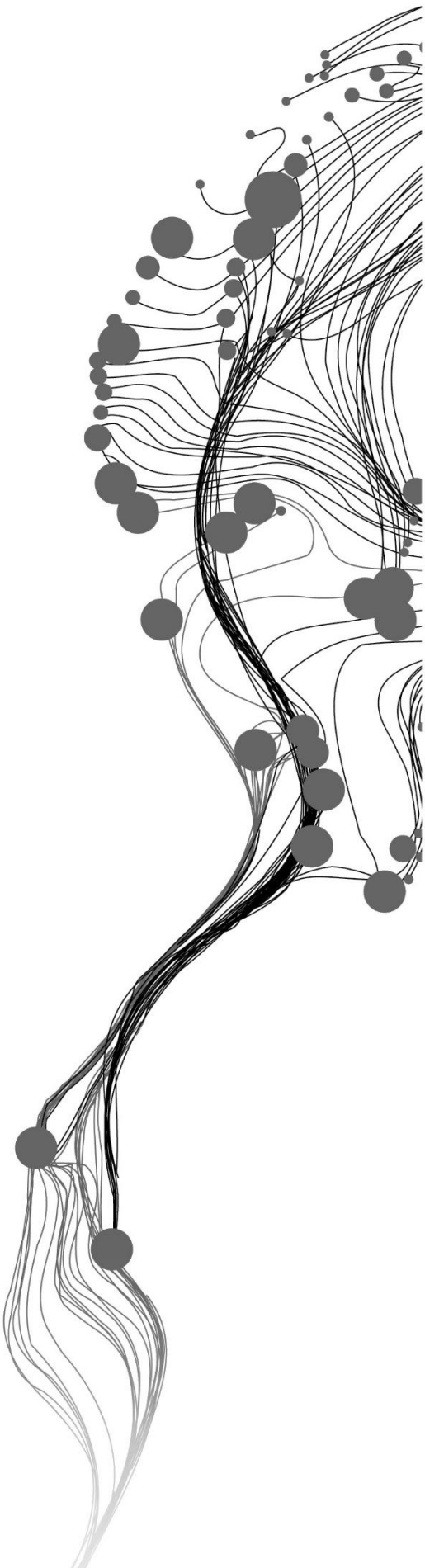
Hotel-related Attributes and Hotel Perceived Value: A Case Study in New York City based on Geodata Science and Machine Learning

YUN XU
June, 2020

SUPERVISORS:
Dr. M. Wang
Dr. F.O. Ostermann



Hotel-related Attributes and Hotel Perceived Value: A Case Study in New York City based on Geodata Science and Machine Learning

YUN XU
Enschede, The Netherlands, June, 2020

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr. M. Wang
Dr. F.O. Ostermann

ABSTRACT

Tourism is one of the industries that drive the world's economic growth, whereas the hotel industry has always occupied a prominent position in tourism-related industries. Visitors usually post reviews to express their impressions on hotels regarding factors both from inside and outside of the hotel. Therefore, not only hotel attributes but also built-environment attributes and reviewer-related attributes should be taken into consideration in hotel research. Reviewers score the hotel in five dimensions in the TripAdvisor platform: overall rating, location, cleanliness, service, and value. Except for value, all the other indicators can be judged from specific aspects. While value, here to be more specific, hotel perceived value, is a more compound concept, whose score could be influenced by multiple factors. Hotel perceived value is a kind of intangible asset of the hotel. It is less studied and often overlooked but plays an important role in the hotel industry. In this study, the relationship between the hotel-related attributes and hotel perceived value is inspected by taking New York City as a case study area. We adopted TripAdvisor platform and NYC Open Data as the data sources, and applied geodata science in data processing, i.e., collecting hotel information data via web crawling, and transforming reviewer address data into coordinates via geocoding. Machine learning was involved in predicting hotel perceived value. Nine machine learning methods are compared: Ridge classifier, Logistic regression, Decision Tree classifier, Bagged Decision Tree, Random Forest classifier, Gradient Boosting Machine, XGB classifier, Support Vector classification, and K-Nearest Neighbors. Among them, the XGB classifier performed best. Indicators' accuracy, F1 score, recall score, and precision score are as high as 0.8. The XGB classifier feature importance function picks hotel ranking and negative review amount as the most prominent features regarding hotel perceived value. The built environment occupies the largest proportion of hotel-related attributes, more than half. However, these issues are suggested as not very significant in this study; only the accessibility/convenience to restaurants, attractions, and airports slightly show the position. Attributes related to the hotel itself display their significance in the importance ranking. One of the reviewer-related attributes, the number of cities that reviewers come from, is relatively important rather than the rest two. The reliability of the hotel's perceived value on the TripAdvisor site is also worthy of consideration. Suggestions for hotel managers are provided that they are supposed to improve the hotel ranking and remove the influence of the negative comments by responding more to these negative feedbacks in order to enhance hotel perceived value. Nevertheless, cautions are needed when generalizing the results from this study, given the potential presence of multicollinearity, which, however, does not affect the overall performance of the prediction. In future work, issues related to reviewer classification, review text analysis, reviewer address, and hotel classification could be considered to build up a more complete and convincing hotel perceived value research framework.

Keywords: hotel attribute, perceived value, geodata science, machine learning, TripAdvisor, NYC Open data

ACKNOWLEDGEMENTS

I would like to express my gratitude to my first supervisor, Dr. Mingshu Wang, and my second supervisor, Dr. Frank Ostermann, for their constant guidance and instructive suggestions during the preparation of my thesis. Their critical thinking is what I have been pursuing, and their professional quality is what I appreciate. Special appreciation goes to Yiming, the software developer of Hawk, who gave me timely instructions for applying Hawk so that my thesis can keep going on smoothly.

I also want to dedicate my thanks to ITC and UT, who offered me the ITC Excellence Scholarship and provided me the opportunity to study in the Netherlands.

What's more, I would like to thank my friends whom I meet in ITC, Fanshu, Yi, Xujiayi, Xinyi, Xin, Xiaojian, Shan, Markie, Coco, Rui, and all the other lovely schoolmates, from whom I get precious friendship, encouragement, and happiness.

Finally, I would owe my deepest love to my angel mother, who is always standing behind me and silently supporting me at any cost. It is a pity that she cannot come to my graduation ceremony and witness the finish of a journey in my life due to the COVID-19. I want to embrace my mother from afar and tell her I will always love her.

Yun Xu
Enschede, 1 June 2020

TABLE OF CONTENTS

1. Introduction 1
   1.1. Background 1
   1.2. Research objectives and questions 2
   1.3. Outline of the thesis 3
2. Literature review 5
   2.1. Hotel perceived value 5
   2.2. Geodata science 6
   2.3. Machine learning 6
3. Study area and data description 9
   3.1. Study area 9
   3.2. Data description 10
4. Methodologies 15
   4.1. Geodata science 16
   4.2. Machine learning 21
5. Results 27
   5.1. Descriptive statistics 27
   5.2. Cross-validation and model selection 28
   5.3. Feature importance 29
6. Discussion 33
   6.1. Geodata processing 33
   6.2. Feature importance 33
7. Conclusion and outlook 37

LIST OF FIGURES

Figure 1 New York City Borough and hotel location distribution map....9
Figure 2 Attribute information from TripAdvisor web page....11
Figure 3 Example of data lens pages....12
Figure 4 An example of datasets....12
Figure 5 An example of external datasets....12
Figure 6 An example of Files & Documents....13
Figure 7 An example of Filtered Views....13
Figure 8 An example of maps....13
Figure 9 Flowchart of the study....15
Figure 10 Examples of irregular address text records in collected TripAdvisor data....16
Figure 11 Flow chart of "Text to Coordinates"....17
Figure 12 Flow map of Waldorf Hotel in the year 2014....17
Figure 13 The interface of Hawk when applying data cleaning....18
Figure 14 Basic process steps in Hawk....18
Figure 15 Machine learning procedure overview....22
Figure 16 Confusion matrix....24
Figure 17 XGB classifier feature importance....30
Figure 18 Logistic regression feature importance....31
Figure 19 Feature importance of Random Forest classifier (a) and Gradient Boosting classifier (b)....32

LIST OF TABLES

Table 1 TripAdvisor review data information 10
Table 2 Selected NYC Open Data introduction 14
Table 3 Target variable label distribution 20
Table 4 Variable information of hotel-related attributes 20
Table 5 List of machine learning methods and Python packages 22
Table 6 Matrics score functions in Python 25
Table 7 Descriptive statistics for the variables of hotel 27
Table 8 Model hyperparameter tuning 28
Table 9 Model performance based on 10-fold cross-validation 29

1. INTRODUCTION

1.1. Background

Tourism is a global force that drives economic growth and promotes employment opportunities. According to the World Travel Organization (2019), the volume of international tourists reached the 1.4 billion mark in 2018, and export revenue from tourism rose to USD 1.7 trillion. The rapid development of tourism has driven the accommodation, food, transportation, entertainment, and other related ancillary industries, making a great contribution to the world's economic development.

Hotels provide accommodation facilities and catering services for tourists, which is a key material basis for the advancement of the tourism industry and an important basis for tourists to complete tourism activities. It is a matter of great importance to promote the development of the two in a coordinated way. As reported by TUI (2019), the market size of the global hotel industry keeps growing from 2014 to 2018. The global occupancy rate (the total number of rooms occupied or rented at a given time) of hotels increased during the five years in most regions, with some continents rising as high as more than 70 percent.

As a place that provides space for guests or tourists to have a rest or live in, hotels vary in size, cost, service, style, location, equipment, infrastructure, surrounding environment, and so forth. Visitors choose a hotel based on their judgment. With the rapid development of information technology, people are now more used to finding information about a particular hotel and booking a room via web pages or applications. Among them, the TripAdvisor platform occupies a significant position. It is known as the most prominent travel platform in the world, containing more than 830 million pieces of opinions and reviews regarding 8.7 million travel-related issues such as accommodation, experiences, and so on (TripAdvisor, n.d.). From the hotel information pages, customers can view various indicators about it, which may influence their choices. Out of personal experience, they may choose a hotel considering and prioritizing price, reviews, pictures, rating values, etc. After staying in a hotel, a visitor can also share his experience in the hotel via TripAdvisor platform, marking the overall property together with a rating of location, cleanliness, service, and value.

In this research, we mainly concentrate on the "value" of the hotel. Hotel value has mainly two types: market value and perceived value. Market value more refers to the brand and attribute value (O'Neill & Xiao, 2006). Perceived value is concluded as the overall evaluation of the product's utility from consumers based on the perceptions received and given, according to Zeithaml (1988), which is an important factor in studying customer satisfaction. It will not only have an impact on consumers' purchase action but also influence their motivation to recommend and return (Al-Sabbahy, Ekinci, & Riley, 2004). Even though for marketing perceived value counts, it was not taken seriously in the related research (Dodds, Monroe, & Grewal, 1991) because of the difficulties with its conceptualization (Al-Sabbahy, Ekinci, & Riley, 2004). In this study, as reviewers mark the value of the TripAdvisor platform, it is closer to the definition of perceived value. As the hotel's perceived value accounts for a number of reasons, it is necessary for the hotel industry to quantify and understand it for operation and promotion.

Reviewers can score the hotel in five dimensions in the TripAdvisor platform, namely overall rating, location, cleanliness, service, and value. Despite value, all other indicators judge hotels from specific aspects. While value, here to be more specific, hotel perceived value, is a more compound and subjective concept, whose score could be influenced by multiple factors. Therefore, those multiple factors from different aspects should be taken into consideration as well.

Many factors from both inside and outside of hotels may determine the visitors' choice. Based on the research of Dolnicar & Otter (2003), several hotel attributes come forward with an "important" ranking. The ranking contents are listed as follows: "convenient location, service quality, reputation, friendliness of staff, price, room cleanliness, value for money, hotel cleanliness, security, room standard, swimming pool, the comfort of the bed, parking facilities and room size". However, as Tsai et al. (2009) clarified, limiting factors are examined in studies and literature that talk over hotel competitiveness. It can also be noticed that most of the factors are only related to the hotel itself. Therefore, it is necessary to examine more factors as well as factors outside the hotels in order to see how hotel perceived value is influenced. According to Go, Pine, & Yu (1994), the development of the community where a hotel located stimulates the performance of the hotel. For example, entertainment facilities may attract visitors and generate needs for hotel rooms (Tsai et al., 2009). Therefore, the built environment of hotels is also worthy of being considered in hotel research. The built environment refers to the combination of artificial surroundings in the modern world, involving the fields of economics, management, geography, design, technology, and so forth (Roof & Oleru, 2008). In the hotel industry, the built environment can be represented by the hotel near-by attributes (e.g., nearest transportation and nearest restaurant) and the located environment, i.e., neighborhood attributes (e.g., neighborhood population and greening rate). We simply name the factors inside hotels as "hotel attribute" and the factors outside hotels as "built-environment attribute". In the meanwhile, we also consider factors that are produced from reviewers, which are named as "reviewer-related attribute", such as the distance between the hotel and reviewers' city, because these attributes are not from the hotel location but are key issues in hotel research as well.

For multivariable studies in the hotel industry, a questionnaire is widely used in the early years (Dolnicar & Otter, 2003). As with the changing of data sources, scientists are not limited to direct investigation. Machine learning becomes the newly developing research method in the hotel industry research area. Because machine learning helps to analyze automatically based on a huge number of disparate data, it also helps to build models for better predictions. In addition to accurate analysis, it saves time, money, and human resources in the meantime. Therefore, to better generate the multiple attributes of the hotel in this study, machine learning is applied for the understanding of the hotel's perceived value.

1.2. Research objectives and questions

1.2.1. General objective

The general objective of this study is to understand how hotel-related attributes, including hotel attributes, reviewer-related attributes, and built-environment attributes, can predict hotel perceived value from visitors with geodata science and machine learning methods.

1.2.2. Sub-objectives and questions

To achieve the general objective, we illustrate sub-objectives with three parts, which are multi-sourced geodata acquisition and processing, hotel perceived value model building, and hotel-related attributes

importance analysis, combing with five research questions. The research sub-objectives in this research and their corresponding questions are as follows:

(1) To acquire, process, and integrate hotel-related attributes from multiple sources.
Data will be collected from the TripAdvisor platform and NYC Open Data website, including hotel attributes, reviewer-related attributes, and built-environment attributes. The data are then processed and integrated into independent variables for machine learning models.

The related research question is:
Q1.1 What should be the variables representing hotel-related attributes?

(2) To predict hotel perceived value by machine learning based on hotel-related attributes.
Machine learning models, including linear, ensembles (bagging and boosting), and other algorithms are applied for hotel perceived value classification. The best machine learning model will be selected based on the accuracy, F1 score, precision score, and recall score.

The related research questions are:
Q2.1 What are the parameter settings regarding the selected models?
Q2.2 What are the accuracy, F1 score, precision score, and recall score of these models?
Q2.3 Which model would be the one that has the highest accuracy-related indicators?

(3) To find out the most important hotel-related attributes that contribute to hotel perceived value. Feature importance ranking is supposed to be listed and studied.

The related research question is:
Q3.1 How are the hotel-related attributes distribute in feature importance ranking?

1.3. Outline of the thesis

In the next chapter 2, related work about hotel perceived value, geodata science, and machine learning will be introduced. In chapter 1, basic information about the study area, and the explanation of the data will be presented. Next, the methodology on how to process multi-sourced data and how to perform machine learning will be described in chapter 1. In chapter 1, the result of this study will be proposed. In chapter 1, the result will be discussed. Finally, the study will be summarised, and the limitation and improvement of the study will be concluded in chapter 1.

2. LITERATURE REVIEW

Hotel perceived value is a relative compound concept for visitors, which can be contributed by all kinds of attributes from hotels, visitors, and the built environment. To generate these multiple attributes in the hotel perceived value model, machine learning is nowadays a widely used method. To process data into machine learning models, geodata science is the must way to handle data from multiple sources. Therefore, the literature review is expended from three aspects: hotel perceived value, geodata science, and machine learning.

