Tong Nie

MACHINE LEARNING FOR FORECASTING FUTURE RESERVATIONS' RATINGS

—*Radisson Blu Seaside in Helsinki*

Master's Thesis in Governance of
Digitalization
Supervisors: Prof. József Mezei
Faculty of Social Sciences, Business and
Economics
Åbo Akademi University

Turku 2023

ABSTRACT

Subject: Governance of Digitalization

Writer: Tong Nie

Title: Machine Learning for Forecasting Future Reservations’ Ratings of a Hotel

Supervisor: Prof. József Mezei

Abstract:

In the current age of internet and big data, it is imperative for hotels to enhance their online reputation to remain competitive and profitable. This research presents a new perspective on how hotels can maintain and improve their online reputation through the use of machine learning techniques to predict the ratings of reservations. The approach involves analysing data that customers provide when booking a room. Additionally, the study explores how insights gleaned from online textual reviews can be used by hotel managers to address negative ratings.

The study's primary objective is to assess the effectiveness of machine learning in predicting negative instances, a critical factor in managing online reputation. The best performing models achieved a 60% accuracy in classifying negative instances. However, increasing the number of predicted true negative instances also increased the number of false negative instances. This result was primarily due to the unpredictability of customer behaviour, making it difficult to accurately predict ratings.

Despite not achieving the desired result, this study presents a novel direction for future research and provides suggestions for future research ideas. By utilizing machine learning algorithms to analyse customer data, hotels can better understand their customer's preferences, allowing them to improve their online reputation and ultimately improve their bottom line.

Keywords: big data, data analysis, predictive analysis, online reputation, hotel industry

Date: 01.05.2023

Number of pages: 102

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

With the continuous popularization of mobile internet and information technology, the amount of data generated by human beings is increasing exponentially. The world has evolved into the big data era, where data has become the new oil of the economy (*The Economist* 2017). Far beyond just a fancy expression, big data has begun to infiltrate and disrupt various industries, especially in contributing to the evolution of management strategies and daily operations of many traditional industries. Against this background, the huge value generated by big data has attracted the interest and attention of numerous practitioners and researchers in the hotel industry. For the hotel industry, big data is both an opportunity and a challenge. As public places, hotels welcome a considerable number of visitors every day, and consequently have a unique advantage in collecting big data. However, discussing and mastering how to leverage big data to serve the operations and management of the hotel industry is also a challenge faced by managers and researchers in the hotel industry today.

A substantial number of scholars have conducted their research regarding big data from multiple perspectives, such as the form of big data, the source of big data, and the feature of big data. However, Line et al. (2020) argue that data do not equal knowledge. Thus, only having big data will not necessarily help a company make better data-driven decisions or create value. Indeed, big data itself is not sufficient to generate relevant knowledge; the analytics accompanying big data also has significant impact. The real core of big data applications lies in mining the intelligence contained in the data, rather than performing simple data calculations. Mariani et al. (2021) broadly classify data analytics into four categories: descriptive analytics, exploratory analytics, predictive analytics, and prescriptive analytics. While descriptive analytics and exploratory analytics aim to generate information and knowledge about the past and present, predictive analytics and prescriptive analytics are conducive to generating information and knowledge about the future (Mariani et al. 2021).

In today's era of big data, big data and analytics together have been playing a leading role from a company's strategy level to operational level. They can help in making better data-driven decisions by the management and create value in the business. Predictive analytics in particular is a strategy that is increasingly important to businesses. Using machine learning to analyse data collected from businesses can be utilized to generate more accurate predictions about the future. Kumar et al. (2018) emphasize that predictive analytics can help organisations to become proactive, forward looking and to anticipate trends or behaviour based on data, and it will also be helpful to identify the risks and opportunities for every individual customer. From a different perspective, hotels work with numerous data sources. Mariani et al. (2021) state that enormous amounts of data are produced by both hospitality & tourism service providers and customers. Furthermore, big data in the hotel industry can be subdivided into many types from different perspectives, such as data before check-in and data after check-out, or user-generated content (UGC) and data collected from devices. Among these, Li et al. (2018) argue that UGC is the dominant type of data in tourism research. Therefore, the focus area of this thesis is performing predictive analytics of UGC for hotels. In this thesis, UGC in the hotel industry refers to the ratings and reviews given by guests after their stay at a hotel.

