Transactions on Data Analysis in Social Science 1(4) (2019) 171–185

ISSN Online: 2821-1936

Transactions on Data Analysis in Social Science

Journal Homepage: https://transoscience.ir

Quantitative Data Analysis of TripAdvisor Reviews for Hotels in Tehran

R. Khorsand[1,*], M. Rafiei[2], V. Keyvanfar[3]

[1] M.Sc. Student, Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran
[2] Assistant Professor, Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran
[3] Ph.D. in Industrial Engineering, Department of Industrial Engineering, Amirkabir University of Technology, Tehran, Iran

ARTICLE INFO

Article History:
Received 11 June 2019
Received in revised form 6 August 2019
Accepted 9 November 2019
Available online 1 December 2019

Keywords:
Big Data, Data Mining, Supply Chain, Tourism Industry

ABSTRACT

This study investigates hotel rating prediction using machine learning techniques, focusing on hotels in Tehran, the capital and largest city of Iran. Data were collected from *TripAdvisor.com*, the world's largest online travel and tourism platform. A total of 64 Tehran-based hotels with official TripAdvisor pages were identified, yielding 4,736 user reviews compiled from the earliest available entries. The primary aim of this research is to predict the ratings that new users may assign to Tehran's hotels based on both user profile characteristics and hotel attributes. To achieve this objective, eight supervised machine learning models were implemented using the R programming language: K-Nearest Neighbors (KNN), Naïve Bayes Classifier, Decision Tree, Logistic Regression, Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM). The models were comparatively analyzed to evaluate their predictive performance and to identify the most accurate approach. The findings of this study contribute to the growing field of predictive analytics in online review systems, demonstrating the potential of data-driven methods to enhance customer satisfaction, support managerial decision-making, and improve service quality within the hospitality and tourism industry.

1. INTRODUCTION

The tourism supply chain is defined as a network of tourism organizations that participate in various activities related to providing components of tourism products or services such as flights and accommodations and extend to the distribution and marketing of the final tourism product within a specific destination. This network involves numerous stakeholders across both the public and private sectors [1].

In this study, data related to the city of Tehran, Iran, were web-scraped from the TripAdvisor website using the Java programming language. These data include user reviews of Tehran hotels (collected from the first available review up to the end of April 2019), user profile features, and hotel-related information such as facilities and services.

* Corresponding Author: raminaakhorsand@gmail.com
M.Sc. Student, Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran

http://dx.doi.org/10.47176/TDASS.2019.171

© 2019 by the authors. Licensee T.D.A.S, Tehran, Iran. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).

In total, 4,736 records were extracted from 64 hotels in Tehran that have official pages on TripAdvisor. The dataset was then used to predict new user ratings based on the characteristics of user profiles and the attributes of hotels. To achieve this, eight supervised machine learning models were applied:

- K-Nearest Neighbors (KNN)
- Naïve Bayes Classifier
- Decision Tree
- Logistic Regression
- Artificial Neural Network (ANN)
- Support Vector Machine (SVM)
- Random Forest (RF)
- Gradient Boosting Machine (GBM)

Finally, the predictive performance of all eight models was compared in detail to identify the most accurate model for user rating prediction.

2. LITERATURE REVIEW

With the widespread development of the Internet, social networks have created an extensive platform for tourists to share various types of tourism-related information, such as travel experiences and personal reviews. For instance, travelers can express their satisfaction or dissatisfaction with tourism products, thereby generating a rich repository of online review data. Tourists also share their experiences through blogs and microblogs such as Twitter, offering valuable insights for potential travelers. These online reviews, blogs, and text-based datasets represent a unique form of big data in tourism research encompassing textual data mining, sentiment transfer, and mood analysis among travelers.

Online data used in tourism studies typically fall into two categories: review data and blog data, each serving distinct research purposes.

- Review data mainly reflect tourists' attitudes toward tourism products and are widely used to assess customer satisfaction. Several studies have explored key features of tourist satisfaction [2-5] and examined its relationship with other factors such as guest experience and competitive positioning [6-7].

- The most frequently studied review subjects include hotels (including rural accommodations), restaurants, and tourist attractions. Hotel reviews are often used to assess and enhance electronic word-of-mouth (eWOM) [2,7,14]; restaurant reviews to evaluate tourist satisfaction [14-15]; and attraction reviews to improve destination management [16-17].

The popularity of these three categories accommodation, dining, and travel activities can be attributed to their central importance in shaping travelers' overall experiences. The present research offers clear implications for the tourism, hospitality, and marketing industries, helping them attract a larger number of tourists through data-driven insights.

Blog data, which capture travel stories and emotional narratives, have primarily been used in tourism recommendation systems and sentiment analysis [18]. For instance, [18] employed travel blogs to explore tourist destinations and travel sequences for optimized itinerary planning. Similarly, [19] extracted valuable information from blog data about where to go and what to do during trips. In terms of sentiment analysis, studies such as [20], [21] utilized Twitter data to explore tourist emotions and sentiment patterns.

Regarding data sources, online review data in tourism research have generally been collected from diverse social media platforms, including TripAdvisor [3], [9], [10], [16], [22-25], Yelp [15], [26], Expedia [7], Ctrip [27], Qunar [28], Booking [5], and Dianping [29]. Among these, TripAdvisor, as one of the largest global tourism social networks, is the most widely used source.

For blog-based data, Twitter and Sina Weibo serve as the two main platforms. For example, [20], [30], and [31] used Twitter data to extract geographical and sentiment information from tourists, while [32] employed Sina Weibo the Chinese equivalent of Twitter to identify potential tourist areas, the lifespan of travel news, and public attitudes toward travel policy changes.

