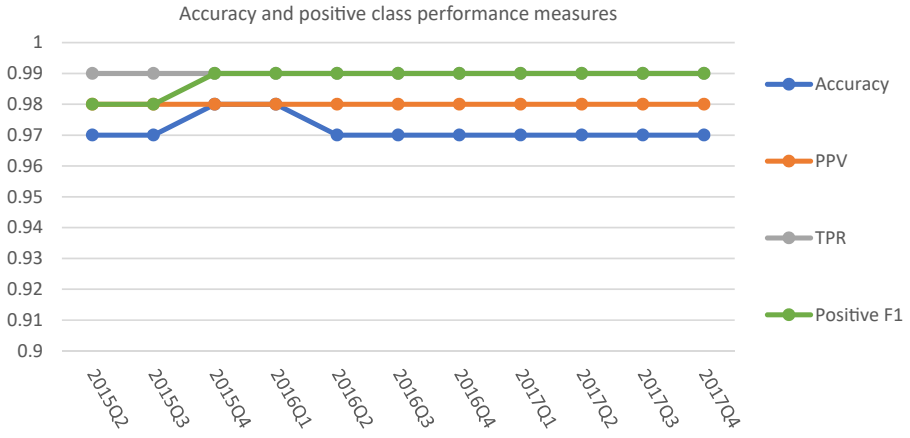
Figures 2 and 3 summarize these findings. In Figure 2, it is apparent that the systematic performance measures related to the positive class, that is, PPV, TPR and $F_1^+$, tend to reach values close to one, whereas in Figure 3 it can be observed that the performance measures related to the

**Table 7** Evaluation measures for each trimester

| *Year* | *Trimester* | *TP* | *TN* | *FP* | *FN* | *ACC* | *PPV* | *TPR* | *NPV* | *TNR* | $F_1^+$ | $F_1^-$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2015 | T2 | 15,938 | 384 | 367 | 148 | 0.97 | 0.98 | 0.99 | 0.72 | 0.51 | 0.98 | 0.60 |
| | T3 | 13,078 | 426 | 273 | 173 | 0.97 | 0.98 | 0.99 | 0.71 | 0.61 | 0.98 | 0.66 |
| | T4 | 12,092 | 158 | 212 | 90 | 0.98 | 0.98 | 0.99 | 0.64 | 0.43 | 0.99 | 0.51 |
| 2016 | T1 | 12,169 | 251 | 196 | 112 | 0.98 | 0.98 | 0.99 | 0.69 | 0.56 | 0.99 | 0.62 |
| | T2 | 12,161 | 240 | 227 | 119 | 0.97 | 0.98 | 0.99 | 0.67 | 0.51 | 0.99 | 0.58 |
| | T3 | 12,166 | 256 | 271 | 98 | 0.97 | 0.98 | 0.99 | 0.72 | 0.49 | 0.99 | 0.58 |
| | T4 | 11,375 | 274 | 214 | 124 | 0.97 | 0.98 | 0.99 | 0.69 | 0.56 | 0.99 | 0.62 |
| 2017 | T1 | 10,036 | 193 | 187 | 100 | 0.97 | 0.98 | 0.99 | 0.66 | 0.51 | 0.99 | 0.57 |
| | T2 | 10,298 | 224 | 217 | 94 | 0.97 | 0.98 | 0.99 | 0.70 | 0.51 | 0.99 | 0.59 |
| | T3 | 10,071 | 245 | 170 | 101 | 0.97 | 0.98 | 0.99 | 0.71 | 0.59 | 0.99 | 0.64 |
| | T4 | 9,147 | 182 | 157 | 88 | 0.97 | 0.98 | 0.99 | 0.67 | 0.54 | 0.99 | 0.60 |