1.1 Objective of the Thesis

Previous researches of predictive analytics in the hotel industry focus on several aspects of hotels from the supply side, such as forecasting arrivals (Sun et al.,2019), forecasting hotel occupancy (Pan and Yang, 2017; Rivera, 2016), forecasting hotel booking cancellations (Sánchez-Medina et al., 2020; Antonio, et al., 2017), forecasting hotel demand (Pereira et al., 2022), forecasting hotel room price (Al Shehhi et al., 2020) and predicting sentiment and rating of tourist reviews (Puh and Babac, 2022). However, forecasting hotels' ratings based on Personal Name Records (PNR) as predictors is not extensively researched. PNR refers to the information provided by customers when they make a reservation, such as, check-in date, check-out date, and room type.

In the era of internet and big data, the ratings and reviews of hotels generated by different consumers online represent the online reputation of the hotels. The online reputation has a huge impact on future acquisition of new reservations, which will ultimately affect the hotels' operational performance and profitability. Therefore, it would be advantageous for a hotel to generate relatively accurate forecasts about a certain future reservation's rating in advance. When the forecasted rating is relatively low, the hotel could perform some actions to improve this future reservation's rating to maintain or even improve the online reputation of the hotel. This thesis addresses this issue by forecasting a future reservation’s rating (UGC) of a hotel based on PNR data, and identifying the most frequent negative textual review of a hotel as the reason for negative ratings. The following research questions are raised:

RQ1: What topics have been covered in previous literatures on using machine learning to make predictions in the hotel industry?

RQ2: Which machine learning models offer the best performance in predicting customer's (numeric) evaluations?

RQ3: What insights can be gained for hotel managers from customers' (text) evaluation?

1.2 Structure of the Thesis

This thesis is divided into seven chapters. This current chapter gives a brief introduction to the thesis topic and the research questions. The remainder of this thesis is structured as follows:

Chapter 2 provides background information about big data and the hotel industry. This chapter discusses the related concepts of big data and data analytics in the context of the hotel industry. How the hotel industry has changed in the era of big data is also discussed in this chapter.

Chapter 3 explains the chosen hotels and the variables of the datasets that are used in this thesis. This chapter presents the reasons for choosing Radisson Blu seaside hotel as the target hotel and choosing booking.com as a data source. The chosen variables are also explained.

Chapter 4 presents the chosen methods used to answer the research questions. In addition, the data pre-processing method is also introduced in this chapter.

Chapter 5 presents the data pre-processing and the results of exploratory data analysis. Furthermore, the results of the research questions are also presents in this chapter.

Chapter 6 discusses the answers to the research questions. Limitations are also described in this chapter.

Chapter 7 summarizes the thesis and presents the main conclusions.

2 BACKGROUND

This chapter aims to provide background information about big data and data analytics, and discusses related concepts in the context of the hotel industry. The changes in the hotel industry and the necessity of forecasting ratings for future reservations in the era of internet and big data will also be discussed.

2.1 Big Data

The advent of big data has brought about significant transformations in numerous industries and has even revolutionized several aspects of our daily routines. However, big data is not mysterious, but an inevitable product of technological development to a certain stage. The concept of "big data" was first proposed in a slide entitled "Big Data and the Next Wave of InfraStress" in the mid-1990s by John Mashey, and was significantly enriched by Douglas Laney at Gartner (Diebold, 2012). Laney (2001) identifies three key features of big data, and defined the 3Vs: *Volume*, *Variety*, and *Velocity*. With the development of technology, new important characteristics of big data are continuously emerging. Many researchers attempt to boost the definition of big data on the basis of Laney's 3Vs by extending with more V features (Wamba et al. 2015). According to Lv et al. (2022), *value* is added to the features of big data by Gantz & Reinsel (2011). Gandomi et al. (2015) state that IBM and SAS added *Veracity* and *Variability* as additional V features to big data respectively, while *Variability* refers to the variation in the data flow rates that indicates the periodic peaks and troughs of big data velocity. Wamba et al. (2015) summarize the details of the first four V features, as are shown in Table 1.