The sample sizes used in previous studies vary considerably, ranging from 373 records [4] to 412,784 records [3], depending on the research scope and the type of textual data analyzed.

To extract and utilize the valuable hidden information embedded in online textual data, various text-mining techniques have been extensively applied in tourism research. These techniques generally follow two main stages: data collection and data mining, with the latter consisting of two sub-stages data preprocessing and pattern discovery.

The first step, data collection, involves gathering online textual data (including tourism-related reviews and blogs) from relevant social networking sites through web crawling technology [7], [13], [19]. Specifically, a *web crawler* (also referred to as a *bot* or *spider*) is a software program or a set of programs that automatically and recursively downloads web pages, extracts URLs from the *HyperText Markup Language (HTML)*, and retrieves them for analysis. For instance, [13] employed a web crawler written in Python and Java to collect hotel-related reviews. Similarly, [2] developed a web crawler to periodically gather review data from TripAdvisor, while [18] utilized this technology to obtain user-generated blogs from travel websites.

The second step, data mining, focuses on analyzing the collected online textual data to extract useful knowledge for tourism research. This process involves two stages: data preprocessing and pattern discovery. In data preprocessing, various techniques are applied depending on research objectives. Common operations include data cleaning, tokenization, stemming, and part-of-speech (POS) tagging, as frequently adopted in existing tourism literature utilizing online textual data.

Data cleaning aims to identify and remove incorrect or irrelevant records from online textual datasets such as spelling errors [7], *stop words* [5], [7], [19], non-target languages, and low-frequency words [2] to ensure the extraction of valuable information in the tourism domain.

Tokenization refers to breaking down travel-related textual information into words, phrases, or other meaningful linguistic units. Through this process, tourism-related keywords concerning destinations, travel sentiments, and visitor experiences can be filtered from large text corpora [2], [5], [13].

Stemming is employed to identify the root forms of words, treating all words sharing the same root as a single *token*, thereby simplifying the modeling process [5].

Part-of-speech tagging involves labeling each word in a sentence according to its grammatical category such as noun, adjective, or adverb. For example, since hotel reviews are predominantly expressed using nouns, adjectives, and negative adverbs, uninformative words and other labels can be excluded from the analysis [2], [9].

Pattern discovery represents another crucial step in text data mining, aiming to uncover meaningful information within textual documents. Typical techniques applied in tourism research include Latent Dirichlet Allocation (LDA), sentiment analysis, statistical analysis, clustering and classification, text summarization, and dependency modeling.

Latent Dirichlet Allocation (LDA) is a probabilistic model used to identify latent topics within textual data. For example, [2] employed LDA to rapidly uncover composite topics such as factors influencing hotel customer satisfaction from a large volume of online reviews.

Sentiment analysis is utilized to determine tourists' attitudes toward tourism products or attractions by categorizing textual data into positive, negative, or neutral emotional classes. Recent studies have increasingly adopted sentiment analysis as a valuable tool for examining travelers' perceptions of hotel services [20] and popular destinations [19].

Statistical analysis represents the most fundamental technique for examining various forms of data, including textual information. In tourism studies, descriptive statistics (e.g., mean, variance) [26], t-tests [33], correlation matrices [26], Kruskal–Wallis tests, Mann–Whitney tests [4], and correspondence analysis [34] have been widely

applied to describe diverse information within online textual data, such as reviewer identity disclosure, review credibility, and hotel rankings.

Clustering and classification are employed to group sets of objects so that items within the same cluster are more similar to each other than to those in other clusters. For instance, [31] used clustering techniques to group travel trajectories whose geolocation sequences belonged to similar classes based on geotagged messages on Twitter. Likewise, [35] applied classification techniques to categorize hidden information in travel blogs and hotel reviews about Barcelona into meaningful groups according to keyword distributions.

Text summarization is employed to automatically generate concise summaries of one or more documents, thereby extracting key information from large textual datasets. [9] used a multi-document summarization method to identify the most informative sentences among hotel reviews.

Dependency modeling seeks to establish relationships between textual data (particularly online reviews) and tourism-related factors such as hotel performance [24], [36], restaurant performance [29], and tourist behaviors [14]. Various regression models, including Bayesian logistic regression [14], linear regression [24], [29], [36], and Tobit regression [16], have been employed for this purpose.

It is noteworthy that numerous powerful data-mining tools and software packages have been developed specifically for text processing. General-purpose tools include WEKA, LingPipe, and TextBlob. Specifically, WEKA, developed by the University of Waikato, provides comprehensive solutions for text processing challenges such as data cleaning, word frequency analysis, clustering, classification, and pattern mining. LingPipe, developed by Alias, is a robust toolkit capable of performing textual data cleaning, tokenization, part-of-speech (POS) tagging, sentiment analysis, clustering, classification, and pattern mining. TextBlob, a Python library, offers an application programming interface (API) for common natural language processing tasks such as POS tagging, noun-phrase extraction, sentiment analysis, classification, and translation.

In addition, several specialized tools with specific functionalities are also available, such as those developed by the Institute of Computational Technology, the Chinese Word Analysis System, and Jieba for Chinese tokenization and POS tagging, as well as the Stanford Log-linear POS Tagger for part-of-speech tagging [9].

Ultimately, the insightful information extracted through data-mining techniques can be transformed into valuable knowledge to further advance tourism research. According to related studies, such knowledge encompasses aspects such as tourist satisfaction structures [14], [37], hotel preferences [22], tourist areas [38], tourism routes [30], and review characteristics [13]. This kind of knowledge has proven instrumental in enhancing tourism management and developing effective tourism recommendation systems.