2.1. Hotel perceived value

In the hotel industry, the most important question for a hotel manager to answer is, "why should I choose the hotel?" As a kind of intangible asset for the hotel industry, hotel perceived value proposition impresses consumers that the hotel is unique and matched, leading consumer's actions such as booking, revisiting, and so on. Hotel perceived value is defined as "what consumers will get (quality) for how much they pay"(Zeithaml, 1988). It shows that service quality and price are the two important issues for perceived value. Oh (1999) assessed the role of the perceived value in an integrated framework of customer satisfaction and service quality in order to understand the consumers' decision-making process. Sweeney & Soutar (2001) suggested quality, emotional, social, and functional value as four sub-dimensions of perceived value.

For more hotel valuation research, Callan & Bowman (2000) proposed that cleanliness, value for money, bedroom comfort, and safety & security as top issues in the hotel attribute importance list. Mattila & O'Neill (2003) identified that besides price, which leads to overall consumer satisfaction, cleanliness, maintenance of the guest room, and staff attentiveness are the three factors regarding consumer satisfaction. Chan & Wong (2005) figured out that price, location, and service are the most influential aspects of hotel selection. Moreover, traveler type induced different emphasis on other hotel attributes. J. Zhang, Ye, & Law (2011) suggest that empirical findings proposed location and room quality as the key indicators for the price of the room, and room design and amenities help to introduce value for hotels. Raza et al. (2012) investigated 125 luxury hotel customers of Pakistan and uncovered that perceived value and service quality are positively connected with satisfaction and revisit intentions.

There are pieces of literature that emphasize the relationship between the built environment and hotel perceived value as well. Yang, Mao, & Tang (2018) classified factors related to location into accessibility to attractions, the convenience of transport, and the environment around to determine guest satisfaction. They suggested that accessibility, green spaces, water body, and local businesses are of significant importance. H. Li, Ye, & Law (2013) illustrated a similar finding that transportation convenience, tourist destination accessibility, and value are important factors for booking hotels.

With the Internet's rise, online review research regarding hotel perceived value becomes popular. Xie, Zhang, & Zhang (2014) inspected reviews online and corresponding responses from managers from 843 hotels and figured out that overall rating, value rating, cleanliness and location, size, and variation of reviews and management response amount contribute to hotel performance. Numerous studies have applied TripAdvisor as a data source from the perspectives of negative comments (Camilla, 2011), hotel

star-classifications (Rhee & Yang, 2015), reviewers' travel experience (Gao, Li, Liu, & Fang, 2018), multiple psychological distances (Huang, Burtch, Hong, & Polman, 2016), and so forth.

2.2. Geodata science

Geodata science is a wide-ranged concept. According to Zuo & Xiong (2020), geodata science is the intersection of geoscience and data science. Geodata science framework contains geoscience data sets collection, earth information mining, geographic knowledge discovery, and spatial decision making.

Geographic information is widely applied in research. Thereinto social media data with geotags can be a tool for analyzing the tourism and hotel industry. Martí, Serrano-Estrada, and Nolasco-Cirugeda (2019) discussed the challenges, limitations, opportunities, and biases related to the adoption of location-based social media data, such as the lack of consistency of geocoded data, space limitation with untransferable locations and so forth. For location-based social media data, it is usually unstructured and massive. In the meanwhile, multi-sourced data are challenging to be unified. Geodata science gives the possibility to collect and process information that is hard to be quantified in a cheaper and large-scale way.

In the hotel and tourism industry, Park, Yang, & Wang (2019) suggested that an inverted U-shaped relationship is shown between service satisfaction and travel distance. During the research, they developed an automated crawler program to collect hotel-related information, where geographic data are involved: hotel location evaluation, user home location, etc. Huang et al. (2016) studied more than 160,000 online reviews of the restaurants, proving the effect of distance boosting. They adopted geocoding by using Google Maps API to obtain geographic coordinates for restaurants based on their address information. Nyaupane, Graefe, & Burns (2003) used GIS software to calculate the distance between respondent home and site based on home zip code and site coordinates to determine the socio-demographic and behavioral attributes. Kisilevich, Keim, & Rokach (2013) worked out a decision-support system based on GIS that helps to use basic hotel and location features to estimate objective hotel rates and predict temporary room prices. Yang, Tang, Luo, & Law (2015a) designed a web GIS application for hotel location selection. The application involved a series of machine learning algorithms for predicting indexes related to locations, which presented the potential power of the combination of geodata science and machine learning.

In this study, we collect geodata from multiple sources: TripAdvisor platform and NYC Open Data site, which contain geographic factors, including location information, accessibility information, etc. Geodata science methods such as geodata collection via web crawling, geocoding, and distance calculation are applied to generate variables for building hotel perceived value models.

2.3. Machine learning

According to Mitchell (1997), machine learning is the research on algorithms that helps a system to teach itself and improve by itself based on experience. It is good at finding rules from complicated data. Usually, the typical steps in a machine learning framework are as follows: data collection, data preprocessing, model selection, model training, model evaluation, parameter tuning, and prediction. With the rise of big data, machine learning is nowadays a method that is widely applied in all areas, and geographical information area is no exception. The specific method will change its applicability according to different cases. To figure out what kind of machine learning algorithms are suitable for this study, we mainly focus on the utilization of machine learning in the tourism and hotel industry.

Machine learning is applied in remote sensing (Belgiu & Drăgu, 2016; Mountrakis, Im, & Ogole, 2011), mapping analysis (Kobler & Adamic, 2000; Rahmati, Pourghasemi, & Melesse, 2016), hazard evaluation (Feizizadeh, Roodposhti, Blaschke, & Aryal, 2017; Mojaddadi, Pradhan, Nampak, Ahmad, & Ghazali, 2017; Tehrany, Pradhan, Mansor, & Ahmad, 2015) and so forth. Hagenauer, Omrani, & Helbich (2019) compared 38 machine learning models to analyze land consumption rates (LCR), where the eXtreme gradient boosting decision tree performed best and support vector machine with polynomial kernel performed worst. In the tourism industry, Zhou, Wang, & Li (2019) modeled transport means based on environment and temporal factors with machine learning methods. They compared 11 machine learning algorithms and selected Random Forest as the best-fitted model with all variables.

Though limited discussion on machine learning algorithm applications can be found in the literature of the hotel industry (Y. Zhang, 2019), it is necessary to fill the vacancy in this area. According to Zhang (2019), hotel online review analysis is the proportion in the hotel industry that applies machine learning methods mostly. Garcı´a-Pablos, Cuadros, & Linaza (2016) applied Conditional Random Fields (CRF) and Support Vector Machine (SVM) to train the annotated hotel reviews. Schmunk, Höpken, Fuchs, & Lexhagen (2014) adopted Naïve Bayes, Support Vector Machines (SVM), and k-nearest neighbor (k-NN) to learn how to express the attributes, subjectivity or emotion derived from the words that appear in the sentence. In spite of review analysis, Yang, Tang, Luo, & Law (2015) applied various machine learning algorithms such as Linear regression, Support vector regression, and Boosted Regression to predict business indicators accompanied by locations.

Based on the research, we consider applying Ridge classifier, Logistic regression, Decision Tree classifier, Bagged Decision Tree, Random Forest classifier, Gradient Boosting Machine, XGB classifier, Support Vector classification, and K-Nearest Neighbors as the machine learning models. The bunch of methods includes linear, ensemble, and other models, where ensemble models can be divided into bagging and boosting approaches. It ensures the model diversity in order to pick up the models with the best performance so that the hotel's perceived value can be better studied.

3. STUDY AREA AND DATA DESCRIPTION

In this chapter, the reasons for choosing New York City as the study area and the brief information of the city are introduced. Meanwhile, the data source of the study and specific data content are presented.

3.1. Study area

New York City is selected as the study area in this research. It is one of the cities with open data access. Besides, New York City is one of the leading tourist destinations, attracting visitors from both domestic and oversea areas all year round. Tourist quantity provides the basis of big data. According to NYC & Company (2019b, 2019a, 2020), the city has experienced ten-year continuous tourism growth, the visitor volume of the New York City keeps increasing year by year, hotel occupancy and average daily rate are staying at a high level during the past five years as well. The hotel marketing in New York City presents a prosperous scene, following by the market competition becomes more intensive. The New York City Department of City Planning (2017) quotes Smith Travel Research (STR), claiming that there are over 630 hotel attributes in the five boroughs where over 80 percent of the hotel rooms are in Manhattan. In the meanwhile, much of the growth of hotel rooms since 2010 happened in all boroughs. Therefore, built-environment issues are worthy of consideration in this area.

All in all, New York City is a desirable place for research. Figure 1 shows the borough distribution and the location of the hotels we collected in New York City. It is noticed that during the data collection, the Manhattan borough, named "New York" at the county level, is representing NYC in TripAdvisor when users try to search "New York". Therefore, all of the hotels we collected are located in Manhattan borough.

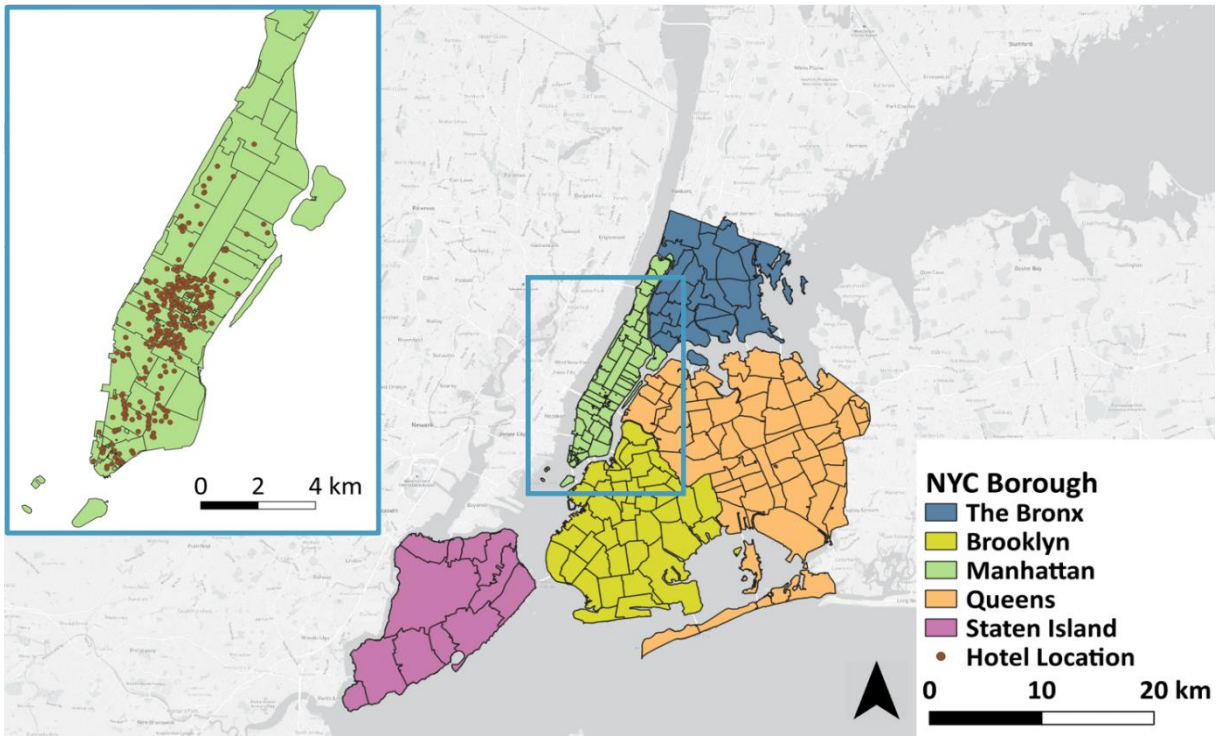


Figure 1 New York City Borough and hotel location distribution map

3.2. Data description

There are two main sources of data in this research, namely TripAdvisor[1] and NYC Open Data[2]. In this section, both datasets and the choice of the possible attributes that have a contribution to hotel perceived value will be introduced.

TripAdvisor is one of the most extensive trip platforms of the world, owning a business on making plans and reviews covering hotels, activities, restaurants, flights, vacation rentals, and cruises. Travelers from all around the world use the TripAdvisor website or application to gain experience from other visitors or share their own experience during the vacation. TripAdvisor data contain most of the useful information in the study. On the one hand, TripAdvisor data of hotel reviews in English within New York City during the year 2004 to 2015 were collected by web crawling. There are 466 hotel data items in total. In each item, 13 attributes are included. The details of the applied attributes in the data processing are specified in Table 1. Based on the hotel address and reviewer address, the distance between the two can be calculated. City information can also be extracted from the reviewer address; therefore, the number of cities where reviewers come from can be computed as well. The specific processing is described in 4.1.1.1. On the other hand, TripAdvisor data from hotel information pages were crawled in order to aggregate as many hotel attributes and built environment attributes as possible. The examples of attribute contents are pointed out in Figure 2, illustrating the attribute information from the TripAdvisor web page where the annotations are the attribute names set in the data frame. The specific explanation of these attributes is listed in Table 4 in 4.1.4, with the data category named "hotel attribute" and "built-environment attribute".

Table 1 TripAdvisor review data information

Column name	Explanation
hotel_name	The name of the hotel for each item
hotel_ranking	The hotel ranking among 466 hotels
hotel_address	The address of the hotel
review_date	The date that the reviewer wrote the review
review_text	The review text that the reviewer wrote for the hotel
reviewer_address	The address that the reviewer wrote in his/her TripAdvisor account profile

[1] The official website of TripAdvisor: https://www.tripadvisor.com/
[2] The official website of of NYC Open Data: https://opendata.cityofnewyork.us/

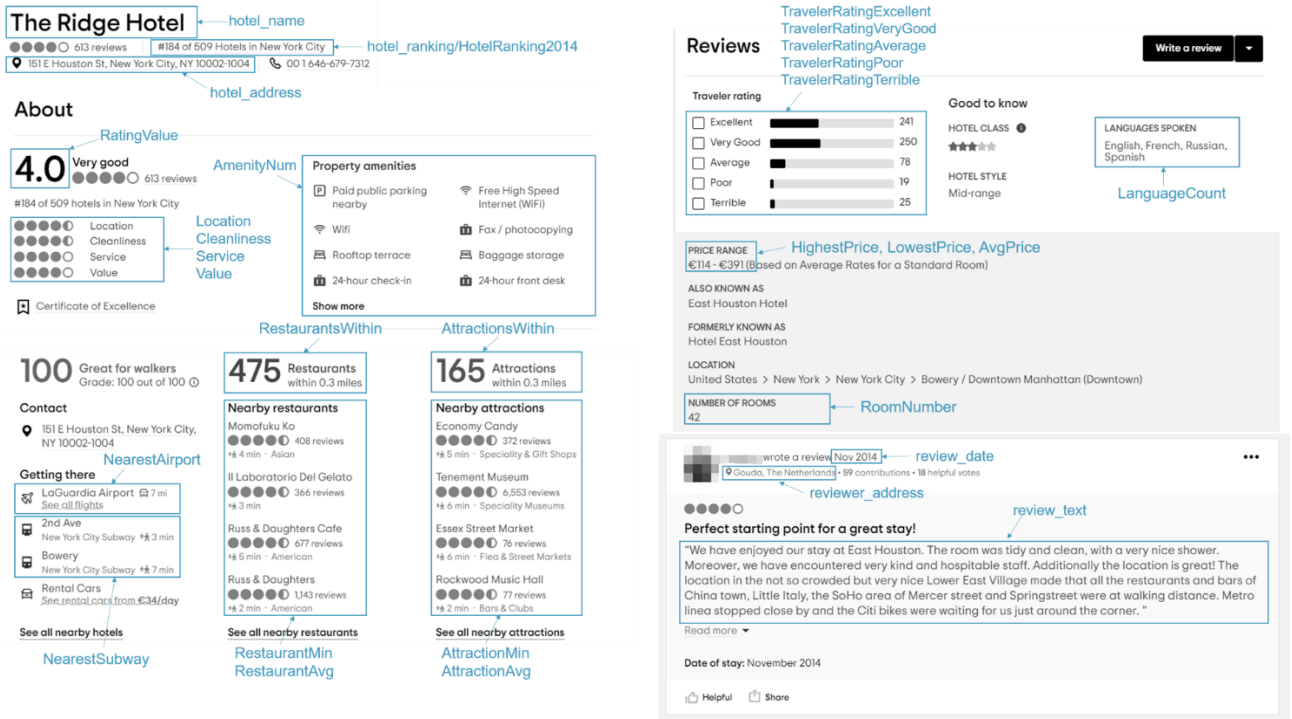


Figure 2 Attribute information from TripAdvisor web page

NYC Open Data is jointly developed by the Mayor's Office of Data Analytics (MODA) and the Department of Information Technology and Telecommunications (DoITT). NYC Open Data covers nearly 2000 categories of data from five categories, including Business, City Government, Education, Environment, and Health. Different data view types are involved in NYC Open Data, such as Data lens pages (shown in Figure 3 Example of data lens pagesFigure 3), datasets (shown in Figure 4), External datasets (shown in Figure 5), File & Documents (shown in Figure 6), Filtered Views (shown in Figure 7) and maps (shown in Figure 8). Built-environment features at neighborhood-level are derived from this site. As is mentioned in 1.1 and 2.1, green space, the community's economic, social, and cultural issues are proved to have an impact on the hotel industry. Based on that, we also take local infrastructure and issues related to the environment into account, together with the consideration of the availability of zip code, coordinates, and borough code in the data. We looked through all the items of data in NYC Open Data and filtered them based on the reasonability and availability. From business category, we chose data of tobacco and electronic cigarette retail dealer, sidewalk café, library, and filming locations; from city government category, we chose data of population, wifi hotspot, and parks zones area; from environment category, we chose data of community gardens, street tree data, and public recycling bins. There are finally ten items of data that are selected to process built-environment attributes at the neighborhood-level. The detailed information of those data is listed in Table 2.