Table 1 Big Data Feature Details

Feature	Details
Volume	Large volume of data that considering either huge storage or large number of records, even calculated in units of PB (1PB = 1024 TB) or billions of records (Russom, 2011; Manyika et al., 2011).

Variety	Greater variety of data sources, formats, and dimensions (Russom, 2011).
Velocity	High frequency of data generation and data delivery, such as clickstream data (Russom, 2011, Manyika et al., 2011).
Value	The extent to which big data generates economically worthy insights and or benefits through extraction and transformation (Wamba et al., 2015).
Veracity	Data quality or truthfulness. The unreliability inherent in some sources of data is another consideration of big data (Elgendy et al., 2014; Gandomi et al., 2015).
Variability	The variation in the data flow rates that indicates the periodic peaks and troughs of big data velocity (Gandomi et al., 2015).

2.2 Data Analytics

"Big data does not automatically translate into better decision-making and performance" (Mariani et al. 2021). Similarly, Line et al. (2020) argue that data does not equal knowledge.

As a result, simply possessing big data will not assist a company in making better data-driven decisions or increase performance. Indeed, big data alone is insufficient to provide valuable insights; the analytics that accompany big data also has a substantial influence. Rather than basic data processing, the true core of big data applications is extracting the intelligence included in the data.

Mariani et al. (2021) broadly classify data analytics into four categories based on two dimensions (time and knowledge type): descriptive analytics, exploratory analytics, predictive analytics, and prescriptive analytics. To be more specific, the time dimension is divided according to past, present, and future, while the type of knowledge is used to distinguish whether the data produces information or knowledge. Moreover, Król et al. (2020) state that cognitive analytics is the next level of prescriptive analytics. Cognitive analytics involves the integration of cognitive computing techniques to extract meaningful insights from a variety of heterogeneous data sources through the implementation of cognitive models that mimic human cognitive processes (Gudivada et al., 2016). In addition, diagnostic analytics is a methodology that aims to identify the underlying root causes of a specific problem by employing a combination of exploratory data analysis and other tools such as visualization techniques to examine the available data, and potentially collect additional data as necessary (Banerjee et al., 2013). Table 2 shows the differences and examples in the hotel industry of the six types of analytics. In this thesis, the focus is on predictive analytics.

Table 2 Differences of the 6 Types of Analytics

Analytics	Aim	Examples in the hotel industry
Descriptive Analytics	Generate information about the past and present that can be used for the future (Mariani et al., 2021; Mathew, 2019).	Buyer summary analysis on data from a chain hotel to identify products bought from each buyer and the total amount paid (Mathew, 2019).
Exploratory Analytics	Generate knowledge about the past and present (Mariani et al., 2021).	Creating a plot to understand the relationship between *special requests* and *cancellation booking status* before forecasting the cancellation rate of online orders (Saputro et al., 2021).
Predictive Analytics	Generate information about the future by learning from real-time and historical data (Mariani et al., 2021; Mathew, 2019).	Forecasting of hotel booking cancellations using PNR as independent variables to identify future cancellation rates and which customer is likely to cancel (Sánchez-Medina et al., 2020).
Prescriptive Analytics	Generate knowledge about the future and propose an action plan (Mariani et al., 2021; Mathew, 2019).	Extracting short-comings related to hotel attributes, examining the quantitative effect of fixing these problems on hotels' online scores, and providing some prescriptions for hotel managers (Rezaei et al., 2022).
Diagnostic Analytics	Identify the underlying root causes of a specific problem by employing a combination of	Listing the number and proportion of hotel properties in different ser-

	exploratory data analysis and other tools (Banerjee et al., 2013).	vice classes and hotel properties under major brands on Yelp, TripAdvisor and Expedia to find the reason of the distribution (Xiang et al., 2017).
Cognitive Analytics	Extract meaningful insights from a variety of heterogeneous data sources through the implementation of cognitive models that mimic human cognitive processes (Gudivada et al., 2016)	Using machine learning to distinguish positive and negative deceptive and non-deceptive reviews and the main topics associated to positive and negative deceptive and non-deceptive reviews across 20 popular hotels in Chicago (Martinez-Torres et al., 2019).