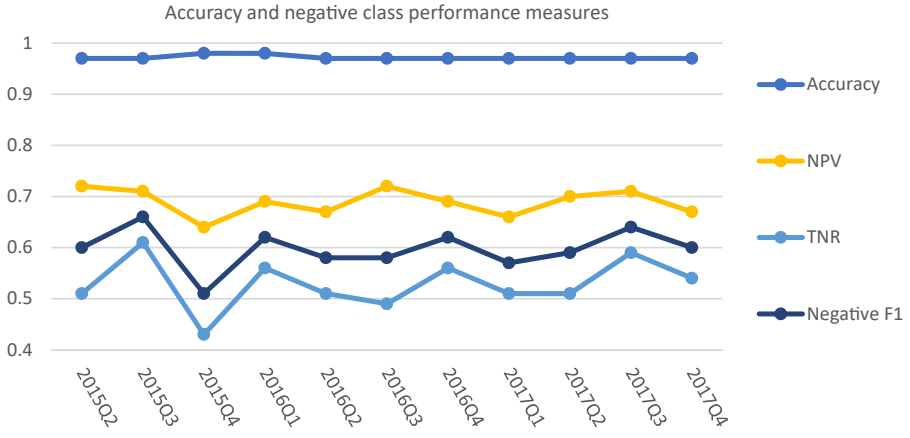
**Source(s):** Prepared by the authors

**Figure 2** Accuracy versus performance measures for the positive class




**Source(s):** Prepared by the authors

**Figure 3** Accuracy versus performance measures for the negative class




**Source(s):** Prepared by the authors

negative class, that is, NPV, TNR and $F_1^-$, tend to reach lower values. However, the overall summary measure, Accuracy (shown in both figures), is clearly aligned with performance measures related to positive class, losing its *a priori* overall descriptive capacity.

Finally, although not included for the sake of space, it is important to remark that an analysis of the data categorize either by countries or by continent provides similar results.

## Discussion and conclusions

The results show the importance of applying an unbalanced data approach when assessing the perceptions of dissatisfied customers and defining differentiating strategies in hotels (Griffin, 1997; Chang *et al.*, 2016) in a context of an increasingly demanding consumer (Dwivedi *et al.*, 2022).

In this regard, the study has checked two main hypotheses regarding the importance of data collection from recommendation surveys in the hotel sector. Both faced current most important challenges to understand the tourism market that affect significantly to the decision-making process such as marketing and promotion strategies. In this vein, Li *et al.* (2010) highlight the importance to incorporate negative association rules in marketing and they provide a novel mining approach to estimate the confidence of targeted association rules (positive and negative) in their analysis of Outbound Tourism in Hong Kong.

Many authors who have analysed attribute asymmetry (Mittal *et al.*, 2001; Davras and Caber, 2019; Bi *et al.*, 2020 and Fong *et al.*, 2016; Kim *et al.*, 2019) have revealed a significant gap between the importance of the attributes and their satisfaction, most important in the intangible attributes against tangible attributes. however, there are hardly any studies of customer satisfaction that take into account the imbalance data in their studies (Fernández-Muñoz *et al.*, 2019; Li and Sun, 2012). Therefore, the importance of the hypotheses we propose.

Concerning to hypothesis 1, which refers to the existence of unbalanced data between satisfied and dissatisfied customers, we have shown that there is a generalized imbalance in the data collected from the opinions of hotel clients. This imbalance is shown to be inherent in the data collection process and therefore seldom addressed in the academic literature. Li and Sun (2012) and Hirokawa and Hashimoto (2018) are good examples of this.

Through the analysis of a representative amount of data, we empirically demonstrate that the imbalance phenomenon is systematically present in hotel services recommendation surveys. In addition, we show that such imbalance exists independently of the period in which the survey is done, which means that it is intrinsic to recommendation surveys on this topic.

Similarly, there are many studies that analyse the relationship between customer satisfaction and recommendation (Bowen and Chen, 2001; Kandampully and Suhartanto, 2000; Oh, 1999; Raza *et al.*, 2012; Oklevik *et al.*, 2018; Sukhu *et al.*, 2019; Luong *et al.*, 2020; Serra-Cantallops *et al.*, 2018) none of them consider the imbalance of the data in the predictive model.

With regards to hypothesis 2, we have shown that imbalance in the dataset clearly affects the predictability that results from construction of models with such data. An inadequately constructed model undoubtedly produces inadequate results for decision making. In this sense, we empirically demonstrate that the quality of the results obtained using statistical prediction procedures is markedly affected by such a phenomenon.