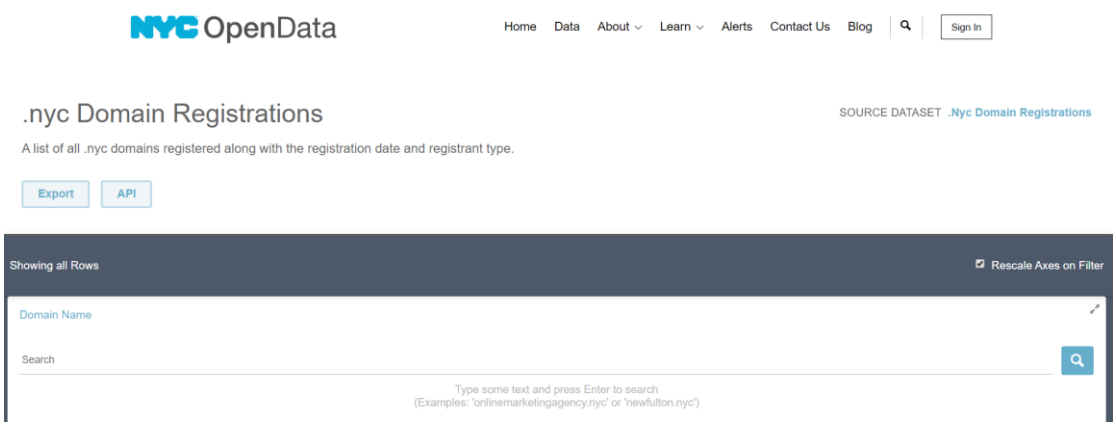


Figure 3 Example of data lens pages

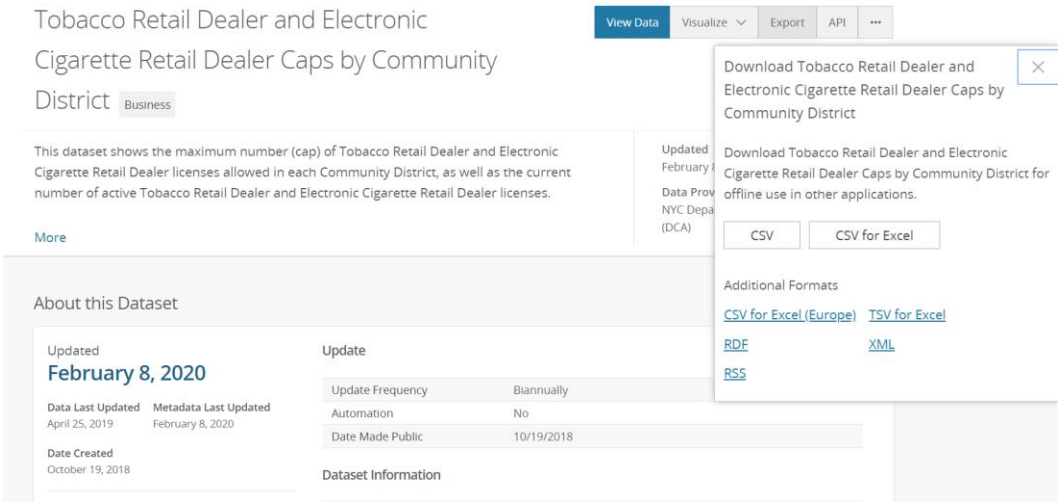


Figure 4 An example of datasets

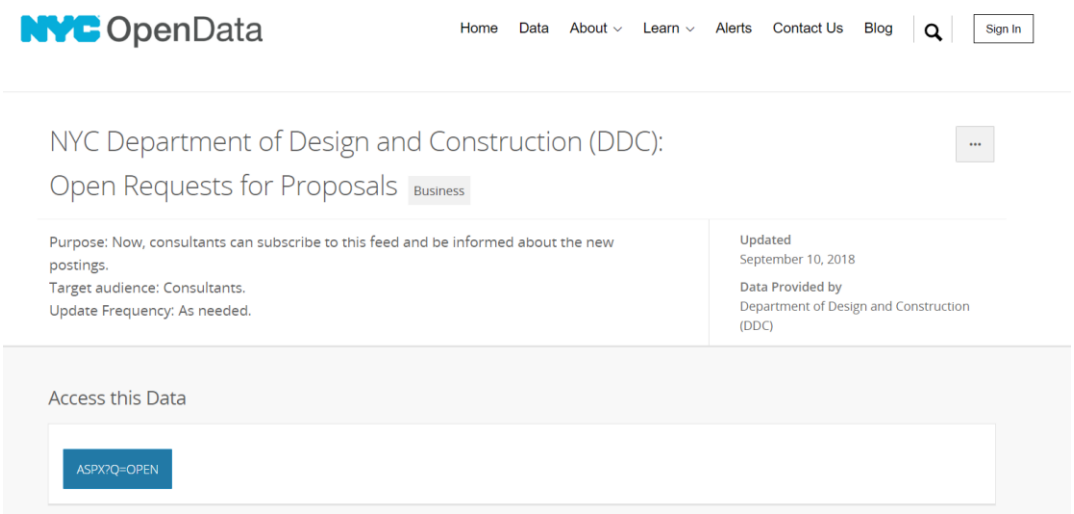


Figure 5 An example of external datasets

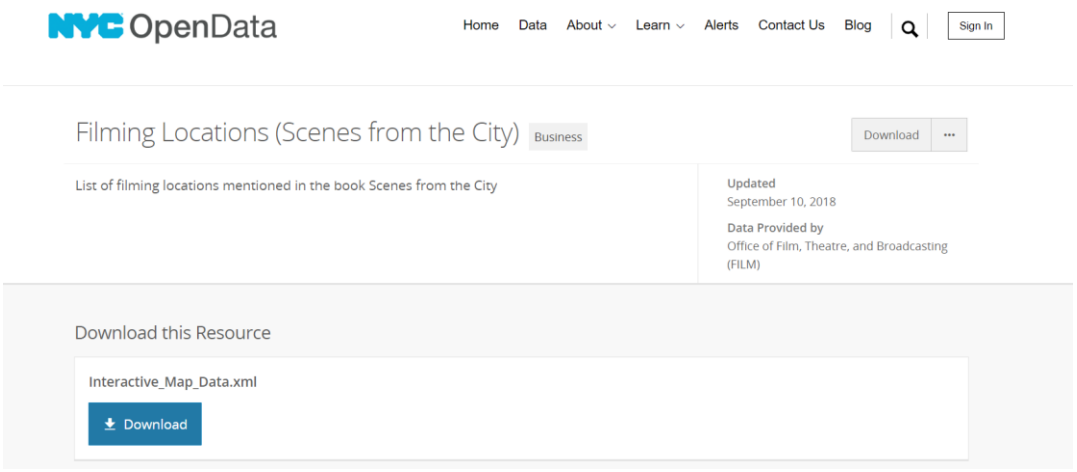


Figure 6 An example of Files & Documents

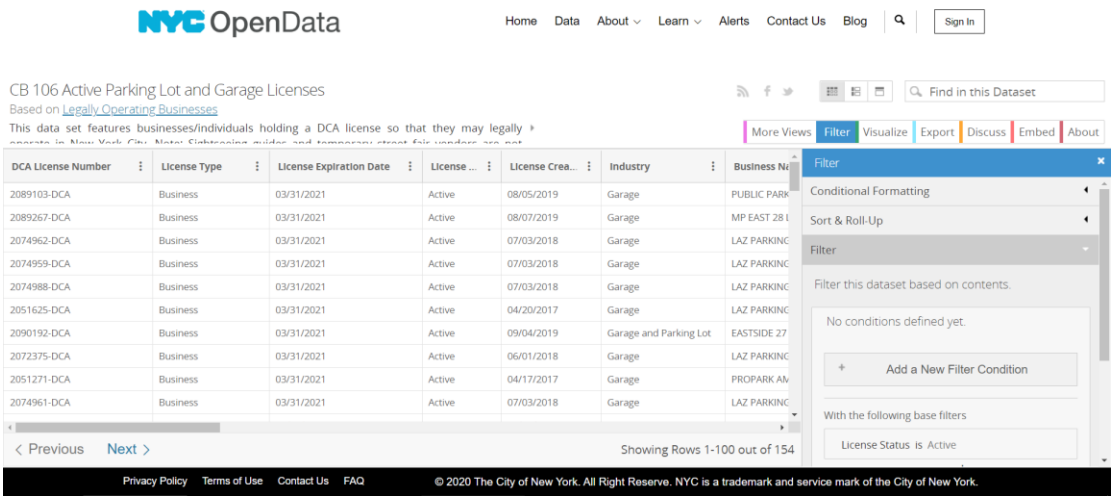


Figure 7 An example of Filtered Views

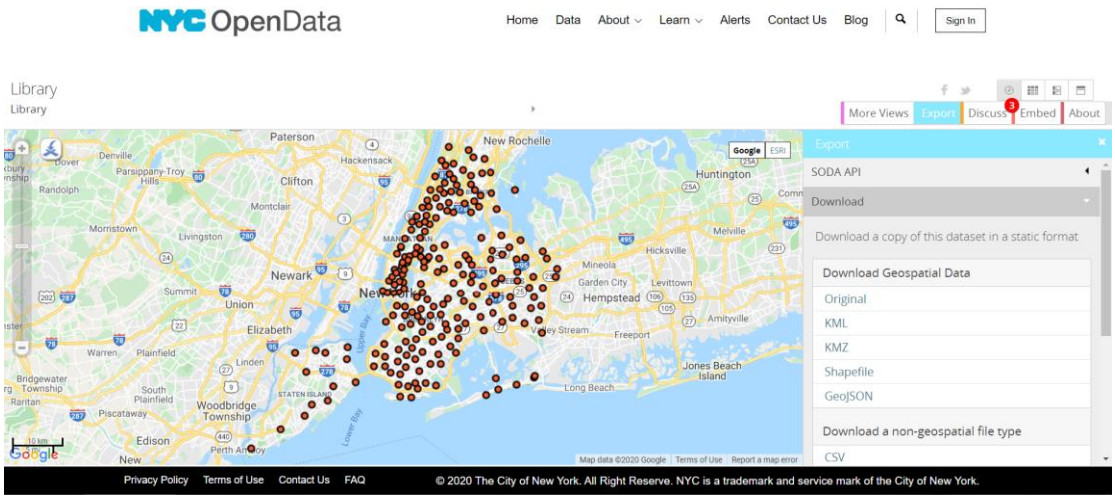


Figure 8 An example of maps

Table 2 Selected NYC Open Data introduction

Category	Data name	URL	Description
Business	Tobacco Retail Dealer and Electronic Cigarette Retail Dealer Caps by Community District	https://data.cityofnewyork.us/Business/Tobacco-Retail-Dealer-and-Electronic-Cigarette-Ret/ymyu-3dbp	The number of tobacco and electronic cigarette retail dealer licenses in each community district
	Sidewalk Café Licenses and Applications	https://data.cityofnewyork.us/Business/Sidewalk-Caf-Licenses-and-Applications/qcdj-rwhu	The number of sidewalk café license applications and issued licenses
	Library	https://data.cityofnewyork.us/Business/Library/p4pf-fyc4	Library locations in New York City
	Filming Locations (Scenes from the City)	https://data.cityofnewyork.us/Business/Filming-Locations-Scenes-from-the-City-/qb3k-n8mm	List of filming locations mentioned in the book Scenes from the city
City government	New York City Population By Community Districts	https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Community-Districts/xi7c-iiu2	NYC population by community districts, from census bureaus' decennial data for the years 1970, 1980, 1990, 2000 and 2010
	NYC Wi-Fi Hotspot Locations Map	https://data.cityofnewyork.us/City-Government/NYC-Wi-Fi-Hotspot-Locations-Map/7agf-bcsq	NYC Wi-Fi hotspot location distribution
	Parks Zones	https://data.cityofnewyork.us/City-Government/Parks-Zones/4j29-i5ry	Large NYC parks are subdivided into smaller sections as zones.
Environment	NYC Greenthumb Community Gardens	https://data.cityofnewyork.us/Environment/NYC-Greenthumb-Community-Gardens/ajxm-kzmj	List of NYC green thumb community gardens
	2015 Street Tree Census - Tree Data	https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh	2015 street tree census data, collected by volunteers and staff from NYC Parks & Recreation and partner organizations
	Public Recycling Bins	https://data.cityofnewyork.us/Environment/Public-Recycling-Bins/sxx4-xhzg	Locations of public recycling bins throughout New York City

4. METHODOLOGIES

Figure 9 provides an overview of the study process. The workflow can be mainly separated into two parts: geodata science and machine learning. The acquisition of geodata from multiple sources, and the data processing and integration procedure are illustrated in 4.1, which belong to the geodata science part. The data are collected from the TripAdvisor platform and NYC Open Data site. TripAdvisor data are divided into review data and hotel information data. The former is handled by the geocoding method, and the latter is generated by web crawling. They are integrated as hotel-related attribute data and trained in different machine learning models. The model training, evaluation, and selection are explained in 4.2, which belong to the machine learning part, and the feature importance results are shown in 5.3. The analysis is discussed in chapter 1. A literature review (chapter 2) is involved throughout the whole process.