Big data and data analytics have increasingly attracted scholars' attention and big data and data analytics together have been playing a leading role in a company's strategy operations. However, they are a double-edged sword, as there are still numerous challenges related to big data and data analytics. The first concept that needs to be considered is privacy; not only the personal data privacy, but also the privacy of people's state and behaviour, since people's state and behaviour can be targeted by analysing people's data, including personal data and data traces on the internet (Zhang, 2018). For example, sports apps can detect people's real-time data through some physical indicators, so as to obtain people's health status after data analysis. If these data are not protected sufficiently, people's privacy may be violated. In addition, data security is another consideration. Smart devices are commonly used nowadays, from smart phones to smart home terminals. Frequent engagement with data through multiple terminals, including the utilization of smartphones for the management of smart home terminals, has become a widespread practice. This may raise security issues; for example, once data of the smart phone is obtained or controlled by others, it will cause serious problems to the smart home terminals (Zhang, 2018). Moreover, according to my experience and observations, people are in general passive in the situation of being dominated by big data, and even worse, big data have changed people's lifestyle and daily habits. With the development of big data and analytics, all kinds of websites and applications are becoming more and more aware of people's own preferences, more so than people themselves. While they analyse people's preferences based on the data they obtain and keep feeding people the information they want to see, they can receive the feedback data. This results in a cycle that optimizes itself, i.e., collect data – analyse data – recommendation – collect feedback data – analyse feedback data – recommendation cycle (presented in Figure 1). For example, TikTok uses its own recommendation algorithms to recommend short videos for users, and again uses users' feedback data on these recommended videos, such as viewing time and likes, to optimize the algorithms and continues to recommend videos of interest to users in order to achieve the purpose of increasing the users' viewing time and generating transactions. This may lead to users spending more time and money on TikTok which, in turn, changes the user's lifestyle and daily habits in disguise.