We identify that the most affected prediction results are those related to unsatisfied customers that provide negative recommendations. It is observed that systematically, throughout all trimesters, the overall accuracy and measures associated to data related to satisfied customers providing positive recommendation (PPV, TPR and $F_1^+$) are very high (close to 1). However, measures associated to data related to unsatisfied customers providing negative recommendations (NPV, TNR and $F_1^-$) are clearly lower.

In fact, TNR takes systematically values ranging from 0.5 to 0.6. This means that the model considers at least 40% of dissatisfied customers as satisfied customers, not taking into account the overall accuracy measure. In a well-posed model, the $F_1^+$, $F_1^-$ and accuracy measures are expected to reach similar values. Since $F_1^+$ and $F_1^-$ describe the performance of the predictive method for each class separately (but accurately), our recommendation is to use both measures jointly to correctly describe the overall performance of predictive methods built from imbalanced samples.

As a consequence of the above, if there is imbalance in the data obtained from recommendations, and this mainly affects negative recommendations, we do not have certainty about the adequacy of the attributes in the definition for those who make negative recommendations (Bi *et al.*, 2020; Davras and Caber, 2019). Moreover, the probability that some of the attributes are not critical (or the most important) for those who make negative recommendations is very high.

Once the imbalance problem has been detected, corrective actions should be carried out. In the tourism environment, a reduced literature is available, usually focused on a particular instance of problem. A recent approach is presented in Fernández-Muñoz *et al.* (2019), where a balanced sample is built by means of a subsample procedure on the majority group, that is, the set of customers providing a positive recommendation. With this technique, the overall error and the particular error for each class become similar. However, the approach is not systematically applied to a long period of trimesters and should therefore be considered a cross sectional design with non-concluding results.

There are more sophisticated strategies such as synthetic minority over-sampling technique (SMOTE) developed by Chawla *et al.* (2002). In this case, the strategy followed by the authors is to build synthetic samples of the minority groups, that is, the set of customers providing a negative recommendation in order to obtain a balanced dataset. As far as we know, this approach has not been applied to data coming from recommendation surveys within the tourism sector and therefore constitutes a promising research line.

### *Implications*

We have shown the importance and the implications of imbalanced data for hotel prediction models and the influence it may have on the decision-making process for hotels.

The analysed data shows that there is a systematic imbalance in all semesters. This imbalance leads to an error in the prediction of dissatisfied customers. That is, if we base the prediction on imbalanced data, we drag that error into the prediction. As it has been proven in studies that analyse customer satisfaction and their intention to recommend, the imbalance data consideration is not usually applied. It would therefore be convenient to do so in order to make predictions that are more adjusted to the reality of the sample.

The hotel prediction models require an accurate design of the attributes to correctly reflect the most important characteristics of their products and services to capture client's attention. Otherwise, if dissatisfied client perception is not well defined (because the imbalance data does not allow to correctly measure that perception), it is possible that hotels will end up developing ineffective campaigns that will not capture client's attention. Thus, clients will remain dissatisfied (or even become more dissatisfied) as there are doubts about the products and services they view negatively and on which they base their negative recommendation.

This situation, which directly affects the predictive capacity of the model, has other implications clearly related to hotel management, development of competitive offers, construction of brand image and development of competitive advantage. The definition of competitive offers based on attributes that affect negative recommendations may be one of the implications. A poor definition of the characteristics of a dissatisfied client would become a pervasive problem.

The application of an imbalanced data approach improves the detection of dissatisfied customers and as expressed in the literature, this can help to detect the main deficiencies in the service and improve customer perception of quality. It can also help to detect those attributes that most influence the customer recommendation in a more precise way, being very useful for the design of customer loyalty campaigns and the design of relationship marketing strategies. These findings could, for instance, be applied to marketing campaigns focused on hotel image improvement among dissatisfied customers.

Regarding further research a similar analysis should be done comparing the results with other machine learning models, and then, select the most accurate one for imbalanced data. Also, the analysis could be extended introducing resampling or pre-processing techniques to handle the imbalance problem.

### *Limitations*

The sample is composed of four attributes, but the model can, however, be enriched with more attributes, as other authors do. In order to reach a larger sample of hotels, the model was reduced

to 4 attributes in order to assess customer satisfaction. However, in future developments, it would be convenient to introduce attributes that are already present in other studies, such as the attitude of the hotel staff, to complement the model.