3. RESEARCH ON TRIPADVISOR DATA

Studies focusing on TripAdvisor data can generally be classified into two major categories: quantitative studies and textual studies.

The first category, quantitative studies, emphasizes the numerical analysis of user-generated ratings. Overall, online reviews are considered much more significant than traditional feedback methods in reflecting hotel performance indicators [39]. Various methods have been applied to TripAdvisor user reviews. For instance, data mining and decision tree algorithms have been used to group customers across different hotels [40]. Another study employed data-mining techniques to understand the behavior and needs of potential hotel customers [41]. Methods such as AutoRegressive Integrated Moving Average (ARIMA) and Support Vector Machine (SVM) have also been implemented to analyze online reviews within the hospitality industry. Indeed, many studies utilizing data mining in tourism still rely primarily on ARIMA models [42].

However, the optimal technique for predicting potential customer ratings has not yet been identified. Determining the best method requires extensive research using diverse datasets, as rating prediction is a long-term and evolving process. Addressing this research gap is one of the main objectives of the present study, which compares multiple machine learning models applied to quantitative data extracted from TripAdvisor to identify the most accurate predictive model.

Nevertheless, this study alone is not sufficient and further research is needed. Firstly, the models employed are limited in number, and secondly, they were applied only to hotel data from two cities. Hence, future studies should extend these analyses to a broader range of cities and additional machine learning models to determine the most accurate rating prediction approach.

The second category, textual studies, employs text-mining and sentiment analysis techniques. The number of published papers in this category far exceeds that of quantitative studies. However, text mining faces several inherent limitations. The first limitation pertains to language diversity, as user reviews are written in multiple languages, making multilingual analysis challenging. Most studies analyze English-language texts, which cannot fully capture the diversity of global user experiences. The second limitation is the incomplete understanding of textual content through current text-mining methods. Another challenge involves errors and noise within user-generated reviews an issue particularly pronounced in countries where English is not the primary language [42].

As mentioned earlier, the number of studies that analyze quantitative TripAdvisor data remains limited, despite such data being more structured, accessible, and interpretable. Identifying users likely to assign low ratings is particularly valuable, as responding to user reviews especially in large-scale businesses is time-consuming [43]. For example, [42] applied the Support Vector Machine algorithm to predict potential hotel user ratings, followed by a data-driven sensitivity analysis to identify the most influential features within the predictive model.

Although such studies are insightful, many other machine learning techniques remain unexplored for analyzing user ratings. Employing diverse learning algorithms enables researchers to identify the most accurate rating prediction models and improve overall prediction performance. On TripAdvisor, users can rate hotels numerically, thereby generating valuable data resources [44]. However, few studies have examined complete city-level datasets or compared outputs across multiple cities. In many prior works, datasets were selectively sampled for instance, by focusing only on 4- and 5-star hotels or analyzing a limited number of user reviews per hotel.

To address this research gap, the present study utilizes real-world TripAdvisor data related to Tehran, collected from the date the city first established an official page on the platform. This approach ensures more comprehensive and representative findings and provides practical insights for managers of all Tehran hotels. Since the main objective of this study is to predict potential user ratings, eight different machine learning algorithms were applied to Tehran's dataset to not only forecast new user ratings but also identify the best-performing model among them. These eight models include:

- *k*-Nearest Neighbors (KNN)
- Naïve Bayes Classifier
- Decision Tree
- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest
- Gradient Boosting Machine (GBM)

4. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

The recent boom in social networking platforms has sparked growing interest in utilizing large-scale online textual data such as reviews and blogs for tourism research. Nevertheless, significant opportunities remain for expanding and deepening such emerging research.

In terms of sample size, most existing studies have relied on relatively small datasets, which are prone to sampling bias and estimation bias, often resulting in inaccurate analytical outcomes inconsistent with real-world patterns. Therefore, future research should employ larger-scale datasets to ensure more generalizable results.

Regarding research focus, previous work has mainly centered on extracting valuable knowledge from large-scale online textual data for instance, studying tourist sentiments or opinions about products or destinations. However, relatively few studies have explored how such insights can be practically applied to tourism product design or marketing strategies.

Moreover, other relevant issues such as data reliability also deserve attention. For example, travelers may occasionally post fake positive reviews to avoid potential conflicts or obtain refunds [23]. Such online behavioral patterns among tourists represent promising directions for future in-depth research.

Given the unstructured and complex nature of online reviews and travel-related blogs, text-processing and analysis methods still face significant challenges. Most current studies focus on a single language, often excluding or ignoring non-target languages [2]. However, travelers from different cultural backgrounds pursue distinct travel interests and exhibit varying preferences [33]. Consequently, developing robust analytical methods for multilingual texts is essential to avoid information loss.

In addition, in sentiment analysis of tourism reviews, nouns, adjectives, and negative adverbs are commonly regarded as key indicators. Yet, sentiment is also shaped by the polarity of verbs (e.g., "like," "hate," "love") and intensity modifiers (e.g., "very," "more," "too much") [9]. Such nuanced linguistic information should therefore be incorporated into future sentiment analysis frameworks for tourism research.

5. RESEARCH METHODOLOGY AND PROBLEM-SOLVING APPROACH

Initially, the data were scraped from the TripAdvisor website, and after the data cleaning process, the refined dataset was stored in the final database in three separate tables. The data were then examined in a preliminary, descriptive manner to gain an overall understanding of their characteristics. Subsequently, eight supervised machine learning models were trained and evaluated. The results obtained from the model evaluation phase were compared using common accuracy assessment metrics in machine learning to identify the best predictive model for estimating potential travelers' ratings of hotels in Tehran.