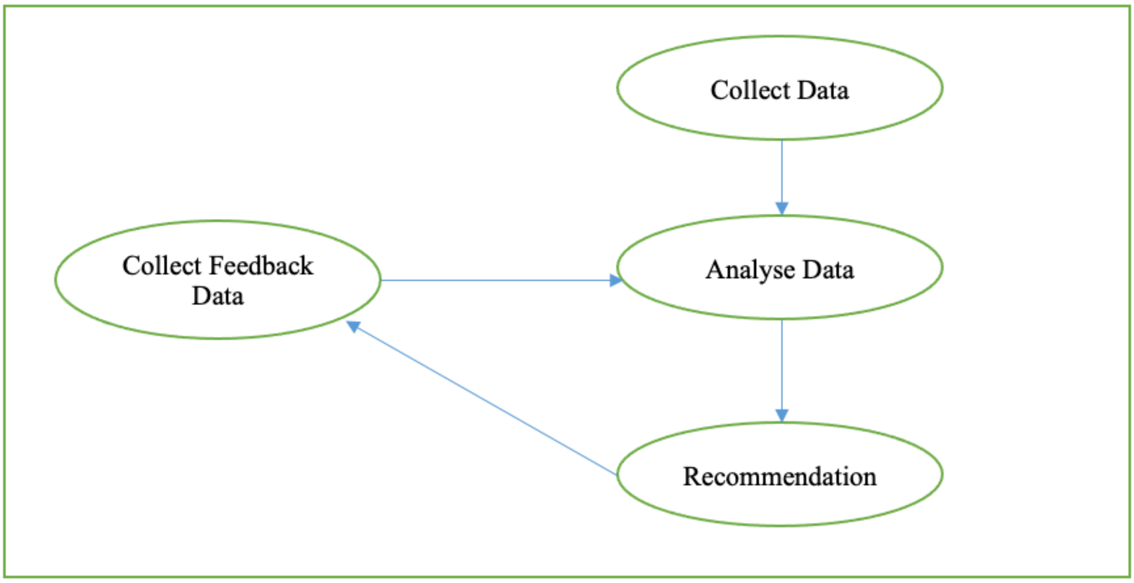


Figure 1 Data Optimization cycle

2.3 The Hotel Industry in the era of Internet and Big Data

As a traditional industry, the hotel industry concerned location and visitor flows most before the era of internet. Before the advent of the Internet, it was difficult for people to obtain information since there was a serious barrier of information asymmetry. Visitors did not even know that there was a certain hotel in a certain place, let alone how to make a reservation in advance or how to perform precision marketing from the hotels' side. The emergence of the Internet, especially the mobile Internet, has eliminated the distance between hotels and the visitors, breaking this information asymmetry. With the advancement of technology and the arrival of the era of big data, hotels have begun to notice a decrease in the number of walk-in customers visiting the hotels, and an increase in the number of customers who make reservations in advance via the Internet.

2.3.1 Hotels in the era of internet and Big Data

One hand, as a service provider in the supply side, hotels' business models will undergo profound changes. The Internet has revolutionized the way hotels do business, as it has made it much easier for travellers to search for, compare, and book accommodations online (He, 2019). Hotels have had to adapt by creating websites, using online travel agencies (OTAs) such as Tripadvisor and Booking.com, and establishing a presence on social media platforms in order to reach and attract customers. In the meantime, hotels are now able to collect and analyse data on customer preferences, booking patterns, customer reviews, and other factors to realize precision marketing and make informed business decisions (Zhang et al., 2019). For example, hotels may use data analytics to optimize room pricing, target marketing campaigns, and improve their online reputation.

2.3.2 Machine learning in the hotel industry

Big data and machine learning have a significant influence the development of the hotel industry from various perspectives. To be more specific, the utilization of machine learning within the hotel industry has been demonstrated to enhance the overall user experiences, customize/personalize services, and provide insights into consumer behaviours which is achieved through the prediction of purchasing behaviours and identification of future trends (Alotaibi, 2020). This thesis will focus on the forecast perspective in the context of the hotel industry. Therefore, improving user experiences and customizing services for users by using machine learning, such as smart hotels with intelligent robot and face recognition, will not be illustrated. According to Alotaibi (2020), a majority of studies by researchers in the hotel industry have focused on the utilization of machine learning techniques within the context of online reviews understanding, demand forecasting, hotel price prediction, booking cancellation prediction, and revenue management. In addition, more applications of machine learning from the perspective of forecast in the context of hotel industry are presented in Table 3.

Table 3 Forecast research using machine learning in the hotel industry

Researcher	Topic	Algorithm
Moon et al., 2015	Predicting the energy consumption in a hotel room	Artificial Neural Network (ANN)
Sánchez-Medina et al., 2020; Antonio, et al., 2017; Antonio, et al., 2019	Predicting the hotel cancellations	random forest, support vector machine (SVM), ANN, XGboost
Sun et al., 2019	Forecasting tourists’ arrivals	kernel extreme learning machine (KELM) models, LSSVR, SVR, ANN and ARIMAX.
Pan and Yang, 2017; Rivera, 2016	Forecasting hotel occupancy	ARMAX, MSDR
Pereira et al., 2022	Forecasting hotel demand	ADE
Al Shehhi et al., 2020	Forecasting hotel room price	the seasonal autoregressive integrated moving average (SARIMA) model, the restricted Boltzmann machine as a deep belief network model, the polynomial smooth support vector machine model, the adaptive network fuzzy interference system (ANFIS) model
Puh and Babac, 2022	Predicting sentiment and rating of tourist reviews	Naïve Bayes, SVM, convolutional neural network (CNN), long short-term memory (LSTM) and bidirectional long short-term memory (BiLSTM)

2.3.3 UGC in the era of internet and big data

On the other hand, from the perspective of customers, the use of the internet and big data has made it easier and more convenient for customers to book and experience their hotel stays. Meanwhile, they also generate ratings and reviews for their stays. As main sources of genuine information, the user-generated content (UGC) has endangered the hegemony of conventional content generators such as one-way advertising and expert critiques (Beer et al., 2010). Online word-of-mouth (WOM) has a greater influence on online customers' assessments of product and their decision making (Brown et al., 2007). In addition, Phillips et al. (2017) propose that the UGC are more successful in influencing consumer behaviour than traditional marketing tools.