The results of this work present some threats to validity that should be mentioned: First, data were collected from a non-probabilistic sampling conditioned to commercial agreements between hotels and the polling contractor. Second, in the survey, the questionnaire was self-administered, and, therefore, only the opinion of volunteers is taken into account. This fact could partly explain the imbalance in the dataset, as satisfied customers tend to participate more responsively in this type of questionnaires, third, the weight of four and five start hotels was significantly higher than the weight of hotels with three stars or less; and finally, after the COVID-19 pandemic, it would be advisable to replicate the analysis in order to check the validity of the findings.

## References


Abdul-Rahman, M. and Kamarulzaman, Y. (2012), "The influence of relationship quality and switching costs on customer loyalty in the Malaysian hotel industry", *Procedia – Social and Behavioral Sciences*, Vol. 62, pp. 1023-1027, doi: 10.1016/j.sbspro.2012.09.174.

Aguilar-Rojas, O., Fandos-Herrera, C. and Flavián-Blanco, C. (2015), "What may lead you to recommend and revisit a hotel after a service failure instead of complaining?", *International Journal of Contemporary Hospitality Management*, Vol. 27 No. 2, pp. 214-235, doi: 10.1108/ijchm-06-2013-0265.

Akinci, S. and Aksoy, S. (2019), "The impact of service recovery evaluation on word-of-mouth intention: a moderated mediation model of overall satisfaction, household income and gender", *Tourism Management Perspectives*, Vol. 31, pp. 184-194, doi: 10.1016/j.tmp.2019.05.002.

Bagherzadeh, S., Shokouhyar, S., Jahani, H. and Sigala, M. (2021), "A generalizable sentiment analysis method for creating a hotel dictionary: using big data on TripAdvisor hotel reviews", *Journal of Hospitality and Tourism Technology*, Vol. 12 No. 2, pp. 210-238, doi: 10.1108/jhtt-02-2020-0034.

Baniya, R. and Thapa, P. (2017), "Hotel attributes influencing international tourists' satisfaction and loyalty", *Journal of Tourism and Hospitality Education*, Vol. 7, pp. 44-61, doi: 10.3126/jthe.v7i0.17689.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, R. (2020), "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible artificial intelligence", *Information Fusion*, Vol. 58, pp. 82-115, doi: 10.1016/j.inffus.2019.12.012.

Berezina, K., Bilgihan, A., Cobanoglu, C. and Okumus, F. (2016), "Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews", *Journal of Hospitality Marketing and Management*, Vol. 25 No. 1, pp. 1-24, doi: 10.1080/19368623.2015.983631.

Bergel, M., Frank, P. and Brock, C. (2019), "The role of customer engagement facets on the formation of attitude, loyalty and price perception", *Journal of Services Marketing*, Vol. 33 No. 7, pp. 890-903, doi: 10.1108/jsm-01-2019-0024.

Bi, J.-W., Liu, Y., Fan, Z.-P. and Zhang, J. (2020), "Exploring asymmetric effects of attribute performance on customer satisfaction in the hotel industry", *Tourism Management*, Vol. 77, 104006, doi: 10.1016/j.tourman.2019.104006.

Bodet, G., Anaba, V. and Bouchet, P. (2017), "Hotel attributes and consumer satisfaction: a cross-country and cross-hotel study", *Journal of Travel and Tourism Marketing*, Vol. 34 No. 1, pp. 52-69, doi: 10.1080/10548408.2015.1130109.

Bowen, J.T. and Chen, S. (2001), "The relationship between customer loyalty and customer satisfaction", *International Journal of Contemporary Hospitality Management*, Vol. 13 No. 5, pp. 213-217, doi: 10.1108/09596110110395893.

Bozzo, C. (2008), "Different reasons why dissatisfied customers stay with their supplier", *24th IMP-Conference in Uppsala*, Sweden.

Cantallops, A.S. and Salvi, F. (2014), "New consumer behavior: a review of research on eWOM and hotels", *International Journal of Hospitality Management*, Vol. 36, pp. 41-51, doi: 10.1016/j.ijhm.2013.08.007.

Chang, E. and Wong, S. (2005), "Identifying and exploiting potentially lucrative niche markets: the case of planned impulse travelers. Hong Kong tourism SMEs", *Service Quality and Destination Competitiveness*, pp. 295-311, doi: 10.1079/9780851990118.0295.