5.1. Data Collection

In this study, Tehran, the capital city of Iran and a major destination for tourists from around the world, was selected as the case city. The data were gathered from TripAdvisor.com, the world's largest travel platform that hosts more than 730 million customer reviews [45]. Information related to all 64 hotels in Tehran listed on the website was collected using the Java programming language and stored in structured tables within a MySQL database.

Among the extensive and valuable information provided by TripAdvisor, three main data categories were selected for analysis:

1. **Hotel Information:** This includes general details such as the hotel name, average customer rating (ranging from 1 to 5), total number of reviews, the hotel's ranking among others in the same city, and its class (number of stars).
2. **Hotel Facilities:** Facilities such as in-room service, restaurant, swimming pool and massage, free breakfast, airport transfer, concierge service, laundry, multilingual staff, sauna, free parking, bar, high-speed Wi-Fi, fitness center, banquet room, business center with internet access, conference facilities, jacuzzi, meeting rooms, public Wi-Fi, and wheelchair accessibility were extracted. These were encoded in a **binary format (0 or 1)** and stored in a separate table.
3. **Reviewer Information:** This category contains data related to the users who posted reviews on hotel pages, including the review date, country or city of origin, level of contribution, rating given to the hotel (1–5), review text, date of stay, type of trip, and year of joining TripAdvisor.

A sample of the extracted data from one hotel and one corresponding review is presented in Figure 1. Each of these data categories was stored in a separate table in MySQL and linked via the hotel ID, a unique numerical identifier.

In total, the 64 hotels generated 4,736 reviews, on which machine learning and data mining processes were conducted using the R programming language.

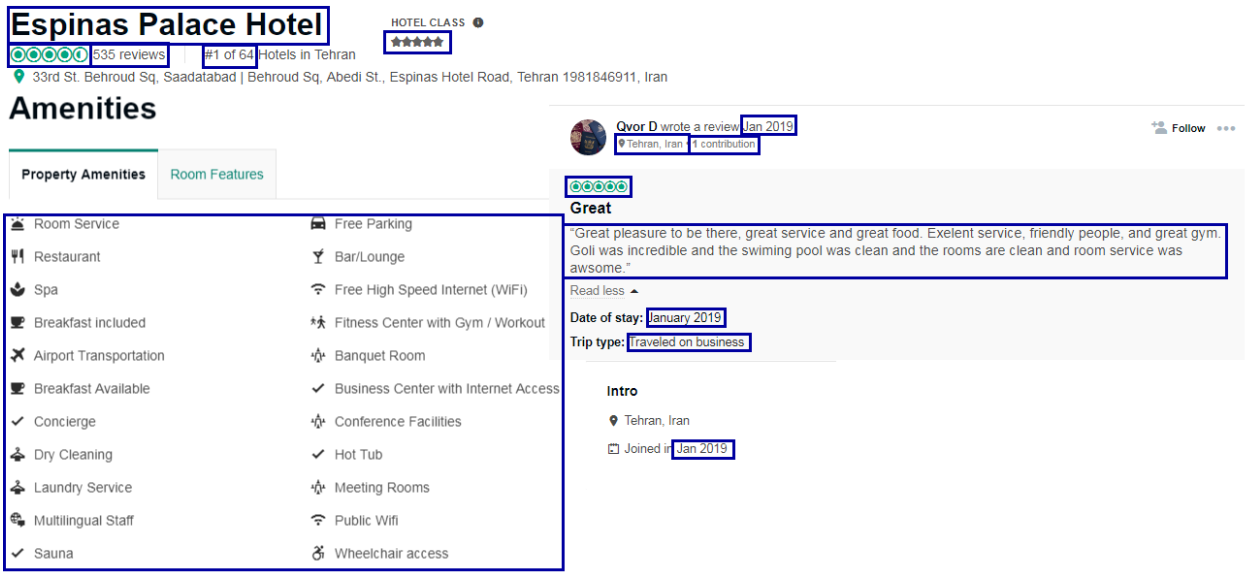


Fig. 1. Data extraction from one sample hotel used in the study

5.2. Data Cleaning

After data collection, a data cleaning phase was conducted to identify and eliminate noise, missing values, outliers, incorrect data, and other inconsistencies. On the TripAdvisor website, the country of origin of a tourist is not selected from predefined options; rather, users manually enter this information. This led to numerous input errors in that field, and therefore, this feature was excluded from the analysis.

The type of trip (e.g., business, leisure, family) is also an optional field on TripAdvisor, and some users chose not to specify it. However, because these data are highly valuable and selected from a fixed set of options, we decided not to remove them unless they introduced measurable bias. In the dataset used for this study, 8.7% of reviews lacked a specified trip type. As this amount of missing data is considered negligible [46], those missing values were imputed using the mode substitution method, where missing entries are replaced by the most frequently occurring category.

Some variables were also transformed to improve interpretability and analytical relevance. The user's registration year on TripAdvisor was subtracted from 2019 to represent the number of years of membership on the platform. Additionally, both the stay date and the review date were converted into month-based formats to facilitate temporal analysis.

An extreme outlier was also identified in the user participation level variable, where one record showed a participation count of 110,134,448, which was far beyond the next highest value (6,313,161). This instance was therefore recognized as an outlier and removed from the dataset.

5.3. Preliminary Examination of Quantitative Data

After cleaning and preprocessing, the final dataset for Tehran contained 28 features and 4,718 records (tuples). Figure 2 presents the distribution of user ratings. The first set of numbers in the chart indicates the rating values, and the second set represents the percentage of each rating relative to the total number of reviews for hotels in Tehran.