To illustrate, the UGC in the hotel industry can be the ratings and textual reviews for the hotels by tourists after each of their stay of a hotel. Leal et al. (2019) state that the UGC is crucial for tourists to choose tourism resources, while according to the World Tourism Organization, the UGC on digital platforms is decisive for people to select places to visit, to stay or to eat (World Committee Tourism Ethics, 2017). In the present day, more and more consumers would refer to the ratings and reviews of the hotel in order to choose a satisfactory hotel. Booking.com surveyed 21,500 tourists around the world, and the results show that 75% of tourists consider other people's reviews to be very influential in helping them find a hotel before making a reservation. The survey also found that 53% of tourists will terminate the reservation of the hotels due to bad ratings and reviews about the hotels on the website (Madison, 2019).

positive reviews or ratings can increase customers' trust in a hotel and make them more likely to book with the hotel. However, negative ratings and reviews may have an even more significant impact on a hotel's online reputation and may discourage potential customers from booking a room. Jin et al. (2016) conclude that positive information in the UGC does not significantly affect consumer evaluation whereas negative information significantly affects consumer evaluation. This conclusion shows that users pay more attention to the negative UGC of other users and reduce the influence of positive UGC when evaluating the quality of hotels. This can make it especially crucial for hotels to address negative ratings and reviews in a timely and effective manner, or hotels could take steps in advance to avoid negative ratings and reviews if they could have a relatively accurate forecast about future reservations' ratings and reviews. In other words, hotels should reduce negative UGC and encourage satisfied customers to leave positive UGC in order to boost the hotel's reputation and attract more business.

2.4 The necessity of forecasting ratings for future reservations

Ye et al. (2009) conclude that hotel managers should passionately consider the UGC about their hotels, especially those posted on third-party websites, as the UGC online has a significant impact on hotels' future reservations because a 10% increase in reviewer ratings increases sales by 4.4%. Thus, a higher online reputation of a hotel positively influences the future reservations and the profitability of the hotel. Simultaneously, when choosing a hotel to stay, customers will not only refer to the positive or negative UGC, but also to the number of UGC. The higher the number of UGC, the more positive the UGC, the more likely customers will choose this hotel. More future reservations could bring more new UGC for a hotel. If a hotel can maintain or even further improve more positive UGC of future reservations, it will keep or improve a better reputation for the hotel, and this will result in more future reservations and profitability of the hotel. Figure 2 shows the forward cycle of a hotel's online reputation and future reservations (profitability).