Caruana, S. and Schembri, C. (2016), "The significance of electronic Word-of-Mouth (e-WOM) content in the shaping of the visitor's perception of quality and value", *Springer Proceedings in Business and Economics*, Springer, pp. 535-550.

Chang, Y., Ko, Y.J. and Leite, W.L. (2016), "The effect of perceived brand leadership on luxury service WOM", *Journal of Services Marketing*, Vol. 30 No. 6, pp. 659-671, doi: 10.1108/jsm-01-2015-0005.

Chawla, N., Hall, L., Bowyer, K. and Kegelmeyer, W. (2002), "SMOTE: synthetic minority oversampling technique", *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321-357.

Dabholkar, P., Thorpe, D.I. and Rentz, J.Q. (1995), "A measure of service quality for retail stores", *Journal of the Academy of Marketing Science*, Vol. 24 No. 1, pp. 3-16.

Davras, Ö. and Caber, M. (2019), "Analysis of hotel services by their symmetric and asymmetric effects on overall customer satisfaction: a comparison of market segments", *International Journal of Hospitality Management*, Vol. 81, pp. 83-93, doi: 10.1016/j.ijhm.2019.03.003.

De Diego, I.M., Redondo, A.R., Fernández, R.R., Navarro, J. and Moguerza, J. (2022), "General performance score for classification problems", *Applied Intelligence*, Vol. 52, pp. 12049-12063, doi: 10.1007/s10489-021-03041-7.

Dinçer, M.Z. and Alrawadieh, Z. (2017), "Negative word of mouse in the hotel industry: a content analysis of online reviews on luxury hotels in Jordan", *Journal of Hospitality Marketing and Management*, Vol. 26 No. 8, pp. 785-804, doi: 10.1080/19368623.2017.1320258.

Dörtyol, İ.T., Varinli, İ. and Kitapci, O. (2014), "How do international tourists perceive hotel quality?", *International Journal of Contemporary Hospitality Management*, Vol. 26 No. 3, pp. 470-495, doi: 10.1108/ijchm-11-2012-0211.

Dwivedi, R.K., Pandey, M., Vashisht, A., Pandey, D.K. and Kumar, D. (2022), "Assessing behavioral intention toward green hotels during COVID-19 pandemic: the moderating role of environmental concern", *Journal of Tourism Futures*, Vol. ahead-of-print No. ahead-of-print, doi: 10.1108/JTF-05-2021-0116.

Fernández Hilario, A., García López, S., Galar, M., Prati, R.C., Krwaczyk, B. and Herrera, F. (2018), *Learning from Imbalanced Data Sets*, Springer, doi: 10.1007/978-3-319-98074-4.

Fernández-Muñoz, J.J., Moguerza, J.M., Martin Duque, C. and Gomez Bruna, D. (2019), "A study on the effect of imbalanced data in tourism recommendation models", *International Journal of Quality and Service Sciences*, Vol. 11 No. 3, pp. 346-356, doi: 10.1108/IJQSS-05-2018-0050.

Fong, L.H.N., Lei, S.I. and Law, R. (2016), "Asymmetry of hotel ratings on tripadvisor: evidence from single- versus dual-valence reviews", *Journal of Hospitality Marketing and Management*, Vol. 26 No. 1, pp. 67-82, doi: 10.1080/19368623.2016.1178619.

Gallarza, M.G., Arteaga-Moreno, F., Del Chiappa, G. and Gil-Saura, I. (2016), "Intrinsic value dimensions and the value-satisfaction-loyalty chain: a causal model for services", *Journal of Services Marketing*, Vol. 30 No. 2, pp. 165-185, doi: 10.1108/jsm-07-2014-0241.

Gazzola, P., Grechi, D., Pavione, E. and Ossola, P. (2019), "Albergo diffuso, model for the analysis of customer satisfaction", *European Scientific Journal*, Vol. 5 No. 25, pp. 1-25, doi: 10.19044/esj.2019.v15n25p1.

Gellerstedt, M. and Arvemo, T. (2019), "The impact of word of mouth when booking a hotel: could a good friend's opinion outweigh the online majority?", *Information Technology and Tourism*, Vol. 21 No. 3, pp. 289-311, doi: 10.1007/s40558-019-00143-4.