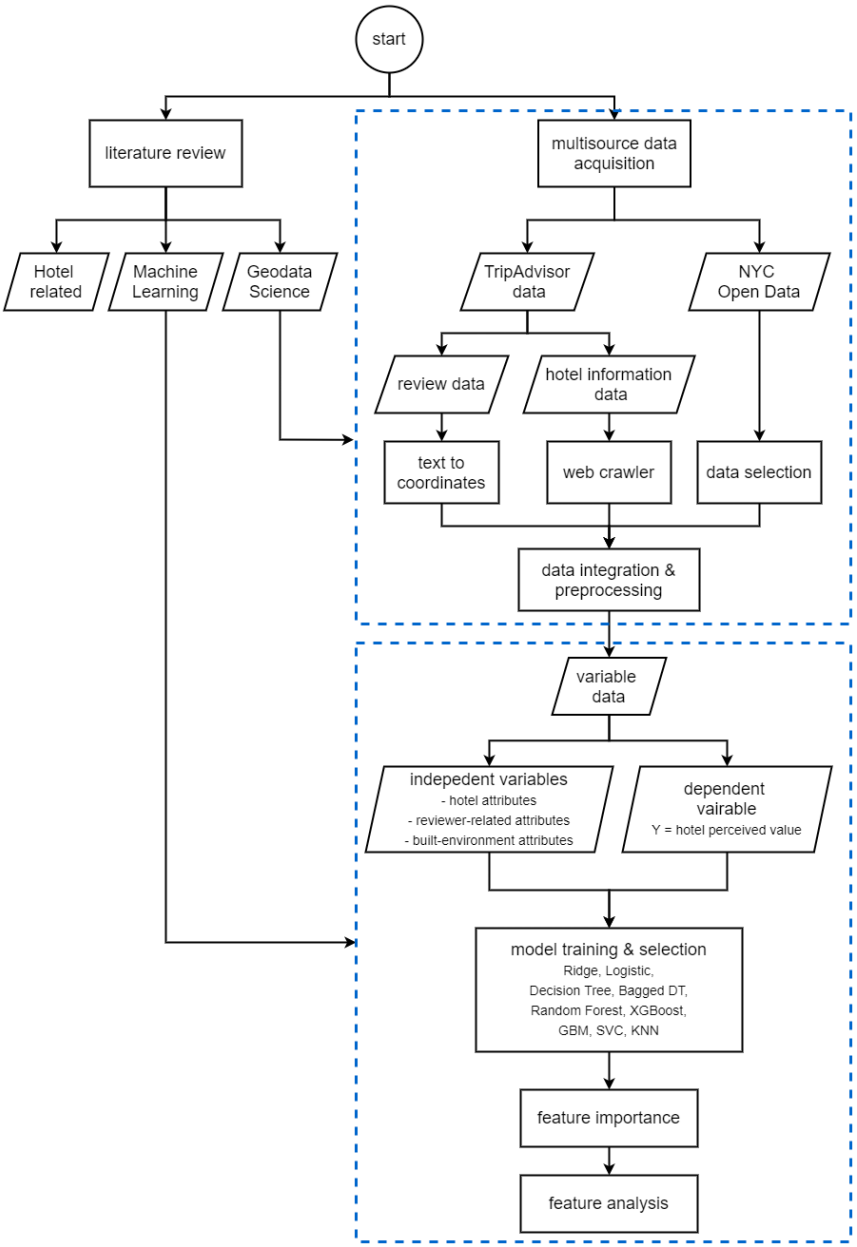


Figure 9 Flowchart of the study

4.1. Geodata science

In this part, two objectives are supposed to be achieved. One is the TripAdvisor data processing, of which review data require geocoding to transform the reviewer address and hotel address into coordinates, and hotel information data require web crawling to be generated. The other one is the NYC Open data processing, where built-environment data are filtered, which are regarded as possibly influential attributes, and basic data processing packages are applied in Python.

4.1.1. TripAdvisor data processing

The TripAdvisor data, including review data and hotel-related attribute data, are both collected via web crawling, while the review data was collected in 2015, and the hotel-related attribute data are newly collected. The specific processing of the hotel-related attribute data is illustrated in 4.1.1.2, whereas 4.1.1.1 mainly focuses on the geocoding method regarding review data.

4.1.1.1. Text to coordinates

The number of cities that visitors come from (namely "CityCount"), the average distance between hotels and visitors (namely "MeanDistance"), and the standard deviation of the distances (namely "StdDistance") are the goals that are taken for the possible components of the reviewer-related attributes. To achieve this goal, the geocoding method should be adopted, where coordinates of hotels and reviewers' addresses are expected to be obtained first. The variable named "reviewer_address" in the TripAdvisor review dataset is in the irregular address format. Because in the early years, the reviewer location recorded in the profile of TripAdvisor can be filled with any words by the reviewers themselves. Problems arose when we try to figure out what is the real city that the reviewers come from. Examples are shown in Figure 10. There are several possibilities of the wrong texts we meet: a fake address, a misspelled address, another orthography, county/state/province/country (the different level from the city), blank, and so forth.

	A	B	C	D	E	F
1	hotel_name	hotel_ranking	hotel_address	review_date	review_text	reviewer_address
2	Dream Midtown	#203 of 466	Address:210 We	September 28, 2015	We had to move ro	NYC
3	Dream Midtown	#203 of 466	Address:210 We	August 25, 2015	Near the theater dis	ct
4	Dream Midtown	#203 of 466	Address:210 We	August 24, 2015	Great location. Cen	Brooklyn
5	Dream Midtown	#203 of 466	Address:210 We	August 11, 2015	It was perfect for m	.
6	Dream Midtown	#203 of 466	Address:210 We	August 10, 2015	I love this city and I	N/A
7	Dream Midtown	#203 of 466	Address:210 We	July 22, 2015	The Dream in Midto	DEERFIELD
8	Dream Midtown	#203 of 466	Address:210 We	June 23, 2015	I loved my room. Th	EQ
9	Dream Midtown	#203 of 466	Address:210 We	May 14, 2015	Everything about my	MA
10	Dream Midtown	#203 of 466	Address:210 We	April 16, 2015	The rooms are ridicu	Yes
11	Dream Midtown	#203 of 466	Address:210 We	April 8, 2015	I've stayed at t	Smith's
12	Dream Midtown	#203 of 466	Address:210 We	March 15, 2015	This hotel was ok. I	USA
13	Dream Midtown	#203 of 466	Address:210 We	February 15, 2015	Overall very nice; O	Destin, FL

Figure 10 Examples of irregular address text records in collected TripAdvisor data

For locating the coordinates of cities and hotels, package "geopy" in Python (https://pypi.org/project/geopy/) is adapted, which is used to standardize the informal address text, recognize the address level (city, county, country and so forth) and provide the latitude and longitude. The package is applied as well to calculate the geodesic distance with two given coordinate pairs. Figure 11 explains the flows of Text to Coordinates part. In this part, 466 hotel data items change into 358, because there exist hotel data with no data of the year 2014 or with several items of data of the year 2014 but remaining no city identification.

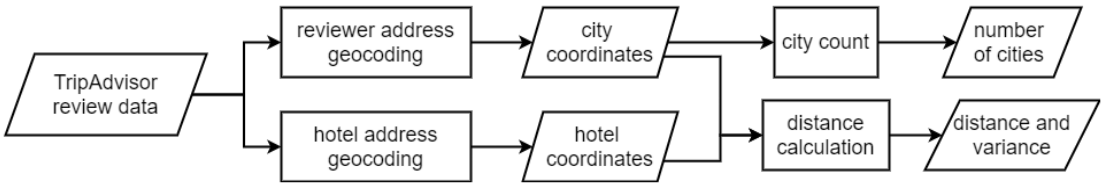


Figure 11 Flow chart of "Text to Coordinates"

Initially, Natural Language Processing (NLP) was considered for analyzing the rest data items which cannot be recognized or are with blank records. In the largest dataset, "Waldorf-Astoria Hotel", those data were filtered for testing, where the GeoText package in Python (https://pypi.org/project/geotext/) was selected to distinguish geographic terms in the text. Geographic terms were counted. While the result shows that "New York" is with the highest frequency, which is far more than the other city name counts, it can be inferred that reviewers are more likely to talk about the hotel and the located city. City name with confusing meaning like Nice, Reading, and York also occupied a particular part in the rest result. In the meanwhile, the whole bunch of city count only takes up a tiny part (more or less ten) within thousands of pieces of data. From this perspective, we decide not to use text analysis (i.e., NLP) in reviewer address analysis because it may not increase accurate information but introduce some errors in it.

Since the coordinates of the hotels and visitor addresses are obtained, how the tourism transfer from their home cities to New York City can be therefore plotted, giving a direct-viewing impression to readers. In this case, a Python script (https://github.com/paulojraposo/FlowMaps) developed by Dr. Paulo is introduced, which helps to draw flows from a start point to an endpoint. It is noted that the scipy, gdal, shapely and pyproj packages in Python are necessary, and it is better to create a new Python environment in case of the influence of other packages. Combining with operation on Python Command Prompt and GIS software, the flow map in shapefile format will be created and visualized.

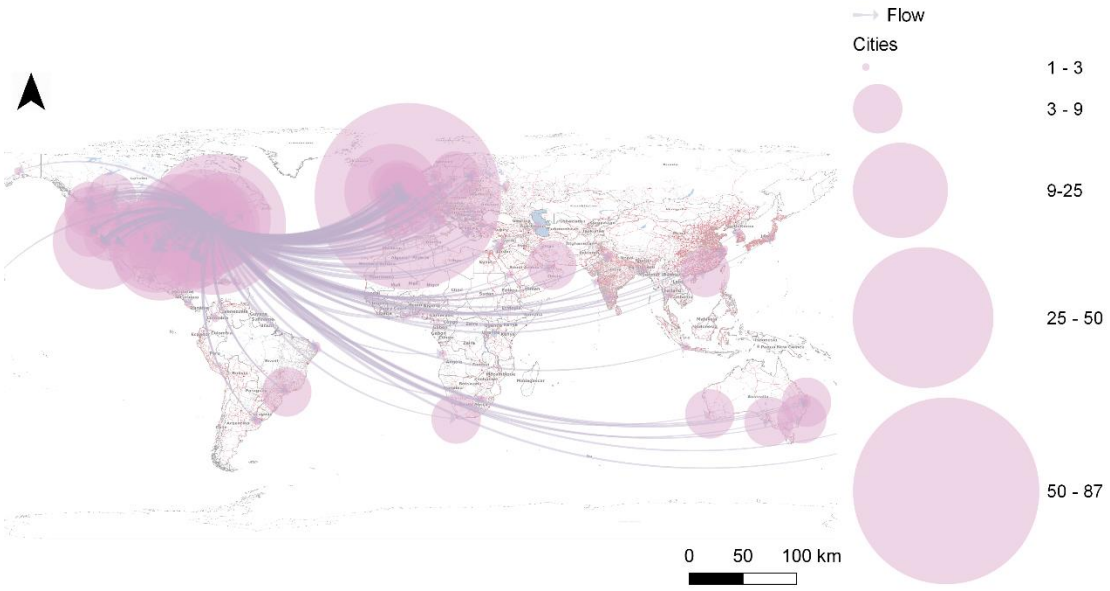


Figure 12 Flow map of Waldorf Hotel in the year 2014

Among the 338 hotels, Waldorf-Astoria contains the most extensive data size. The overall flow map of the Waldorf-Astoria Hotel in the year 2014 is shown in Figure 12 as a sample of tourism flow distribution.

The hotel attracts visitors from dozens of cities all around the world. The European and North American cities are the areas that lead to the biggest volume of tourism towards Hotel Waldorf-Astoria.

4.1.1.2. Web crawling

There are three components for hotel-related attributes: hotel attributes, built-environment attributes, and reviewer-related attributes. The hotel attributes, reviewer-related attributes, and part of built-environment attributes are sourced from the TripAdvisor platform. Except that CityCount, MeanDistance, and StdDistance are computed based on the TripAdvisor review dataset, the remaining hotel-related attributes are all collected from TripAdvisor hotel information pages by web crawling. All the built-environment attributes at the neighborhood-level are resourced from the NYC Open Data website, of which the processing procedure will be explained in 4.1.2.

Hawk (https://github.com/ferventdesert/Hawk) is an open-source software aiming at graphically crawling webpage with cleaning, processing, and saving data. There is a companion library named "etlpy" (https://github.com/ferventdesert/etlpy) in Python as well, which is a profile-based data acquisition and cleaning tool.

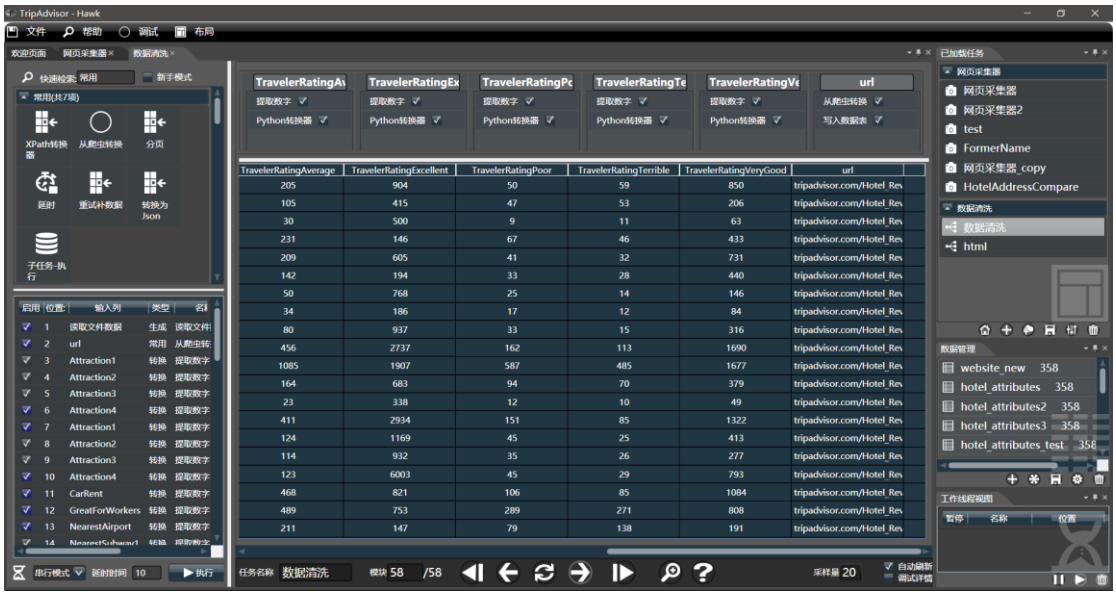


Figure 13 The interface of Hawk when applying data cleaning

Tasks in Hawk can be divided into two parts: web crawling and data cleaning. Hawk is applied in this study combining with Python for TripAdvisor hotel information collection and aggregation. Figure 13 shows the interface of Hawk when data cleaning is being applied. The basic processing steps of the data collection and aggregation are shown in Figure 14. Other processing methods are also joined, such as regular expression, character range extraction, and so forth.

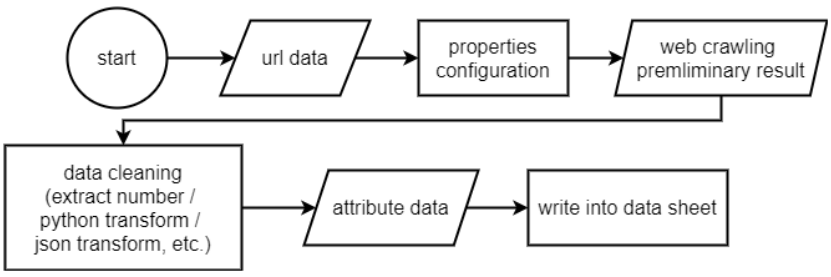


Figure 14 Basic process steps in Hawk

In this part, the size of hotel data decreased from 358 to 338. There are 20 unavailable URLs, whose representing hotels do not exist in the TripAdvisor platform anymore. In addition, it does not mean that the hotels just change their names so that the URLs change correspondingly and become unconnectable. As it is noticed that there still exist 67 URLs available whose hotels change the hotel name during the years, and the URLs will be automatically transformed into the new ones when they are inputted into the address bar of the browser. Out of the consideration of the influence of hotel name or location change, we also listed all the 338 hotels with their names and address in the year 2014 and the year 2020. Fortunately, no hotels move to another place during the years, whether their names have been changed or not.