As illustrated, the majority of users rated Tehran hotels with 4 or 5 stars, meaning that approximately 70% of reviewers on TripAdvisor awarded these top ratings. This suggests an overall positive perception of hotel services and facilities among visitors to Tehran.

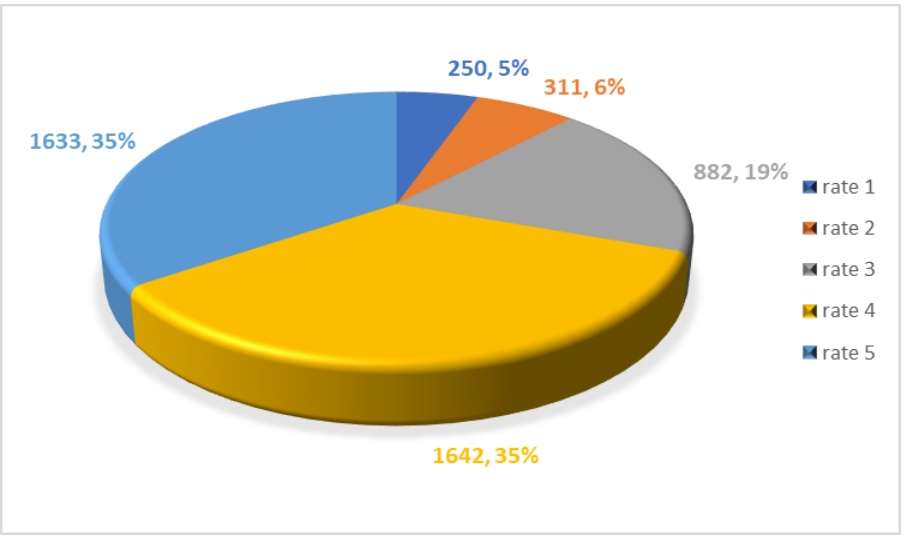
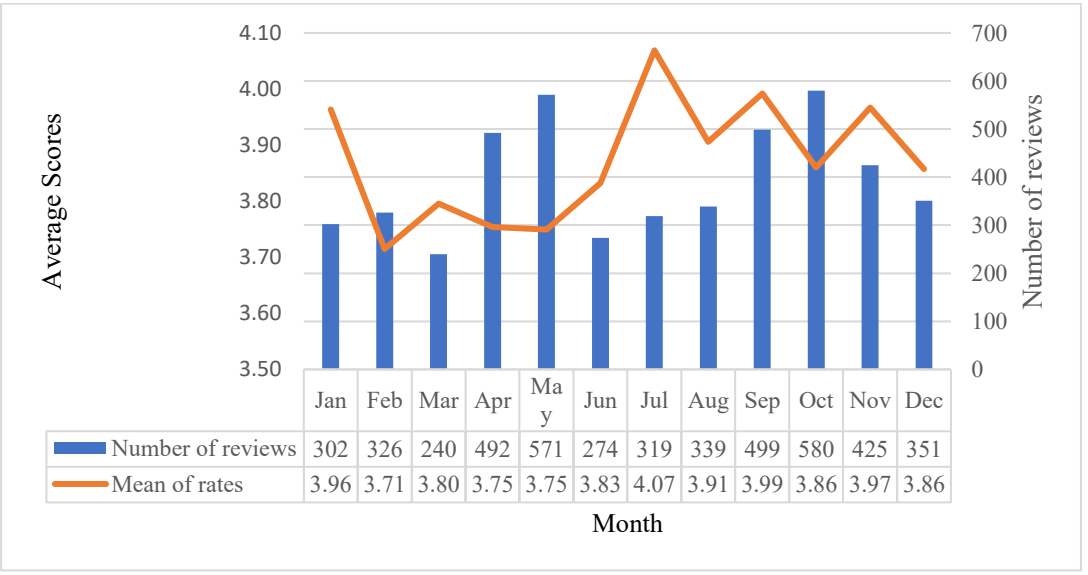


Fig. 2. Overall evaluation of Tehran hotels based on user ratings

Figure 3 illustrates the number of travelers by month along with their average ratings for hotels in Tehran. As shown, October and May are the most popular months for visiting Tehran, with 580 and 571 reviews, respectively. This popularity is likely due to the pleasant weather during these periods, which makes traveling more favorable.

However, a noteworthy finding from Figure 4 is that although these months attract the highest number of visitors, they do not correspond to the highest satisfaction levels. Specifically, May recorded one of the lowest average ratings (3.75) among all months. This indicates a potential gap between tourist volume and service satisfaction during high-demand seasons.

Therefore, hotel managers in Tehran should consider developing strategies to enhance service quality and guest satisfaction, particularly during May and October, when demand peaks. Conversely, July shows the highest average rating among all months, suggesting that travelers during this period are generally more satisfied with their hotel experiences. Hotel managers could analyze the factors contributing to this success and apply similar practices to other months to achieve a more consistent level of customer satisfaction throughout the year.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Number of reviews	302	326	240	492	571	274	319	339	499	580	425	351
Mean of rates	3.96	3.71	3.80	3.75	3.75	3.83	4.07	3.91	3.99	3.86	3.97	3.86

Fig. 3. Number of travelers to Tehran and their average ratings by month

6. RESEARCH FINDINGS

After preliminary cleaning and examination of the data, the next step was data segmentation. The final dataset was split into training and evaluation (or test) sets with a ratio of 70% to 30%, respectively. The training data were used to train the models, while the evaluation data were employed to validate the models' performance. Specifically, each supervised machine learning model was first trained on the training data to learn the patterns, and then the trained models were applied to the evaluation data to assess their accuracy on unseen data.