González, M.E.A., Comesaña, L.R. and Brea, J.A.F. (2007), "Assessing tourist behavioral intentions through perceived service quality and customer satisfaction", *Journal of Business Research*, Vol. 60 No. 2, pp. 153-160, doi: 10.1016/j.jbusres.2006.10.014.

Griffin, J. (1997), *Customer Loyalty: How to Earn it, How to Keep it*, Lexington Books, New York, NY.

Guerra-Montenegro, J., Sanchez-Medina, J., Laña, I., Sanchez-Rodriguez, D., Alonso-González, I. and Del Ser, J. (2021), "Computational Intelligence in the hospitality industry: a systematic literature review and a prospect of challenges", *Applied Soft Computing*, Vol. 102, 107082, doi: 10.1016/j.asoc.2021.107082.

Han, H. and Hyun, S.S. (2015), "Customer retention in the medical tourism industry: impact of quality, satisfaction, trust, and price reasonableness", *Tourism Management*, Vol. 46, pp. 20-29, doi: 10.1016/j.tourman.2014.06.003.

Harrison-Walker, L. (2001), "E-complaining: a content analysis of an Internet complaint forum", *Journal of Services Marketing*, Vol. 15 No. 5, pp. 397-412, doi: 10.1108/EUM0000000005657.

Heinonen, K. and Strandvik, T. (2020), "Reframing service innovation: COVID-19 as a catalyst forimposed service innovation", *Journal of Service Management*, Vol. 32 No. 1, pp. 101-112, doi: 10.1108/JOSM-05-2020-0161.

Hirokawa, S. and Hashimoto, K. (2018), "Simplicity of positive reviews and diversity of negative reviews in hotel reputation", *International Joint Symposium on Artificial Intelligence and Natural Language Processing*, iSAI-NLP, Pattaya, pp. 1-6, doi: 10.1109/iSAI-NLP.2018.8692973.

Hoffmann, F., Braesemann, F. and Teubner, T. (2022), "Measuring sustainable tourism with online platform data", *EPJ Data Science*, Vol. 11 No. 1, doi: 10.1140/epjds/s13688-022-00354-6.

Hosmer, D.W. and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd ed., Wiley, Hoboke, NJ.

Jasinskas, E., Streimikiene, D., Svagzdiene, B. and Simanavicius, A. (2016), "Impact of hotel service quality on the loyalty of customers", *Economic Research-Ekonomska Istraživanja*, Vol. 29 No. 1, pp. 559-572, doi: 10.1080/1331677x.2016.1177465.

Kandampully, J. and Suhartanto, D. (2000), "Customer loyalty in the hotel industry: the role of customer satisfaction and image", *International Journal of Contemporary Hospitality Management*, Vol. 12 No. 6, pp. 346-351, doi: 10.1108/09596110010342559.

Kim, J.J., Lee, Y. and Han, H. (2019), "Exploring competitive hotel selection attributes among guests: an importance-performance analysis", *Journal of Travel and Tourism Marketing*, Vol. 36 No. 9, pp. 998-1011, doi: 10.1080/10548408.2019.1683484.

Kleinbaum, D.D. and Klein, M. (2002), *Logistic Regression; A Self-Learning Text*, Springer, New York.

Kotler, P., Kartajaya, H. and Setiawan, I. (2010), *Marketing 3.0*, doi: 10.1002/9781118257883.

Kunja, S.R., Kumar, A. and Rao, B.M. (2021), "Mediating role of hedonic and utilitarian brand attitude between eWOM and purchase intentions: a context of brand fan pages in Facebook", *Young Consumers: Insight and Ideas for Responsible Marketers*, Vol. 23 No. 1, pp. 1-15, doi: 10.1108/yc-11-2020-1261.

Ladhari, R. and Michaud, M. (2015), "eWOM effects on hotel booking intentions attitudes trust and Website perceptions", *International Journal of Hospitality Management*, Vol. 46, pp. 36-45, doi: 10.1016/j.ijhm.2015.01.010.