4.1.2. NYC Open data processing

Mainly three data formats are involved in NYC Open Data collection: csv, shapefile, and xml. The general data cleaning purpose is to generate one object into one data column. For the csv data, it is initially in the table sheet format. Pandas package in Python (https://pandas.pydata.org/) is mainly applied for data processing. The shapefile data can also be extracted from their map attributes into the sheet from GIS software. As for the xml data, xml.etree.ElementTree (ET in short) module (https://docs.python.org/3.7/library/xml.etree.elementtree.html?highlight=xml#xml.etree.ElementTree. XML) is used for data analysis and extraction.

4.1.3. Data preprocessing

4.1.3.1. Nan values/missing data handling

Among the variables, several attributes contain Nan values due to data missing during data collection: RoomNumber, LowestPrice, HighestPrice, AvgPrice, FilmLocationCount, GardenCount, LibraryCount, ParkAreaACRES, RecyclingBinCount, and SidewalkCafeCount. For variables listed from FilmLocationCount to SidewalkCafeCount, Nan value does represent as zero because if there is a Nan, there is no film location, and all else follows. Therefore, zeros are filled into those Nan value positions. For the first four attributes, zero cannot represent anything. Based on Musil, Warner, Yobas, & Jones (2002), casewise deletion, mean substitution, regression imputation, and the expectation-maximization (EM) algorithm are compared to handle missing data. In this case, the four categories of data are missing completely at random (MCAR). Among these methods, as the size of the data set is not big, and the variables are not highly correlated, casewise deletion and regression imputation are not recommended. The EM algorithm is conceived preferable for missing data handling (Acock, 2005), and mean substitution is commonly used when data are MCAR. Mean substitution is then applied so that the mean values are computed to replace Nan values for the first four attributes.

4.1.3.2. Normalization

In fact, not all machine learning methods require normalization, such as Tree-based algorithms, because the percentage of correctly classified labels determines the split point, which means the feature scaling is resilient. However, some algorithms are easy to be influenced by normalization, such as K-Nearest Neighborhood for its high dependence on distance calculation. To unify the data, we applied data normalization on independent variables where Z-Score Normalization (aka Standardization) is selected. The equation is shown below (Equation 1):

$$Z = \frac{(X_i - \bar{X})}{S} \qquad \text{Equation 1}$$

Where X_i represents as feature values, $\bar{X}$ represents the mean value of X_i, and S represents the standard deviation value of X_i.

4.1.3.3. Imbalanced class handling

Generally, for the labeled classes, there is a package named "preprocessing" in sklearn library (https://scikit-learn.org/stable/modules/preprocessing.html#), in which function "LabelEncoder()" can be applied for label transformation from continuous to categorical. However, Table 3 displays the original label distribution of the target variable, where 2.0, 2.5, 3.0, and 5.0 occupy only a tiny part of the total, which belongs to imbalanced classes. Commonly the third-party library "imblearn" (https://imbalanced-learn.readthedocs.io/en/stable/api.html#) in Python provides various ways to handle issues with imbalanced classes, such as over-sampling (leads to overfitting) and under-sampling (leads to loss of other important information). In this case, considering the excessively small number of several categories, classes are rearranged into three parts: value 5.0 and 4.5 are marked as Class 2, value 4.0 is marked as Class 1, and value 3.5 to 2.0 are marked as Class 0, representing Good, Average and Bad respectively. The updated class distribution is also shown in Table 3.

Table 3 Target variable label distribution

Original Label	Frequency	Percentage%	New Label	Frequency	Percentage%
5.0	1	0.3	2	91	26.9
4.5	90	26.6			
4.0	196	58.0	1	196	58.0
3.5	43	12.7	0	51	15.1
3.0	5	1.5			
2.5	2	0.6			
2.0	1	0.3			

4.1.4. Data integration

Table 3Table 4 shows the variable information of hotel-related attributes from TripAdvisor data and NYC Open Data processing, which are prepared as independent and dependent variables in the machine learning part. There are 39 variables in total, in which 34 are applied as independent variables, "Value" is applied as the dependent variable, and RatingValue, Location, Cleanliness, and Service will be removed in the data set for machine learning models, because these five variables are all the dependent variable candidates that TripAdvisor reviewers judge at the same time, and they are based on reviewers' perception. For RatingValue, Location, Cleanliness, and Service, they are not as compound as value, as they already get more reliable rating scores from the reviewers, whereas value is chosen as the more difficult target variable that we would like to have more insight into.

Table 4 Variable information of hotel-related attributes

Variable Name	Variable Description	Data Category
RoomNumber	the number of hotel rooms	hotel attribute
TravelerRatingExcellent	the number of Excellent rating by the traveler	
TravelerRatingVeryGood	the number of Very Good rating by the traveler	
TravelerRatingAverage	the number of Average rating by the traveler	
TravelerRatingPoor	the number of Poor rating by the traveler	

TravelerRatingTerrible	the number of Terrible rating by the traveler	
LowestPrice	the lowest price of price range based on average rates for a standard room (unit: $)	
HighestPrice	the highest price of price range based on average rates for a standard room (unit: $)	
AvgPrice	average price based on the lowest and the highest price (unit: $)	
RatingValue	the overall rating score of the hotel	
Location	the rating score of hotel location	
Cleanliness	the rating score of hotel cleanliness	
Service	the rating score of hotel service	
Value	the rating score of hotel perceived value	
AmenityNum	the number of hotel amenities listed	
HotelRanking2014	the rank of the hotel within 466 hotels in New York City in the year 2014	
LanguageCount	the number of languages that the hotel service contains	
CityCount	the number of cities that one hotel's reviewers come from	reviewer-related attribute
MeanDistance	the average distance between one hotel and reviewer cities	
StdDistance	the standard deviation of the distance between one hotel and reviewer cities	
NearestAirport	time cost to LaGuardia Airport by car (unit: min)	built-environment attribute
NearestSubway	average time cost to the two nearest subway stations on foot (unit: min)	
RestaurantsWithin	the number of restaurants within 0.3 miles	
RestaurantMin	the minimal time cost to the four nearest restaurants on foot (unit: min)	
RestaurantAvg	average time cost to the four nearest restaurants on foot (unit: min)	
AttractionsWithin	the number of attractions within 0.3 miles	
AttractionMin	minimal time cost to the four nearest attractions on foot (unit: min)	
AttractionAvg	average time cost to the four nearest attractions on foot (unit: min)	
TobaccoLicenseCount	the number of active Tobacco Retail Dealer Licenses based on community district code within the NYC area	
ElecCigaLicenseCount	the number of active Electronic Cigarette Retail Dealer Licenses based on community district code within the NYC area	
SidewalkCafeCount	the number of Active Sidewalk Café Licenses based on zip code within the NYC area	
LibraryCount	the library number count based on zip code within the NYC area	
FilmLocationCount	the film location number count based on zip code within the NYC area	
WifiCount	the Wi-Fi Hotspot number count based on zip code within the NYC area	
2010 Population	population in the year 2010 based on community district code within the NYC area	
ParkAreaACRES	the park area summation based on zip code within the NYC area (unit: acre)	
GardenCount	the Greenthumb Community Garden number count based on zip code within the NYC area	
TreeCount	the Street Tree number count in the year 2015 based on zip code within the NYC area	
RecyclingBinCount	the recycling bin number count based on zip code within the NYC area	

4.2. Machine learning

Once the variable data are prepared, machine learning methods are applied to deal with the problems of multiple variables. This part aims at figuring out the optimal machine learning method and finding out the

influential variables regarding hotel perceived value. Figure 15 shows the overall procedure of the machine learning part.

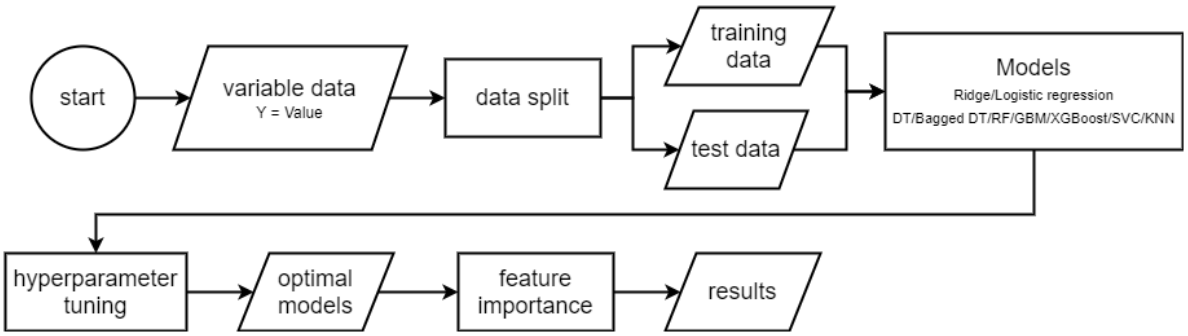


Figure 15 Machine learning procedure overview

4.2.1. Model selection

There are nine machine learning algorithms involved based on the experience of predecessors' research in 2.3: Ridge classifier, Logistic regression, Decision Tree classifier, Bagged Decision Tree, Random Forest classifier, Gradient Boosting Machine, XGB classifier, Support Vector classification, and K-Nearest Neighbors. Table 5 shows the nine machine learning methods and their corresponding Python packages (both for classification and cross-validation) applied in this study.

Table 5 List of machine learning methods and Python packages

Method	Package in Python
Ridge Classifier	sklearn.linear_model.RidgeClassifier sklearn.linear_model.RidgeClassifierCV
Logistic Regression	sklearn.linear_model.LogisticRegression sklearn.model_selection.RepeatedStratifiedKFold sklearn.model_selection.GridSearchCV
Decision Tree	sklearn.tree.DecisionTreeClassifier sklearn.model_selection.RepeatedStratifiedKFold sklearn.model_selection.GridSearchCV
Bagged Decision Tree	sklearn.ensemble.BaggingClassifier sklearn.model_selection.RepeatedStratifiedKFold sklearn.model_selection.GridSearchCV
Random Forest	sklearn.ensemble.RandomForestClassifier sklearn.model_selection.RepeatedStratifiedKFold sklearn.model_selection.GridSearchCV
Gradient Boosting Machine (GBM)	sklearn.ensemble.GradientBoostingClassifier sklearn.model_selection.RepeatedStratifiedKFold sklearn.model_selection.GridSearchCV
XGBoost	xgboost.XGBClassifier sklearn.model_selection.RepeatedStratifiedKFold sklearn.model_selection.GridSearchCV
Support Vector Classification (SVC)	sklearn.svm.SVC sklearn.model_selection.RepeatedStratifiedKFold sklearn.model_selection.GridSearchCV
K-Nearest Neighbors (KNN)	sklearn.neighbors.KNeighborsClassifier sklearn.model_selection.RepeatedStratifiedKFold sklearn.model_selection.GridSearchCV

The first two algorithms belong to linear models, and the rest are non-linear models. Decision Tree refers to tree algorithm. Bagged Decision Tree and Random Forest can both be regarded as a combination of random decision trees and belong to the bagging method. The only difference is the feature selection within the models. Further, XGBoost is the evolution result of Random Forest, as with the Gradient Boosting Machine (GBM), developing as boosting methods. SVC comes from the Support Vector Machine algorithm, of which kernel is the core to handle issues of linear inseparability. K-Nearest Neighbors (KNN) is one of the straightforward classification methods in the data mining area by computing K samples, which are closest or most similar to test samples.

Fundamental theories of those algorithms are explained below:

(1) Ridge Classifier

Ridge classifier is a classification model using Ridge Regression (Hoerl & Kennard, 1970). The basic idea of the Ridge classifier is treating the task as a regression problem by converting the target values into binary values. While in this case, multiclass requires training in a one-versus-all approach.

(2) Logistic Regression

Logistic regression (Yu, Huang, & Lin, 2011) is not a regression model but a commonly-used classification model. It is also called the Logit classifier or MaxEnt classifier. Logistic regression is used for modeling binary variables (0 or 1) to estimate the possibility. In this case, a one-vs-rest (OvR) scheme or cross-entropy loss is applied in the training algorithm for multiclass.

(3) Decision Tree

The Decision Tree model (Breiman, Friedman, Olshen, & Stone, 1984) is a tree structure used for classification and regression. The decision tree is composed of nodes and directed edges. Generally, a decision tree contains a root node and a number of leaf nodes. The decision process of the decision tree needs to start from the root node, test the corresponding feature attributes in the item to be classified, and select the output branch according to its value until it reaches the leaf node, and use the class stored in the leaf node as the decision result.

(4) Bagged Decision Tree

A bagging classifier (Breiman, 1996) is a kind of ensemble algorithms. The basic idea of a bagging classifier is to aggregate basic classifiers' predictions to form a final prediction. It helps to reduce the variance when Decision Tree or other classifiers are applied as the estimator by introducing randomization in its construction process.

(5) Random Forest

Random Forest (Breiman, 2001) is an ensemble algorithm based on the Decision Tree. It is designed to reduce overfitting and variance by using guided aggregation algorithms (bagging algorithms). It is simple and easy to implement, with low computing cost.

(6) Gradient Boosting Machine

Gradient Boosting classifier is the classification model of Gradient Boosting (Friedman, 1999, 2001), which forms a final prediction model out of an ensemble of weak prediction models. Different from the Bagging method, boosting assign weights to the observations trying to reduce variance.

(7) XGBoost

XGBoost is the abbreviation for Extreme Gradient Boosting (Chen & Guestrin, 2016), which is a kind of algorithms as gradient boosted decision tree implementation. It is an improvement to the boosting algorithm based on GBDT (Gradient Boosted Decision Tree).

(8) Support Vector Classification

Support Vector Classification (SVC) is the classification model of Support Vector Machine (SVM)(Platt, 1999). The main point of SVM is to build an optimal decision hyperplane to make the two classes that are closest to the plane have the maximal distance on either of the plane, thereby generalize the classification problems better. Compared with other training classification algorithms, SVM requires relatively fewer samples under the same problem complexity. Furthermore, since kernel functions are introduced into SVM, it is also easy for SVM to deal with high-dimensional samples.

(9) K-Nearest Neighbors

K-Nearest Neighbors (KNN)(Coomans & Massart, 1982) classifies by measuring the distance between feature values. The basic idea of KNN is that: for a certain sample, if most of the nearest K samples in the feature space pertain to a certain class, then the sample also pertains to the class. That is to say, the method only acts on the class of samples according to the nearest one or several (K) samples. It also shows that the result of the KNN algorithm depends largely on the choice of K.

4.2.2. Cross-validation and model evaluation

RidgeClassifierCV is adopted as the built-in function of Ridge classifier for cross-validation. For most algorithms, there are no built-in cross-validation functions in their python modules. Therefore, the RepeatedStratifiedKFold function of sklearn is adopted to apply stratified 10-fold three times with different randomization. The GridSearchCV is introduced to perform an exhaustive search on the specified parameter values of the estimator. The specific evaluation indicator for the classification problem is the confusion matrix shown in Figure 16.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 16 Confusion matrix

Typically, classification accuracy is used to evaluate the model performance, but it is not enough to have a complete judgment of the model. In this case, accuracy score, precision score, recall score, and F1-score are all introduced for model evaluation. The equations of these four scores are listed below (Equation 2 to Equation 5):

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}$$ Equation 2

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$ Equation 3

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$ Equation 4

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$ Equation 5

From the equation composition, it can be noticed that accuracy assigns an equal cost for both True Positive and True Negative, which cannot really tell how good the model regarding the problem. Therefore, Precision and Recall are introduced to evaluate how a sample classified as positive is truly positive and how the class is correctly labeled. F1 is then applied to measure both of the Precision and Recall, whose score will always be closer to the smaller one between the Precision and Recall. Table 6 shows the functions used for the four assessment methods in Python.