As previously mentioned, the objective of this study is to predict a new user's rating for a hotel based on the hotel's features and the user's TripAdvisor profile. Since ratings on this website are numerical values ranging from 1 to 5, supervised machine learning models are required. In this study, eight supervised machine learning models k-Nearest Neighbors (k-NN), Naive Bayes, Decision Tree, Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and Neural Network were trained on 70% of the final dataset and then evaluated on the remaining 30%, which contained entirely unseen data for these models.

Table 1 presents a comparison of these eight machine learning models based on different performance metrics for the city of Tehran. The calculation methods for the four metrics in Table 1 are provided in Equations (1)–(4), all of which are derived from the confusion matrix.

Accuracy is the simplest metric to evaluate a model's correctness, and its calculation is shown in Equation (1). As shown in Table 1, the k-Nearest Neighbors model achieves the highest accuracy among all methods. Following that, Random Forest, Decision Tree, and Gradient Boosting models exhibit the next highest accuracy levels for the Tehran dataset. These three models are fundamentally based on the Decision Tree approach, while Random Forest and Gradient Boosting are designed to enhance the efficiency and accuracy of the Decision Tree method.

$$\text{Accuracy} = \frac{\text{Number of correctly predicted instances by the model}}{\text{Total number of evaluation instances}} \times 100 \quad (1)$$

Table 1. Comparison of error metrics for different machine learning models on Tehran data

Model	Accuracy	MAE	MAPE	MSE
k-Nearest Neighbor	98.87%	0.011	0.49%	0.011
Naive Bayes	41.77%	0.949	45.29%	2.050
Decision Tree	49.29%	0.732	35.40%	1.407
Logistic Regression	47.45%	0.741	33.43%	1.351
Support Vector Machine	46.31%	0.745	33.31%	1.314
Neural Network	46.38%	0.783	37.00%	1.481
Random Forest	49.50%	0.699	32.80%	1.254
Gradient Boosting	49.29%	0.732	35.40%	1.407

Mean Absolute Error (MAE) represents the average prediction error of the model and is calculated using Equation (2), where $n$is the total number of observations, $y_j$is the actual observed value, and $\hat{y}_j$is the value predicted by the model.

$$\text{MAE}=\frac{1}{n}\sum_{j=1}^{n}\left|y_j-\hat{y}_j\right| \quad (2)$$

Mean Absolute Percentage Error (MAPE), which is an indicator for comparing the predictive accuracy of machine learning models, expresses the difference between the actual and predicted values as a percentage. The formula for calculating MAPE is presented in Equation (3).

$$MAPE=\frac{100}{n}\sum_{j=1}^{n}\left|\frac{y_j-\hat{y}_j}{y_j}\right| \tag{3}$$

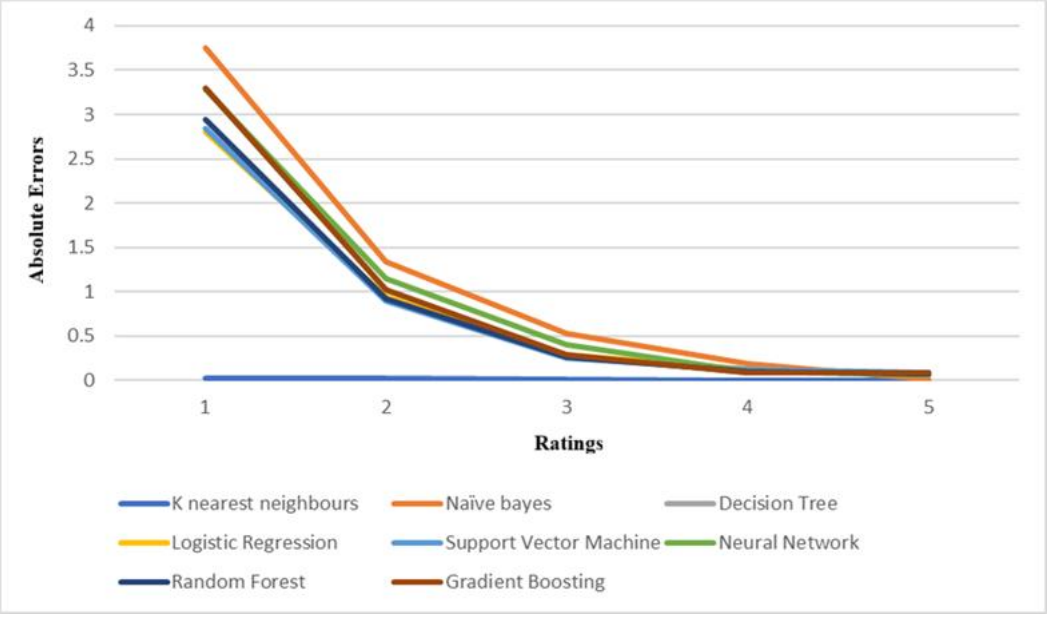


Fig. 4. Absolute Errors of Ratings Predicted by Machine Learning Models for Hotels in Tehran

Mean Squared Error (MSE) represents the average squared difference between the actual observed values and the values predicted by the model, as expressed in Equation (4).

$$MSE=\frac{1}{n}\sum_{j=1}^{n}\left(y_j-\hat{y}_j\right)^2 \tag{4}$$

All three of these methods were used to calculate prediction errors, with lower values indicating better performance. As shown in Table 1, the KNN method exhibited the lowest error among all methods. Figure 4 illustrates the absolute errors for each rating from 1 to 5 across the eight machine learning models applied in this study. As can also be observed from this figure, KNN demonstrates lower prediction errors. An interesting insight from this figure is that all models performed particularly well in predicting higher ratings, especially ratings of 4 and 5. This outcome can be attributed to two possible reasons. First, as seen in Figure 3, the number of ratings of 4 and 5 is considerably higher than that of ratings 1 to 3 (accounting for 70% of all Tehran data). A larger number of observations likely enables the models to learn more effectively, resulting in better evaluation performance. Second, lower ratings may have been assigned with specific intent, potentially leading to biased or skewed ratings.