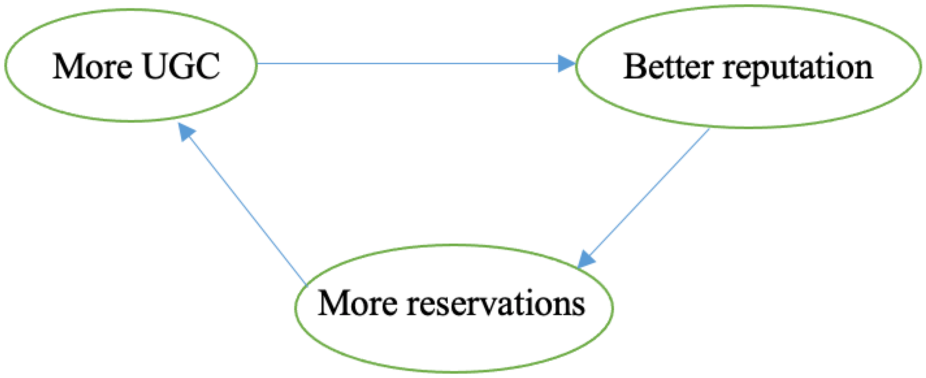


Figure 2 Forward cycle of a hotel's online reputation and future reservations

It is imperative for a hotel to improve a good online reputation for a hotel. Hotels should pay more attention to the UGC of the hotels on different online hotel booking platforms, since the UGC determines the online reputation of the hotels, which in turn determines the profitability and customer acquisition capabilities of the hotels. Therefore, it is advantageous that a hotel could have a relatively accurate forecast about a certain future reservation’s rating in advance. To be more specific, if the forecasted rating of a certain future reservation is relatively low, the hotel could perform some actions, such as upgrading the room to a higher floor, paying more attention to the room cleanness, or making corresponding improvements for the keywords of common negative textual reviews, in order to improve this future reservation’s rating and review to maintain or even improve the online reputation of the hotel.

In light of the current climate, characterized by the proliferation of the internet and the abundance of big data, it is essential for hotels to implement a future reservations' low rating early warning model in order to maintain a competitive edge and consistently deliver superior performance.

3 DATASET

This chapter explains the dataset of this research, including the chosen variables of the datasets that are used in this thesis. This chapter also explains the reasons for choosing booking.com as a data source and choosing Radisson Blu seaside hotel in Helsinki as the target hotel.

3.1 Data source and variables selection

With the development of the internet and big data, hotel distribution channels have also changed with times. The emergence of these channels, such as hotel official websites, OTAs' websites, B2B websites, social media platforms, various map applications, enables customers to obtain more information when choosing their preferred hotels. For example, customers can gain a comprehensive understanding of the hotel they want to choose by the relatively objective hotel reviews and ratings generated by other guests.

As mentioned above, the objective of this thesis is to forecast a future reservation's rating (UGC) of a hotel based on PNR. The variables which will be used as predictors in the machine learning models should be selected from the PNR, and the target variable is specified as the UGC which is the reservation's rating. On one hand, the rating as the target variable is public data, and it can be easily acquired in most of the emerging online distribution channels. In addition, after a general observation of different hotel official websites, OTAs' websites, and map applications, it can be concluded that the number of ratings and reviews on OTAs' websites are the largest among those channels. Therefore, in order to obtain a relatively larger dataset for building machine learning models with less bias, the OTAs' websites have been chosen to obtain the dataset. On the other hand, Sánchez-Medina et al. (2020) state that PNR that exist in historical booking records and is composed of the information provided by guests at the time a reservation is placed. On OTAs' websites, guests may provide different information through different websites when they make a reservation.

Among those emerging OTAs' websites, Hotels.com is a leading provider of hotel accommodation worldwide, and it also gives travellers one of the widest selections of accommodation on the net, including both independent and major chain hotels as well as self-catering in over hundreds of thousands of properties worldwide (Hotels.com, 2023). Hotels.com uses a rating system that allows guests to rate their stays based on different aspects of the hotel, including the quality of the room, the cleanliness of the property, and the service they received. Guests can also leave written reviews about their experience. Worth noting is that Hotels.com may verify the authenticity of reviews before they are published on the website, to ensure that they are legitimate and not spam or fake. This helps to maintain the integrity of the review system and ensure that the reviews are helpful for other users.

Booking.com as one of the largest OTAs' websites in the world, facilitates the reservation of accommodation, transportation, and other travel-related services. According to Booking.com (2023), Booking.com has more than 28 million reported accommodation listings, and it is available in 43 languages. Booking.com allows guests to leave ratings and textual reviews of their stays at properties booked through the website and uses a rating scale up to 10. The ratings that are showed on the websites will be a separate general rating in each individual review instead of ratings of different aspects. Additionally, Booking.com has a sizable reviews database with over 140 million verified evaluations left by visitors following their stay (Martin-Fuentes, 2018). When compared to other OTAs' websites, Booking.com has a distinct advantage in terms of the number of reviews, hosting 39% of all reviews globally (Murphy, 2017), and perhaps a greater percentage if we concentrate on the European market (Martin-Fuentes, 2018).