Table 6 Matrics score functions in Python

Evaluation method	Function in Python
accuracy	sklearn.metrics.accuracy_socre
precision	sklearn.metrics.precision_score
recall	sklearn.metrics.recall_score
F1	sklearn.metrics.f1_score

5. RESULTS

In this section, the results of this study are illustrated from three aspects: descriptive statistics of the variables, including the dependent variables in three categories and the independent variable "value", cross-validation and model selection result, and feature importance ranking result.

5.1. Descriptive statistics

Table 7 provides the descriptive statistics for 34 independent variables and the dependent variable of the hotel. The 34 independent variables are the variables we picked, representing the hotel-related attributes, which are divided into three categories: hotel attribute, reviewer-related attribute, and built-environment attribute. The hotel attribute occupies 12 in 34 features. The reviewer-related attribute only takes up 3 in 34 features, whereas the built-environment attribute dominates the feature sample. The ranges of the variables differ a lot as well, and most of the variables belong to ratio data.

Table 7 Descriptive statistics for the variables of hotel

	Mean	Median	Min.	Max.	Std.	Nature	Unit
Dependent variable							
Value	1.12	1	0	2	0.64	Ordinal	
Hotel attribute							
RoomNumber	243.32	166.5	14	1957	271.54	Ratio	
TravelerRatingExcellent	1298.32	829	0	7326	1284.71	Ratio	
TravelerRatingVeryGood	888.23	611	1	8206	983.59	Ratio	
TravelerRatingAverage	352.06	198.5	0	4942	513.25	Ratio	
TravelerRatingPoor	144.51	69	1	2561	249.99	Ratio	
TravelerRatingTerrible	120.28	54	1	2487	223.66	Ratio	
LowestPrice	175	142.5	54	1144	124.4	Ratio	$
HighestPrice	489.48	423	89	3833	309.96	Ratio	$
AvgPrice	332.24	285	79.5	1994	198.39	Ratio	$
AmenityNum	20.91	21	2	54	8.89	Ratio	
HotelRanking2014	189.59	185.5	1	446	110.83	Interval	
LanguageCount	2.59	2	1	5	1.24	Ratio	
Reviewer-related attribute							
CityCount	23.07	18	2	86	16.43	Ratio	
MeanDistance	5239.58	5223.1	1769.51	12525.26	1453.86	Ratio	km
StdDistance	4412.9	4457.8	203.01	7066.64	1022.12	Ratio	km
Built-environment attribute							
NearestAirport	6.26	6	4	9	0.98	Ratio	min
RestaurantsWithin	398.15	439	43	629	143.46	Ratio	
AttractionsWithin	104.49	95	3	270	60.87	Ratio	
NearestSubway	3.84	3.5	1.5	11.5	1.52	Ratio	min
RestaurantMin	2.84	3	1	6	1.31	Ratio	min
RestaurantAvg	4.35	4.25	1.75	6	0.85	Ratio	min
AttractionMin	2.96	3	1	6	1.19	Ratio	min
AttractionAvg	4.4	4.5	2.25	6	0.83	Ratio	min
ElecCigaLicenseCount	125.87	117	26	167	39.01	Ratio	

FilmLocationCount	7.63	7	0	16	5.27	Ratio	
GardenCount	1.36	1	0	14	1.88	Ratio	
LibraryCount	1.24	1	0	3	1.07	Ratio	
ParkAreaACRES	21.56	0.69	0	421.22	74.02	Ratio	acre
RecyclingBinCount	4.18	2	0	20	4.77	Ratio	
SidewalkCafeCount	25.43	28	0	79	19.95	Ratio	
TobaccoLicenseCount	204.99	174	106	285	73.05	Ratio	
TreeCount	1317.87	1132	48	3570	729.91	Ratio	
WifiCount	55.72	50	1	110	29.34	Ratio	
2010Population	94269.33	60978	51673	219920	52349.57	Ratio	

5.2. Cross-validation and model selection

In this section, cross-validation was applied for hyperparameter tuning. Table 8 shows the parameter setting and selection result of the nine models, which are Ridge classifier, Logistic regression, Decision Tree classifier, Bagged Decision Tree, Random Forest classifier, Gradient Boosting Machine, XGB classifier, Support Vector classification, and K-Nearest Neighbors.
Table 9 illustrates the comparison result of the performance of the nine machine learning models. XGB classifier is proved to be the best-performed model.

As is mentioned in 4.2.2, RidgeClassifierCV and GridSearchCV function are adopted for hyperparameter tuning. The "parameter selection" column in Table 8 is the result of the 10-fold cross-validation running three times, concerning accuracy as the evaluation indicator of model performance with a different combination of parameters. The model.best_score_ and model.best_params_ function is applied.

The model performance comparison results in
Table 9 are rearranged the column order based on the F1-score. Among them, the XGB classifier shows the best performance of which indicators are as high as 0.8. Logistic regression, Random Forest, and Gradient Boosting classifier present good outcomes, where the indicator range is between 0.75 and 0.8. Bagged Decision Tree offers medium results, whereas Decision Tree, Support Vector classification, Ridge classifier, and K-Nearest Neighbors perform less well.

Table 8 Model hyperparameter tuning

Model	Parameter setting	Parameter selection
Ridge Classifier	alphas = [0.1,1,5,10,20,30]	Alpha=1.0
Logistic Regression	solvers = ['newton-cg', 'lbfgs', 'liblinear'] penalty = ['l2'] c_values = [100, 10, 1.0, 0.1, 0.01]	C=100, penalty='l2', solver='lbfgs'
Decision Tree	criterion = ['gini', 'entropy'] splitter = ['best', 'random'] max_depth = [*range(1,10)] max_features = [None, 'auto', 'sqrt', 'log2'] min_impurity_decrease = [*np.linspace(0,0.5,21)]	criterion='gini', max_depth=3, max_features=None, min_impurity_decrease=0.0, splitter='best'
Bagged Decision Tree	n_estimators = [*range(5,100,5)]	n_estimators=70
Random Forest	n_estimators = [*range(100,1000,100)] max_features = [*range(1,20)]	n_estimators=500, max_features=15, criterion='gini',

	criterion = ['gini','entropy'] max_depth = [*range(1,10)]	max_depth=4
Gradient Boosting Machine (GBM)	n_estimators = [*range(100,1000,100)] learning_rate = [0.001, 0.01, 0.1] subsample = [0.5, 0.7, 1.0] max_depth = [*range(1,10)]	learning_rate=0.01, max_depth=5, n_estimators=500, subsample=0.5
XGBoost	max_depth = [*range(3,9)] learning_rate = [0.001,0.01,0.02,0.03,0.04,0.05,0.08,0.1,0.2,0.3] verbosity = [0,1,2,3] n_estimators = [*range(100,1000,100)] subsample = [0.5, 0.7, 1.0] gamma = [*np.linspace(0,0.1,1)] min_child_weight = [1, 2, 3] colsample_bytree = [*np.linspace(0.2,0.1,1)]	max_depth=6, colsample_bytree=1.0, n_estimators=200, learning_rate=0.01, min_child_weight=2, subsample=0.6, verbosity=0
Support Vector Classification (SVC)	kernel = ['poly', 'rbf', 'sigmoid'] C = [50, 10, 1.0, 0.1, 0.01] gamma = ['scale']	C=1.0, gamma='scale', kernel='sigmoid'
K-Nearest Neighbors (KNN)	n_neighbors = [*range(1, 21, 1)] weights = ['uniform', 'distance'] metric = ['euclidean', 'manhattan', 'minkowski']	metric='euclidean', n_neighbors=13, weights='uniform'

Table 9 Model performance based on 10-fold cross-validation

	XGBoost	Logistic	RF	GBM	Bagged DT	DT	SVC	Ridge	KNN
Accuracy	0.8000	0.7765	0.7803	0.7491	0.7255	0.6706	0.6941	0.6706	0.6471
F1	0.8002	0.7798	0.7769	0.7514	0.7275	0.6686	0.6566	0.6421	0.5837
Recall	0.8000	0.7765	0.7803	0.7491	0.7255	0.6706	0.6941	0.6706	0.6471
Precision	0.8009	0.7941	0.7765	0.7551	0.7321	0.6711	0.7001	0.6948	0.6333

Ridge: Ridge classifier; Logistic: Logistic regression; DT: Decision Tree classifier; Bagged DT: Bagging classifier with Decision Tree; RF: Random Forest classifier; GBM: Gradient Boosting classifier; XGBoost: XGB classifier; SVC: Support Vector classification; KNN: K-Nearest Neighbors.

5.3. Feature importance

The variable data applied in this study can be divided into three parts: hotel attributes, reviewer-related attributes, and built-environment attributes. Therefore, our results are expanded based on the three data categories. Figure 17, Figure 18 and Figure 19 classified the data category with different colors, which intuitively shows the distribution of three types of hotel-related attributes.

In order to figure out which feature plays an important role in assessing hotel perceived value, feature importance function is applied based on the four models that are doing better. According to Elith, Leathwick, & Hastie (2008), the feature importance score of the boosted algorithm (in this case it is appropriate for XGBoost, RF and GBM) is the relative variable importance, of which measures are based on the number of splits of the selected variables, weight by the square improvement of the model caused by each split, and all trees are averaged. The contribution of each variable is resized, and the total value of each impact is 1. A larger number indicates a greater impact on the response.

Figure 17 demonstrates the feature importance of the XGB classifier, where HotelRanking2014, TravelerRatingTerrible, and TravelRatingPoor seem like the most important features regarding hotel perceived value, all of which belong to hotel attributes. The following are the reviewer-related attribute CityCount and a series of hotel attributes TravelerRatingExcellent, AvgPrice, TravelerRatingVeryGood, TravelerRatingAverage, RoomNumber, LowestPrice, and HighestPrice. The built-environment attribute RestaurantAvg goes after, which ranks between the two price range features. For the rest ranking features, only LanguageCount and AmentyNum are related to hotel attributes with MeanDistance and StdDistance belonging to reviewer-related attributes. The remaining features all refer to built-environment attributes.

In this case, the hotel attribute occupies an important position among these features regarding hotel perceived value. The reviewer-related attribute shows part importance, whereas the built-environment attribute does not show its prominent significance according to the XGB classifier model result.

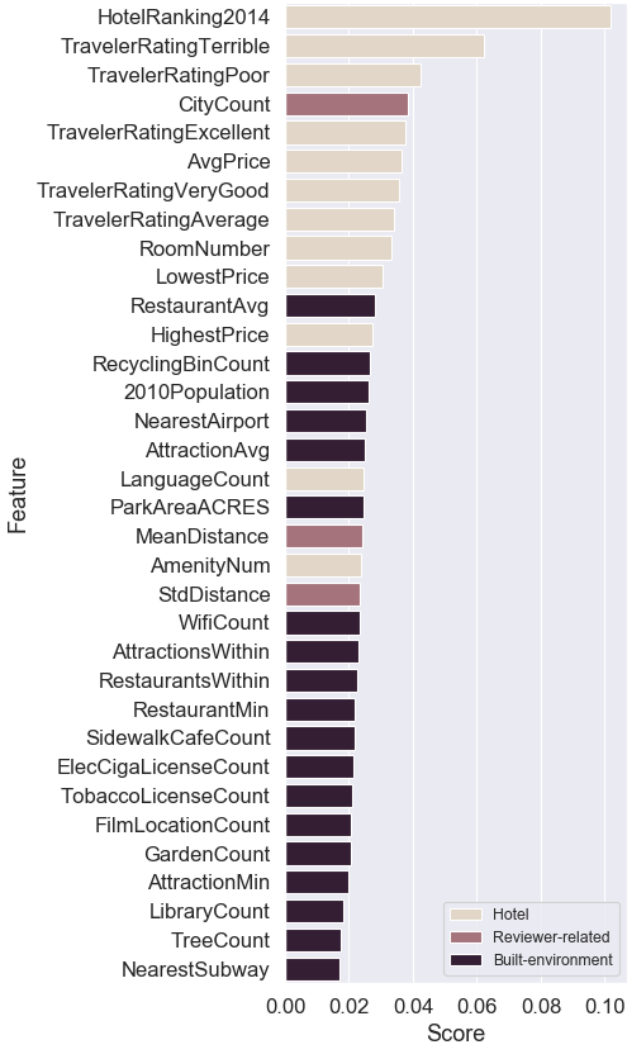


Figure 17 XGB classifier feature importance

Of particular interest here is the Logistic regression algorithm. It returns a multi-dimensional array for feature importance attribute in the shape of (n_classes, n_features) so that it works out a coefficient array for specific labels. Figure 18 shows the feature importance ranking of Logistic Regression for three classes,

where (a) represents Class 0, i.e., Bad Value; (b) represents Class 1, i.e., Average Value; (c) represents Class 2, i.e., Good Value. For the Logistic Regression, the feature importance score can also be called the coefficient. The higher the coefficient, the more important the feature. In the meanwhile, large negative coefficient signifies higher importance in the classification of the negative class and vice versa.

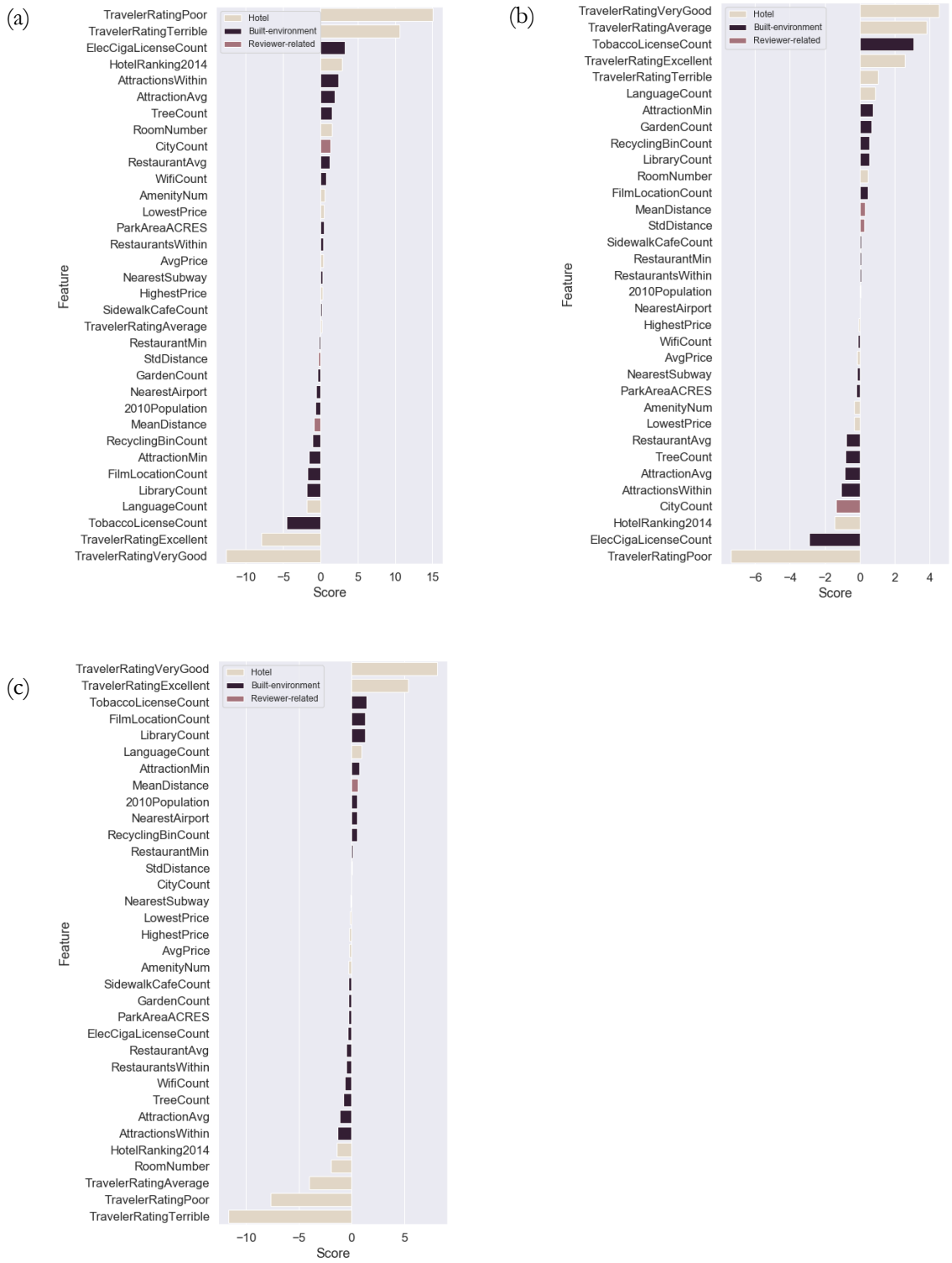


Figure 18 Logistic regression feature importance

By comparing the three separate importance ranking diagrams, we found that TravelerRatingPoor and TravelerRatingTerrible are the top two positively important features for Class 0, namely Bad Value. TravelerRatingVeryGood and TravelerRatingAverage are the most important positive features for Class 1, representing Average Value, followed by TobaccoLicenseCount and TravelerRatingExcellent. As for Class 2, i.e., Good Value, TravelerRatingVeryGood, and TravelerRatingExcellent are ranking in the first two positive positions. In the other direction, TravelerRatingVeryGood and TravelerRatingExcellent are the top two negatively influential features for Class 0, TravelerRatingPoor is the most significant negative feature for Class 1, and TravelerRatingTerrible and TravelerRatingPoor are ranking in the first two negative positions for Class 2. The three ranking diagrams show consistent correspondence regarding the hotel's perceived value and traveler rating value.