Li, H. and Sun, J. (2012), "Forecasting business failure: the use of nearest – neighbour support vectors and correcting imbalanced samples – evidence from the Chinese hotel industry", *Tourism Management*, Vol. 33 No. 3, pp. 622-634, doi: 10.1016/j.tourman.2011.07.004.

Li, G., Law, R., Rong, J. and Vu, H.Q. (2010), "Incorporating both positive and negative association rules into the analysis of Outbound tourism in Hong Kong", *Journal of Travel and Tourism Marketing*, Vol. 27 No. 8, pp. 812-828, doi: 10.1080/10548408.2010.527248.

Long, J.S. (1997), *Regression Models for Categorical and Limited Dependent Variables*, Sage, Thousand Oaks, CA.

Luong, D.B., Wu, K.-W. and Vo, T.H.G. (2020), "Difficulty is a possibility: turning service recovery into e-WOM", *Journal of Services Marketing*, Vol. 35 No. 8, pp. 1000-1012, doi: 10.1108/jsm-12-2019-0487.

Ma, Z. (2022), "Construction of tourism management engineering based on data mining technology", *Journal of Electrical and Computer Engineering*, Vol. 2022, pp. 1-11, doi: 10.1155/2022/1982462.

Malik, S.A., Akhtar, F., Raziq, M.M. and Ahmad, M. (2020), "Measuring service quality perceptions of customers in the hotel industry of Pakistan", *Total Quality Management and Business Excellence*, Vol. 31 Nos 3-4, pp. 263-278, doi: 10.1080/14783363.2018.1426451.

Mittal, V., Katrichis, J. and Kumar, P. (2001), "Attribute performance and customer satisfaction over time: evidence from two field studies", *Journal of Services Marketing*, Vol. 15 No. 5, pp. 343-356, doi: 10.1108/eum0000000005655.

Moliner, B., Gallarza, M.G., Gil, I. and Fuentes, M. (2015), "Causas y consecuencias sociales de la satisfacción de los clientes con hoteles", *Cuadernos de Turismo*, Vol. 36, p. 295, doi: 10.6018/turismo.36.231021.

Nguyen, H.T., Le, A.T.T., Phan, A.C. and Hoang, T.D.L. (2022), "A multi-perspective approach of international tourist satisfaction in tourism service: from big data perspective", *Journal of Asia Business Studies*, Vol. 17 No. 4, pp. 850-872, doi: 10.1108/JABS-03-2022-0090.

Nunkoo, R., Teeroovengadum, V., Ringle, C.M. and Sunnassee, V. (2019), "Service quality and customer satisfaction: the moderating effects of hotel star rating", *International Journal of Hospitality Management*, Vol. 91, p. 102414, 102414, doi: 10.1016/j.ijhm.2019.102414.

Oh, H. (1999), "Service quality, customer satisfaction, and customer value: a holistic perspective", *International Journal of Hospitality Management*, Vol. 18 No. 1, pp. 67-82, doi: 10.1016/s0278-4319(98)00047-4.

Oklevik, O., Nysveen, H. and Pedersen, P.E. (2018), "Influence of design on tourists' recommendation intention: an exploratory study of fjord cruise boats", *Journal of Travel and Tourism Marketing*, Vol. 35 No. 9, pp. 1187-1200, doi: 10.1080/10548408.2018.1487367.

Oliver, R.L. (1999), "Whence consumer loyalty?", *Journal of Marketing*, Vol. 63 No. 33, doi: 10.2307/1252099.

Pampel, F.C. (2000), *Logistic Regression: A Primer*, Sage, Thousand Oaks, CA.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, available at: www.R-project.org

Raza, M.A., Nabeel, A., Muhammad, H. and Bukhari, K. (2012), "Relationship between service quality, perceived value, satisfaction and revisit intention in hotel industry", *Interdisciplinary Journal of Contemporary Research in Business*, Vol. 4 No. 8, pp. 788-805.

Reichheld, F.F. (2003), "The one number you need to grow", *Harvard Business Review*, Vol. 81 No. 12, pp. 46-54.

Samara, D., Magnisalis, I. and Peristeras, V. (2020), "Artificial intelligence and big data in tourism: a systematic literature review", *Journal of Hospitality and Tourism Technology*, Vol. 11 No. 2, pp. 343-367, doi: 10.1108/jhtt-12-2018-0118.