7. CONCLUSION AND RECOMMENDATIONS

In this study, data regarding hotels in Tehran were collected from TripAdvisor. General information for 64 hotels in Tehran, including their amenities and the characteristics of reviewers' profiles, was scraped using the Java programming language from the platform's launch until April 2019 and stored in three separate tables in MySQL.

After data cleaning, a total of 28 features and 4,718 rows remained. The dataset was then split into 70% training and 30% evaluation subsets to ensure that the models could be validated on entirely new data.

Eight supervised machine learning models k-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree, Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting Machine (GBM) were trained on the training data. These trained models were then validated using the evaluation data, and their prediction accuracy and errors were compared.

For future studies, the following recommendations are proposed:

1. Use datasets from different sources, such as other cities in Iran or cities in other countries.
2. Explore additional machine learning methods to predict potential travelers' ratings.
3. Combine machine learning with text mining techniques to extract more information from customer reviews.
4. Employ additional comparative evaluation methods to validate the top-performing model (e.g., Receiver Operating Characteristic (ROC) analysis and Precision-Recall metrics).

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Zhang, X., Song, H., & Huang, G. Q. (2009). Tourism supply chain management: A new research agenda. Tourism Management, 30(3), 345–358. https://doi.org/10.1016/j.tourman.2008.12.010

[2] Guo, Y., Barnes, S. J., & Jia, Q. (2017). No title. Tourism Management, 59, 467–483. https://doi.org/10.1016/j.tourman.2016.09.009

[3] Liu, Y., Teichert, T., Rossi, M., Li, H., & Hu, F. (2017). Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews. Tourism Management, 59, 554–563. https://doi.org/10.1016/j.tourman.2016.08.012

[4] Lu, W., & Stepchenkova, S. (2012). Ecotourism experiences reported online: Classification of satisfaction attributes. Tourism Management, 33(3), 702–712. https://doi.org/10.1016/j.tourman.2011.08.003

[5] Xu, X., & Li, Y. (2016). The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. International Journal of Hospitality Management, 55, 57–69. https://doi.org/10.1016/j.ijhm.2016.03.003

[6] Crotts, J. C., Mason, P. R., & Davis, B. (2009). Measuring guest satisfaction and competitive position in the hospitality and tourism industry. Journal of Travel Research, 48(2), 139–151.

https://doi.org/10.1177/0047287508328795

[7] Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? International Journal of Hospitality Management, 44, 120–130. https://doi.org/10.1016/j.ijhm.2014.10.013

[8] Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews. Journal of Hospitality Marketing & Management, 25(1), 1–24. https://doi.org/10.1080/19368623.2015.983631

[9] Hu, Y.-H., Chen, Y.-L., & Chou, H.-L. (2017). Opinion mining from online hotel reviews: A text summarization approach. Information Processing & Management, 53(2), 436–449. https://doi.org/10.1016/j.ipm.2016.12.002

[10] Ma, J., Luo, S., Yao, J., Cheng, S., & Chen, X. (2016). Efficient opinion summarization on comments with online-LDA. International Journal of Computers, Communications & Control, 11(3), 414–427. https://doi.org/10.15837/ijccc.2016.3.700

[11] Melián-González, S., Bulchand-Gidumal, J., & González López-Valcárcel, B. (2013). Online customer reviews of hotels. Cornell Hospitality Quarterly, 54(3), 274–283. https://doi.org/10.1177/1938965513481498

[12] Phillips, P., Zigan, K., Santos Silva, M. M., & Schegg, R. (2015). The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. Tourism Management, 50, 130–141. https://doi.org/10.1016/j.tourman.2015.01.028

[13] Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management, 58, 51–65. https://doi.org/10.1016/j.tourman.2016.10.001

[14] Zhang, Y., & Cole, S. T. (2016). Dimensions of lodging guest satisfaction among guests with mobility challenges: A mixed-method analysis of web-based texts. Tourism Management, 53, 13–27. https://doi.org/10.1016/j.tourman.2015.09.001

[15] Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. Annals of Tourism Research, 50, 67–83. https://doi.org/10.1016/j.annals.2014.10.007

[16] Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. Tourism Management, 52, 498–506. https://doi.org/10.1016/j.tourman.2015.07.018

[17] Pearce, P. L., & Wu, M.-Y. (2018). Entertaining international tourists: An empirical study of an iconic site in China. Journal of Hospitality & Tourism Research, 42(5), 772–792. https://doi.org/10.1177/1096348015598202

[18] Yuan, H., Xu, H., Qian, Y., & Li, Y. (2016). Make your travel smarter: Summarizing urban tourism information from massive blog data. International Journal of Information Management, 36(6), 1306–1319. https://doi.org/10.1016/j.ijinfomgt.2016.02.009

[19] Xu, H., Yuan, H., Ma, B., & Qian, Y. (2015). Where to go and what to play: Towards summarizing popular information from massive tourism blogs. Journal of Information Science, 41(6), 830–854. https://doi.org/10.1177/0165551515603323

[20] Philander, K., & Zhong, Y. Y. (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. International Journal of Hospitality Management, 55, 16–24.

https://doi.org/10.1016/j.ijhm.2016.02.001

[21] Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of Twitter posts. Expert Systems with Applications, 40(10), 4065–4074. https://doi.org/10.1016/j.eswa.2013.01.001