As references of XGB classifier, Figure 19 displays the feature importance of Random Forest Classifier (a) and Gradient Boosting Classifier (b), in which HotelRanking2014 and TravelerRatingTerrible are ranking the top two as well, followed by other traveler rating values. RoomNumber, CityCount, and LowestPrice are commonly occupying a space. In general, the ranking diagrams present consistency.

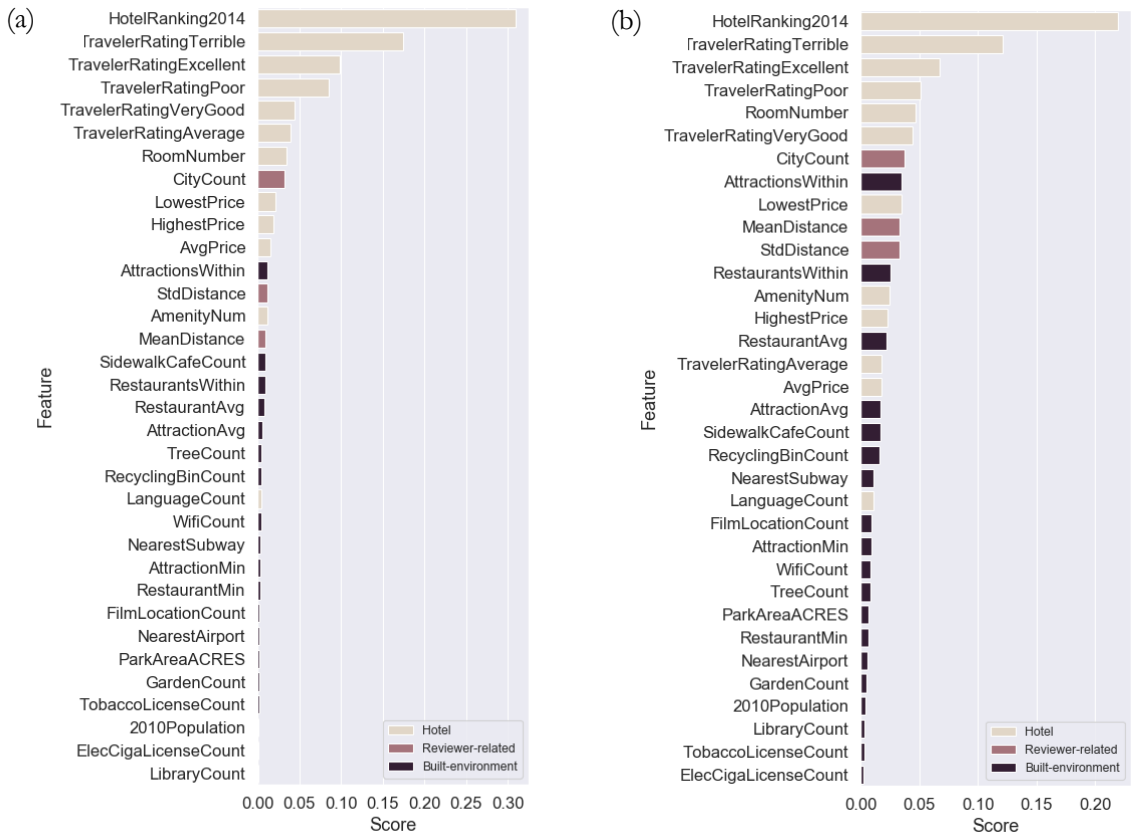


Figure 19 Feature importance of Random Forest classifier (a) and Gradient Boosting classifier (b)

6. DISCUSSION

In this section, two aspects of reflection are displayed regarding the sub-objectives, which are the processing procedure of geodata and the analysis of feature contribution to hotel perceived value.

6.1. Geodata processing

In this study, we applied geodata from TripAdvisor and NYC Open Data and processed them in different ways, including geocoding, measuring distance, web crawling, different data type handling in Python, and so forth. The geocoding method helps to transform the address text into coordinates and the corresponding city names. Distances are measured based on the coordinates. To collect as many hotel attributes as possible through the TripAdvisor platform, we applied web crawling as information collectors. Web crawling is challenging with lots of things to be noticed, such as the changeable XPath due to the regular update of TripAdvisor site, the combined usage of Python and Hawk due to the limited functions of Hawk, and so on. The method is efficient, and the collection results are complete. As for the NYC Open data processing, the procedure follows a normal routine because most of the data are in csv format. Pandas package is simply applied to process these data, whereas handling xml format data (film location data) is complicated owing to the complete unified element tag of it. In 4.1.3.3, value reclassification is applied to handle imbalanced classes. It does solve the imbalanced issue while also causes high similarity among the three classes, which could be one reason for the low accuracy. Due to the lack of data, the problem cannot be solved. However, it also gives us some reflection that during the data processing, we need to keep an eye on data balance and distinction at the same time.

We applied nine machine learning algorithms when modeling the hotel's perceived value classification: Ridge classifier, Logistic regression, Decision Tree classifier, Bagged Decision Tree, Random Forest classifier, Gradient Boosting Machine, XGB classifier, Support Vector classification, and K-Nearest Neighbors. Generally, the boosting and bagging algorithms performed better, whereas Logistic regression surprisingly showed a relatively good performance because it is more commonly used when the dependent variable is binary. The XGB classifier surpasses all the other algorithms by not only having the highest accuracy but also being highly efficient as it runs faster than Random Forest classifier and Gradient Boosting classifier in the model training part. It is proved that model training and selection are of vital importance regarding different kinds of data. Researchers cannot merely pick a widely used machine learning method as their study model without any preparatory work.

We also considered the reasons for those low-performing algorithms. K-Nearest Neighbors displays the worst results due to its high dependence on the distance, which is a disadvantage for high dimensional data. Ridge classifier used to be applied for binary class rather than multiclass data sets. The results also verify that the Support Vector classifier does not perform well on skewed/imbalanced data sets, while Logistic regression is just the reverse. Though the imbalanced classes were handled in 4.1.4.3, it still cannot be called equilibrium. Meanwhile, Decision Tree usually performs unstable and sensitive: tiny perturbations could lead to a different tree. It is also easier to overfit than other tree algorithms, which is also one of the reasons that we apply ensemble algorithms in the experiment.

6.2. Feature importance

The discussion of feature importance is expanded based on the three data categories as well. Table 4 in 4.1.4 shows detailed information about the 34 variables in total that represents hotel-related attributes in

three categories: hotel attributes, reviewer-related attributes, and built-environment attributes. Hotel attributes, reviewer-related attributes, and part of the built-environment attributes are mainly collected and calculated from the information on the TripAdvisor site. The rest of the built-environment attributes are obtained from the NYC Open Data site. Among these attributes, the number of rooms, price, rating score, accessibility to attractions/restaurants/traffic stations, and green space are all studied in previous hotel industry research, and the rest attributes are considered likely to contribute to hotel perceived value judgment.

For many machine learning methods, multicollinearity is not an issue because it does not influence the prediction effect of a model as long as the model is not over-fitting, which is also one of the reasons why we apply machine learning algorithms rather than simple regression methods in this study. However, multicollinearity should be taken into consideration when individual features' impact is judged. Therefore, we should be careful when analyzing feature contribution as it might not be reliable enough. It is also noticeable that when we try to predict the perceived value of the hotel, the features that involve in the machine learning part should all be put into the model as a "work together" effect, in order to predict the same accurate results. Since all the features that we have are from open-sourced data, it is not difficult to obtain them and apply them in the model. In addition, since the number of the data source is limited in this study, and the case is only taken in New York City, the user should be careful when generalizing the results.

Since the general objective of this study is to understand how hotel-related attributes can predict hotel perceived value, three categories of features are now focused. Figure 17, Figure 18, and Figure 19 all show that most of the hotel attributes rank by the front. Among them, HotelRanking2014 in Figure 17 and Figure 19 comes top of the list, indicating that hotel ranking is possibly instrumental in assessing hotel perceived value in the model. The ranking seems like a compound variable similar to hotel perceived value. However, they are not an approximation. According to the introduction from the TripAdvisor website, the hotel ranking defined as "Popularity Ranking", which is based on the quality of reviews, the number of reviews and recency of reviews ("Tripadvisor Popularity Ranking: Key Factors and How to Improve | TripAdvisor Insights," n.d.). In addition, the principal component analysis (PCA) is applied, and the result shows that HotelRanking2014 is a relatively independent variable among all these features, proving that the variable is not impacted by multicollinearity. Therefore, improving the hotel ranking could be a suggestion for a hotel regarding its perceived value.

The results also show that TravelerRatingTerrible ranks higher than TravelerRatingExcellent and much higher than TravelerRatingVeryGood and TravelerRatingAverage; it suggests that positive and negative reviews might be related to hotel perceived value evaluation. In contrast, medium ratings are likely to be overlooked easily. Meanwhile, the number of negative reviews occupy principal positions. One of the possibilities could be that visitors are more concerned about negative reviews. Research of Gavilan, Avello, & Martinez-Navarro (2018) also come up with a similar conclusion that the trust of reviewers on rating varies: the trust of good rating depends on review number, which does not affect the trust of a bad rating. Price and room number are the most extensively studied factors and proved to be primary in the early hotel research model (Zeithaml, 1988; Dodds et al., 1991; Bojanic, 1996; O'Neill, 2004), while in recent years, reviews are getting more attention and concern (Xie et al., 2014; Phillips, Zigan, Silva, & Schegg, 2015; Gavilan et al., 2018), which shows the trend that the importance of price declines and people rely more on ranking and rating as a result of the expanding use of social media (Phillips et al., 2015). Language is the least important issue among the hotel attributes, from which we can infer that language is not a problem for most of the visitors. "Mother tongue" might be a plus for hotel service, while it may not be a requirement and will have little influence on the visitors' experience in the hotels.

There are only three reviewer-related features. CityCount ranks slightly ahead of the other two, which seems associated with hotel perceived value contribution. It is a relatively new feature in the hotel research area. Normally, researchers list specific cities or countries to understand the basic situation of the tourism and hotel industry (Go et al., 1994), which is helpful in providing suggestions for hotels to formulate future plans. In this case, larger CityCount represents more cities, meaning stronger preception from visitor groups, which shows the city diversity. Regarding the results of the Logistic regression model, the number of cities shows an inverse contribution in the ranking of Figure 18 (a) and Figure 18 (b), i.e., positive for bad value and negative for the average value. We infer that hotels with higher city diversity from visitors would contribute positively to bad perceived value and positively to average perceived value. One possibility for this result could be that reviewers feel fine with meeting visitors from fewer cities; in other words, they may have more sense of belonging when they meet more visitors from the same places. As for the two distance-related attributes, i.e., MeanDistance and StdDistance, they show no significant relationship with hotel perceived value classification, which means that distance does not visibly relate to the visitors' judgment on hotel's perceived value.

As for the Built-environment attributes that make up the largest share of the features, they are placed on the lower half of the ranking. Among them, the accessibility/convenience to restaurants, attractions, and airports instead of their numbers slightly show the position in the feature importance ranking, which correspond with the research of Yang et al. (2018) and Li et al. (2013). Simultaneously, no clear evidence shows that green space is essential to reviewers' satisfaction, which shows discrepant results regarding the research of Yang et al. (2018). Other local facilities data such as library number, Wi-Fi hotspot number, Park area, and population do not visually suggest the connection with the perceived value of the hotel as well. One possibility could be that visitors care more about transportation convenience rather than near-by infrastructures and local livelihoods. Another inference could be that visitors do not count these aspects into the hotel's perceived value components. The overall built-environment attributes do not display a significant relationship regarding hotel perceived value.

To connect with the literature findings, as many of them adopted the research method by doing investigation such as making a questionnaire for consumers (Callan & Bowman, 2000; Al-Sabbahy et al., 2004; Casidy, Wymer, & O'Cass, 2018), there also exist the remaining problem of the reliability of "value" which we posted at the very beginning. As we all know, doing an investigation is the most direct way to get perceived opinions on specific target issues from people, whereas writing a review or scoring hotel value is not as straightforward as that. "Value" is such a compound and subjective assessment for the visitors, that they may not carefully think about what "Value" means. However, they may directly mark a score simply based on the existed comments and scores when visitors are writing reviews, which might be the reason why the traveler rating numbers show a good correlation with it. Phillips et al. (2015) and Blal & Sturman (2014) also pointed out that Electronic word-of-mouth (eWOM) reviews have had a major impact on the decisionmakers. Therefore, to improve the online "value" feature, another suggestion could be that managers pay more attention to the management of online comments.

7. CONCLUSION AND OUTLOOK

In summary, this study demonstrated the influential features towards hotel perceived value in New York City, seeking the relationship between the hotel-related attributes and hotel perceived value by approaches of geodata processing and machine learning. The hotel-related attributes contain three components (sub-objective(1)): hotel attributes, reviewer-related attributes, and built-environment attributes, with some have been studied, and some have not, in order to verify the previous studies and discover new possibilities. We compare nine machine learning methods in total (sub-objective(2)), including Ridge classifier, Logistic regression, Decision Tree classifier, Bagged Decision Tree, Random Forest classifier, Gradient Boosting Machine, XGB classifier, Support Vector classification, and K-Nearest Neighbors. After tuning hyperparameter via cross-validation, the XGB classifier is selected as the best-performed algorithm for modeling value classification, of which the accuracy and other indicators reach up to 0.8. Feature importance of the well-performed models is displayed, where hotel ranking and negative review amount show a relatively strong relationship with hotel perceived value classification (sub-objective(3)). It suggests that reviewers might be more susceptible to the hotel ranking and negative comments when they are judging the value of the hotel. As price and service quality used to be of much concern in visitor satisfaction (Bojanic, 1996; Oh, 1999; Callan & Bowman, 2000; Sweeney & Soutar, 2001), we infer that the price's importance is declining as time goes by. Part of built environment attributes plays a bit role, such as the transportation convenience, while most of them take a back seat based on the importance ranking. The possibility could be that the contribution to hotel perceived value from the built environment is not as strong as the power of the hotels' attributes. In addition, the number of cities that reviewers come from is also possibly related to hotel perceived value. Considering the impact of multicollinearity, a similar assessment on the TripAdvisor hotel's perceived value requires the whole set of features applied in this study. The portability of the results should also be taken care of due to space restriction and data source limitations. We could also infer that the "value" score on the TripAdvisor platform is not as reliable the perceived value collected directly from people via investigation. It can be concluded that, with the expansion of social media, if hotel managers in New York City want to attract visitors, improve visitor satisfaction and raise revisit rate, possible solutions could be mainly focusing on factors inside the hotel, trying to improve the hotel ranking and removing the influence from negative comments by responding more to the negative reviewers (Xie et al., 2014; Gavilan et al., 2018).