Serra-Cantallops, A., Ramon-Cardona, J. and Salvi, F. (2018), "The impact of positive emotional experiences on eWOM generation and loyalty", *Spanish Journal of Marketing – ESIC*, Vol. 22 No. 2, pp. 142-162, doi: 10.1108/sjme-03-2018-0009.

Shin, H. and Kang, J. (2020), "Reducing perceived health risk to attract hotel customers in the COVID-19 pandemic era: focused on technology innovation for social distancing and cleanliness", *International Journal of Hospitality Management*, Vol. 91, p. 102664, doi: 10.1016/j.ijhm.2020.102664.

Sukhu, A., Choi, H., Bujisic, M. and Bilgihan, A. (2019), "Satisfaction and positive emotions: a comparison of the influence of hotel guests' beliefs and attitudes on their satisfaction and emotions", *International Journal of Hospitality Management*, Vol. 77, pp. 51-63, doi: 10.1016/j.ijhm.2018.06.013.

Terzić, A., Petrevska, B. and Demirović Bajrami, D. (2022), "Personalities shaping travel behaviors: post-COVID scenario", *Journal of Tourism Futures*, Vol. ahead-of-print No. ahead-of-print, doi: 10.1108/JTF-02-2022-0043.

Xu, Y.H., Li, H., Le, L.P. and Tian, X.Y. (2014), "Neighborhood triangular synthetic minority over-sampling technique for imbalanced prediction on small samples of Chinese tourism and hospitality firms", *2014 Seventh International Joint Conference on Computational Sciences and Optimization*, IEEE, pp. 534-538.

Yoon, Y. and Uysal, M. (2005), "An examination of the effects of motivation and satisfaction on destination loyalty: a structural model", *Tourism Management*, Vol. 25, pp. 45-56, doi: 10.1016/j.tourman.2003.08.016.

Zeithaml, V.A., Berry, L.L. and Parasuraman, A. (1996), "The behavioural consequences of service quality", *Journal of Marketing*, Vol. 60 No. 2, 31, doi: 10.2307/1251929.

## Author affiliations

Clara Martin-Duque is based at the Department of Business Organization, Complutense University of Madrid, Madrid, Spain.

Juan José Fernández-Muñoz is based at the Department of Psicology, Rey Juan Carlos University, Madrid, Spain.

Javier M. Moguerza is based at the Department of Computer Sciences and Statistics, Rey Juan Carlos University, Madrid, Spain.

Aurora Ruiz-Rua is based at the Department of Economic Theory and Mathematics Economy, UNED, Madrid, Spain.

## About the authors

Clara Martin-Duque hold a Ph.D. in Tourism and is a Professor of Business Organization at the Faculty of Commerce and Tourism of the Complutense University of Madrid. At present, she works as an external researcher in the Tourism Intelligence and Innovation Group of the Nebrija University (Smarttour-INN). Her research areas are linked both to the organization of tourism companies and to innovation processes in tourism companies and destinations. Clara Martin-Duque is the corresponding author and can be contacted at: cmartinduque@ucm.es

Juan José Fernández-Muñoz holds a Ph.D. in Social Psychology and is Associate Professor of Social Psychology at Rey Juan Carlos University. He has also been a visiting lecturer in Satakunta University in Finland. His main research interests are in the fields of Social Psychology: early retirement psychosocial processes, educational quality, self-efficacy and destination-image perception and methodology of behaviour sciences.

Javier M. Moguerza holds a Ph.D. in Mathematical Engineering and is full professor of Statistics at Rey Juan Carlos University. His professional expertise is focused on Computational and Applied Mathematics, Optimization, Six Sigma Quality, Data Mining and Pattern Recognition Methods. He has publications in highly ranked scientific journals such as *Information Fusion, Computational Management Science or Annals of Operations Research.*

Aurora Ruiz-Rua holds a PhD. in Economics and a M.A. in Industrial Economics and Regulated Markets is a lecturer of Introduction to Microeconomics at Universidad Nacional de Educación a Distancia (UNED), Spain. At present, she is part of the research group at UNED: Microeconomic Applications Research Group to companies and Househol. Implications for Individual Well-Being (Group Code: 110). Her main research interests are in the fields of industrial economics, individual well-being, tourism and transport.