[22] Li, G., Law, R., Vu, H. Q., Rong, J., & Zhao, X. (Roy). (2015). Identifying emerging hotel preferences using emerging pattern mining technique. Tourism Management, 46, 311–321. https://doi.org/10.1016/j.tourman.2014.06.015

[23] Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. Journal of Travel & Tourism Marketing, 32(5), 608–621. https://doi.org/10.1080/10548408.2014.933154

[24] Xie, K. L., Zhang, Z., & Zhang, Z. (2014). The business value of online consumer reviews and management response to hotel performance. International Journal of Hospitality Management, 43, 1–12. https://doi.org/10.1016/j.ijhm.2014.07.007

[25] [25] Ye, Q., Li, H., Wang, Z., & Law, R. (2014). The influence of hotel price on perceived service quality and value in e-tourism. Journal of Hospitality & Tourism Research, 38(1), 23–39. https://doi.org/10.1177/1096348012442540

[26] Racherla, P., & Friske, W. (2012). Perceived "usefulness" of online consumer reviews: An exploratory investigation across three services categories. Electronic Commerce Research and Applications, 11(6), 548–559. https://doi.org/10.1016/j.elerap.2012.06.003

[27] Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. International Journal of Hospitality Management, 28(1), 180–182. https://doi.org/10.1016/j.ijhm.2008.06.011

[28] Zhang, Z., Zhang, Z., & Yang, Y. (2016). The power of expert identity: How website-recognized expert reviews influence travelers' online rating behavior. Tourism Management, 55, 15–24. https://doi.org/10.1016/j.tourman.2016.01.004

[29] Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. International Journal of Hospitality Management, 29(4), 694–700. https://doi.org/10.1016/j.ijhm.2010.02.002

[30] Chua, A., Servillo, L., Marcheggiani, E., & Vande Moere, A. (2016). Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. Tourism Management, 57, 295–310. https://doi.org/10.1016/j.tourman.2016.06.013

[31] Bordogna, G., Frigerio, L., Cuzzocrea, A., & Psaila, G. (2016). Clustering geo-tagged tweets for advanced big data analytics. In Proceedings - 2016 IEEE International Congress on Big Data, BigData Congress 2016 (pp. 42–51). https://doi.org/10.1109/BigDataCongress.2016.78

[32] Cheng, M., & Edwards, D. (2015). Social media in tourism: A visual analytic approach. Current Issues in Tourism, 18(11), 1080–1087. https://doi.org/10.1080/13683500.2015.1036009

[33] Schuckert, M., Liu, X., & Law, R. (2015). A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently. International Journal of Hospitality Management, 48, 143–149. https://doi.org/10.1016/j.ijhm.2014.12.007

[34] Költringer, C., & Dickinger, A. (2015). Analyzing destination branding and image from online sources: A web content mining approach. Journal of Business Research, 68(9), 1836–1843.

https://doi.org/10.1016/j.jbusres.2015.01.011

[35] Marine-Roig, E., & Anton Clavé, S. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. Journal of Destination Marketing & Management, 4(3), 162–172. https://doi.org/10.1016/j.jdmm.2015.06.004

[36] Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications, 36(3), 6527–6535. https://doi.org/10.1016/j.eswa.2008.07.035

[37] Peng, G., Liu, Y., Wang, J., & Gu, J. (2017). Analysis of the prediction capability of web search data based on the HE-TDC method - prediction of the volume of daily tourism visitors. Journal of System Science and Systems Engineering, 26(2), 163–182. https://doi.org/10.1007/s11518-016-5311-7

[38] Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations - A case from Sweden. Journal of Destination Marketing & Management, 3(4), 198–209. https://doi.org/10.1016/j.jdmm.2014.08.002

[39] Kim, W. G., & Park, S. A. (2017). Social media review rating versus traditional customer satisfaction. International Journal of Contemporary Hospitality Management, 29(2), 784–802. https://doi.org/10.1108/IJCHM-11-2015-0627

[40] Min, H., Min, H., & Emam, A. (2002). A data mining approach to developing the profiles of hotel customers. International Journal of Contemporary Hospitality Management, 14(6), 274–285. https://doi.org/10.1108/09596110210436814

[41] Magnini, V. P., Honeycutt, E. D., & Hodge, S. K. (2003). Data mining for hotel firms: Use and limitations. Cornell Hotel and Restaurant Administration Quarterly, 44(2), 94–105. https://doi.org/10.1016/S0010-8804(03)90022-X

[42] 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. https://doi.org/10.1016/j.tmp.2017.04.003

[43] Nguyen, K. A., & Coudounaris, D. N. (2015). The mechanism of online review management: A qualitative study. Tourism Management Perspectives, 16, 163–175. https://doi.org/10.1016/j.tmp.2015.08.002

[44] O'Connor, P. (2010). Managing a hotel's image on Tripadvisor. Journal of Hospitality Marketing & Management, 19(7), 754–772. https://doi.org/10.1080/19368623.2010.508007

[45] TripAdvisor. (n.d.). Investor relations. http://ir.tripadvisor.com/investor-relations

[46] Bennett, D. A. (2001). How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health, 25(5), 464–469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

[47] Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1–17. https://doi.org/10.1016/j.ins.2012.10.039

[48] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

[49] Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

[50] Tinoco, J., Gomes Correia, A., & Cortez, P. (2011). Application of data mining techniques in the estimation of the uniaxial compressive strength of jet grouting columns over time. Construction and Building Materials, 25(3), 1257–1262. https://doi.org/10.1016/j.conbuildmat.2010.09.027