In this study, limitations exist in several aspects, which can be concluded in four dimensions: reviewer classification, review text, reviewer address, and hotel classification. The corresponding improvements for future work are noted as well.

(1) Reviewer classification

We differed the reviewer based on their city location/hometown, while research proved that they vary in many other aspects: gender, age group, country, travel purpose, travel experience, review distribution, etc. (Knutson, 1988; Nyaupane et al., 2003; Petrick, 2004; Gao et al., 2018). In future work, we can also consider the influence of reviewers' difference in hotel perceived value judgment.

(2) Review text

In this study, we only collected the review information in English for better understanding, which means there exist area bias (e.g., Asian area) that may influence the reviewer-related attributes such as LanguageCount, CityCount, and distance-related attributes. In the meanwhile, we did not have a deep

analysis of review text which is already widely studied in lots of research. In future work, we can add semantic analysis for visitor satisfaction determination.

(3) Reviewer address

In this study, we applied the geocoding method combining with Open Street Map API for the reviewer address transformation. The definition of the city caused a drop in numerous data. In future work, we may take rural areas into account, adding variety for the reviewer group so that it could be more representative of the real crowd.

(4) Hotel classification

Due to the information lack and technology limitation, we did not collect more hotel attributes like hotel class and hotel type, which are both considerable issues in hotel research (Cser & Ohuchi, 2008; L. Zhou, Ye, Pearce, & Wu, 2014; Rhee & Yang, 2015; Mariani & Borghi, 2018). In the meanwhile, the hotel attribute AmenityNum could have provided more information about the hotel facility. In future work, we could look into hotel class, hotel type, and hotel amenity in more detail to examine the influence of class, type, and particular hotel facilities on hotel perceived value.

LIST OF REFERENCES

Acock, A. C. (2005). Working With Missing Values. *Journal of Marriage and Family*, *67*(November), 1012–1028.

Al-Sabbahy, H. Z., Ekinci, Y., & Riley, M. (2004). An investigation of perceived value dimensions: Implications for hospitality research. *Journal of Travel Research*, *42*(3), 226–234. https://doi.org/10.1177/0047287503258841

Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Blal, I., & Sturman, M. C. (2014). The Differential Effects of the Quality and Quantity of Online Reviews on Hotel Room Sales. *Cornell Hospitality Quarterly*, *55*(4), 365–375. https://doi.org/10.1177/1938965514533419

Bojanic, D. C. (1996). Consumer Perceptions of Price , Value and Satisfaction in the Hotel Industry: An Exploratory Study. *Journal of Hospitality & Leisure Marketing*, *4*(1), 5–22.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1017/CBO9781107415324.004

Breiman, Leo. (1996). Bagging predictions. *Machine Learning*, *24*(2), 123–140.

Breiman, Leo, Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*.

Callan, R. J., & Bowman, L. (2000). Selecting a Hotel and Determining Salient Quality Attributes: A Preliminary Study of Mature British Travellers. *INTERNATIONAL JOURNAL OF TOURISM RESEARCH*, *2*, 97–118.

Camilla, V. (2011). Complaints online : The case of TripAdvisor. *Journal of Pragmatics*, *43*(6), 1707–1717. https://doi.org/10.1016/j.pragma.2010.11.007

Casidy, R., Wymer, W., & O'Cass, A. (2018). Enhancing hotel brand performance through fostering brand relationship orientation in the minds of consumers. *Tourism Management*, *66*, 72–84. https://doi.org/10.1016/j.tourman.2017.11.008

Chan, E. S. W., & Wong, S. C. K. (2005). Hotel selection: When price is not the issue. *Journal of Vacation Marketing*, *12*(2). https://doi.org/10.1177/1356766706062154

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Coomans, D., & Massart, D. L. (1982). Alternative k-nearest neighbour rules in supervised pattern recognition. Part 1. k-Nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta*, *136*, 15–27. https://doi.org/10.1016/S0003-2670(01)95359-0

Cser, K., & Ohuchi, A. (2008). World practices of hotel classification systems. *Asia Pacific Journal of Tourism Research*, *13*(4), 379–398. https://doi.org/10.1080/10941660802420960

Dodds, W. B., Monroe, K. B., & Grewal, D. (1991). Effects of Price, Brand, and Store Information on Buyers' Product Evaluations. *Journal of Marketing Research*, *28*(3), 307. https://doi.org/10.2307/3172866

Dolnicar, S., & Otter, T. (2003). Which Hotel attributes Matter ? A review of previous and a framework for future research. In *Proceedings of the 9th Annual Conference of the Asia Pacific Tourism Association (APTA)* (pp. 176–188).

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Feizizadeh, B., Roodposhti, M. S., Blaschke, T., & Aryal, J. (2017). Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping. *Arabian Journal of Geosciences*, *10*(5). https://doi.org/10.1007/s12517-017-2918-z

Friedman, J. H. (1999). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.2307/2699986

Gao, B., Li, X., Liu, S., & Fang, D. (2018). How power distance affects online hotel ratings: The positive moderating roles of hotel chain and reviewers' travel experience. *Tourism Management*, *65*, 176–186. https://doi.org/10.1016/j.tourman.2017.10.007

Garcı´a-Pablos, A., Cuadros, M., & Linaza, M. T. (2016). Automatic analysis of textual hotel reviews.

Information Technology & Tourism, *16*, 45–69. https://doi.org/10.1007/s40558-015-0047-7
Gavilan, D., Avello, M., & Martinez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. *Tourism Management*, *66*, 53–61. https://doi.org/10.1016/j.tourman.2017.10.018
Go, F., Pine, R., & Yu, R. (1994). Hong Kong:Sustaining Competitive Advantage in Asia's Hotel Industry. *Cornell Hotel and Restaurant Administration Quarterly*, *35*, 50–61.
Hagenauer, J., Omrani, H., & Helbich, M. (2019). Assessing the performance of 38 machine learning models: the case of land consumption rates in Bavaria, Germany. *International Journal of Geographical Information Science*, *33*(7), 1399–1419. https://doi.org/10.1080/13658816.2019.1579333
Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, *12*(1), 69–82. https://doi.org/10.1080/00401706.1970.10488635
Huang, N., Burtch, G., Hong, Y., & Polman, E. (2016). Effects of Multiple Psychological Distances on Construal Level: A Field Study of Online Reviews. *Journal of Consumer Psychology*. https://doi.org/10.1016/j.jcps.2016.03.001
Kisilevich, S., Keim, D., & Rokach, L. (2013). A GIS-based decision support system for hotel room rate estimation and temporal price prediction: The hotel brokers' context. *Decision Support Systems*, *54*(2), 1119–1133. https://doi.org/10.1016/j.dss.2012.10.038
Knutson, B. J. (1988). Frequent Travelers: Making Them Happy and. *The Cornell Hotel and Restaurant Administration Quarterly*, *29*(1), 82–87.
Kobler, A., & Adamic, M. (2000). Identifying brown bear habitat by a combined GIS and machine learning method. *Ecological Modelling*, *135*(2–3), 291–300. https://doi.org/10.1016/S0304-3800(00)00384-7
Li, H., Ye, Q., & Law, R. (2013). Determinants of Customer Satisfaction in the Hotel Industry: An Application of Online Review Analysis. *Asia Pacific Journal of Tourism Research*, *18*(7), 784–802. https://doi.org/10.1080/10941665.2012.708351
Mariani, M. M., & Borghi, M. (2018). Effects of the Booking.com rating system: Bringing hotel class into the picture. *Tourism Management*, *66*, 47–52. https://doi.org/10.1016/j.tourman.2017.11.006
Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2019). Social Media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, *74*(May 2018), 161–174. https://doi.org/10.1016/j.compenvurbsys.2018.11.001
Mattila, A. S., & O'Neill, J. W. (2003). Relationships between Hotel Room Pricing , Occupancy , and Guest Satisfaction: A Longitudinal Case of a Midscale Hotel in the United States. *Journal of Hospitality & Tourism Research*, *27*(3), 328–341. https://doi.org/10.1177/1096348003252361
Mitchell, T. M. (1997). *Machine Learning*. Singapore: McGraw-Hill.
Mojaddadi, H., Pradhan, B., Nampak, H., Ahmad, N., & Ghazali, A. H. bin. (2017). Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS. *Geomatics, Natural Hazards and Risk*, *8*(2), 1080–1102. https://doi.org/10.1080/19475705.2017.1294113
Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001
Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. (2002). A Comparison of Imputation Techniques for Handling Missing Data. *Western Journal of Nursing Research*, *24*(7), 815–829. https://doi.org/10.1177/019394502237390
New York City Department of City Planning. (2017). *NYC Hotel Market Analysis*.
Nyaupane, G. P., Graefe, A. R., & Burns, R. C. (2003). Does distance matter ? Differences in characteristics, behaviors, and attitudes of visitors based on travel distance. In *Proceedings of the 2003 Northeastern Recreation Research Symposium* (pp. 74–81).
NYC & Company. (2019a). *NYC Hotel Occupancy, ADR & Room Demand*. Retrieved from https://assets.simpleviewinc.com/simpleview/image/upload/v1/clients/newyorkcity/FYI_Hotel_reports_February_2019_8607015b-b32a-4c7f-9fbd-84cd2a93cbe6.pdf
NYC & Company. (2019b). *Travel & Tourism Trend Report*.
NYC & Company. (2020). *Hotel Development in NYC*.
O'Neill, J. W. (2004). An automated valuation model for hotels. *Cornell Hotel and Restaurant Administration Quarterly*, *45*(3), 260–268. https://doi.org/10.1177/0010880404265322
O'Neill, J. W., & Xiao, Q. (2006). The role of brand affiliation in hotel market value. *Cornell Hotel and Restaurant Administration Quarterly*, *47*(3), 210–223. https://doi.org/10.1177/0010880406289070

Oh, H. (1999). Service quality, customer satisfaction, and customer value: A holistic perspective. *International Journal of Hospitality Management*, *18*(1), 67–82. https://doi.org/10.1016/s0278-4319(98)00047-4

Park, S., Yang, Y., & Wang, M. (2019). International Journal of Hospitality Management Travel distance and hotel service satisfaction: An inverted U-shaped relationship. *International Journal of Hospitality Management*, *76*(May 2018), 261–270. https://doi.org/10.1016/j.ijhm.2018.05.015

Petrick, J. F. (2004). First Timers' and Repeaters' Perceived Value. *Journal of Travel Research*, *43*(August), 29–38. https://doi.org/10.1177/0047287504265509

Phillips, P., Zigan, K., Silva, M. M. S., & 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

Platt, J. C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, *10*(3), 61–74.

Rahmati, O., Pourghasemi, H. R., & Melesse, A. M. (2016). Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran. *Catena*, *137*, 360–372. https://doi.org/10.1016/j.catena.2015.10.010

Raza, M., Siddiquei, A., Awan, H., & Bukhari, K. (2012). Relationship Between Service Quality, Perceived Value, Satisfaction and Revisit Intention in Hotel Industry. *Interdisciplinary Journal of Contemporary Research in Business*, *4*(8), 788–805.

Rhee, H. T., & Yang, S. (2015). Does hotel attribute importance differ by hotel? Focusing on hotel star-classifications and customers' overall ratings. *Computers in Human Behavior*, *50*, 576–587. https://doi.org/10.1016/j.chb.2015.02.069

Roof, K., & Oleru, N. (2008). Public health: Seattle and King County's push for the built environment. *Journal of Environmental Health*, *71*(1), 24–27.

Schmunk, S., Höpken, W., Fuchs, M., & Lexhagen, M. (2014). Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC. In *Information and Communication Technologies in Tourism 2014*. https://doi.org/10.1007/978-3-319-03973-2

Sweeney, J. C., & Soutar, G. N. (2001). Consumer perceived value: The development of a multiple item scale. *Journal of Retailing*, *77*(2), 203–220. https://doi.org/10.1016/S0022-4359(01)00041-0

Tehrany, M. S., Pradhan, B., Mansor, S., & Ahmad, N. (2015). Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena*, *125*, 91–101. https://doi.org/10.1016/j.catena.2014.10.017

TripAdvisor. (n.d.). US Press Center | About Tripadvisor. Retrieved from https://tripadvisor.mediaroom.com/us-about-us

Tripadvisor Popularity Ranking: Key Factors and How to Improve | TripAdvisor Insights. (n.d.). Retrieved from https://www.tripadvisor.com/TripAdvisorInsights/w722

Tsai, H., Song, H., & Wong, K. K. F. (2009). Tourism and hotel competitiveness research. *Journal of Travel and Tourism Marketing*, *26*(5–6), 522–546. https://doi.org/10.1080/10548400903163079

TUI. (2019). *Annual Report Tui Group*.

World Travel Organization. (2019). *International Tourism Highlights*. https://doi.org/https://www.e-unwto.org/doi/pdf/10.18111/9789284421152?download=true

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

Yang, Y., Mao, Z., & Tang, J. (2018). Understanding Guest Satisfaction with Urban Hotel Location. *Journal of Travel Research*, *57*(2), 243–259. https://doi.org/10.1177/0047287517691153

Yang, Y., Tang, J., Luo, H., & Law, R. (2015a). Hotel location evaluation: A combination of machine learning tools and web GIS. *International Journal of Hospitality Management*, *47*, 14–24. https://doi.org/10.1016/j.ijhm.2015.02.008

Yang, Y., Tang, J., Luo, H., & Law, R. (2015b). Hotel location evaluation: A combination of machine learning tools and web GIS. *International Journal of Hospitality Management*, *47*, 14–24. https://doi.org/10.1016/j.ijhm.2015.02.008

Yu, H. F., Huang, F. L., & Lin, C. J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, *85*(1–2), 41–75. https://doi.org/10.1007/s10994-010-5221-8

Zeithaml, V. A. (1988). Consumer Perceptions of Price , Quality, and Value: A Means-End Model and Synthesis of Evidence. *Journal of Marketing*, *52*(3), 2–22.

Zhang, J., Ye, Q., & Law, R. (2011). Determinants of hotel room price An exploration of travelers ' hierarchy of. *International Journal of Contemporary Hospitality Management*, *23*(7), 972–981. https://doi.org/10.1108/09596111111167551
Zhang, Y. (2019). *Forecasting Hotel Demand using Machine Learning Approaches*.
Zhou, L., Ye, S., Pearce, P. L., & Wu, M. (2014). Refreshing hotel satisfaction studies by reconfiguring customer review data. *International Journal of Hospitality Management*, *38*, 1–10. https://doi.org/10.1016/j.ijhm.2013.12.004
Zhou, X., Wang, M., & Li, D. (2019). Bike-sharing or taxi? Modeling the choices of travel mode in Chicago using machine learning. *Journal of Transport Geography*, *79*. https://doi.org/10.1016/j.jtrangeo.2019.102479
Zuo, R., & Xiong, Y. (2020). Geodata science and geochemical mapping. *Journal of Geochemical Exploration*, *209*(106431). https://doi.org/10.1016/j.gexplo.2019.106431