Tripadvisor.com was founded in 2000 and has since become one of the largest travel review websites in the world, with millions of reviews and ratings for places all over the world, including hotels, restaurants, and attractions. It is available in 43 markets and 22 languages as travel guidance company (tripadvisor.com, 2023). In addition to reviews and ratings, it also allows users to research and book travel experiences, including hotels. Hotel reviews on Tripadvisor are written by travellers who have stayed at a particular hotel and want to share their experiences with others.

The most popular OTAs' websites, including Hotels.com, Booking.com, Tripadvisor.com are reviewed in order to find out what mandatory PNR data guests must provide when they book a hotel, and what UGC data is contained in the review section. The results after summarized can be found in Table 4 and 5.

Table 4 PNR from different OTAs' websites

Websites	**Variables**	**Description**
Booking Tripadvisor Hotels	Name	The name of person who make the reservation. (Textual)
Booking Tripadvisor Hotels	Email	Email address. (Textual)
Booking Tripadvisor Hotels	Phone	Phone number. (Textual)
Booking Tripadvisor Hotels	Adult_Number	The number of adults in the reservation. (Numeric)
Booking Tripadvisor Hotels	Child_Number	The number of children in the reservation. (Numeric)
Booking Tripadvisor Hotels	Check_In	Check in date (Date)
Booking Tripadvisor Hotels	Check_Out	Check out date (Date)
Booking Tripadvisor Hotels	Room_Type	The room type of the reservation. (Categorial and vary according to different hotels)
Booking Tripadvisor Hotels	Room_Number	The number of rooms of the reservation. (Numeric)
Booking	Address	The Address of the person who make the reservation. (Textual)
Booking	City	The city of the person who make the reservation. (Textual)
Booking Tripadvisor	Country	The country of the person who make the reservation. (Categorial)
Booking Hotels	Payment_Method	The method of the payment (Categorical: such as Card, PayPal, Google Pay. May vary according to different country where the person who make the reservation locate in)

Booking Tripadvisor Hotels	Payment_Details	The details of the payment, such as the card information. (Textual)

Table 5 Data in review section of different OTAs' websites

Websites	**Variables**	**Description**
Booking Tripadvisor Hotels	Name	The name of the guest. (Textual)
Booking Tripadvisor Hotels	Country	The country of the guest. (Categorial)
Tripadvisor	City	The country of the guest. (Categorial)
Booking	Room_Type	The room type of the stay. (Categorial and vary according to different hotels)
Booking Hotels	Nights	The number of nights of the stay. (Numeric)
Booking Tripadvisor	Check_In_Month	The month in which the guest check in the hotel. (Categorical)
Booking TripAdvisor	Check_In_Year	The month in which the guest check in the hotel. (Numeric)
Booking TripAdvisor Hotels	Travel_Type	The type of this travel for the guest. (Categorical)
Tripadvisor	Contribution	The number of reviews the guest has written. (Numeric)
Tripadvisor	Helpful_Votes	The number of reviews written by the guest that other people considered useful. (Numeric)
Booking Tripadvisor Hotels	Rating	The guest rating for this stay. (Numeric)
Booking Tripadvisor Hotels	Review	The guest textual review guest for this stay. (Textual)
Booking Tripadvisor Hotels	Reviewed_Date	The date of this review. (Date)

The mandatory data that the guests must provide when they book a hotel, has subtle differences according to Table 4. However, most of the mandatory data what the guests must provide when they book a hotel is mostly the same.

The PNR data that are required when booking a hotel appears on all three OTAs' websites is:

- Name
- Email
- Phone
- Adult_Number
- Child_Number
- Check_In
- Check_Out
- Room_type
- Room_Number
- Payment_Details

In order to accurately predict ratings in this research, it is important to consider a range of variables as predictors. Additionally, there exist certain data points that were not included in the aforementioned list. These may include optional information that is mandated during the reservation process, as well as mandatory data that is not displayed on all three OTAs' websites. It can be seen as missing data in the final dataset, since the guests may not provide these two types of data, and the hotel may not receive the data when a reservation is placed. The inclusion of these types of data in the analysis is not deemed necessary for the purposes of this research, as they are not expected to contribute significantly to the predictive capabilities of the model. Therefore, when building the rating-prediction model in this research, the variables as predictors should be the PNR data that are required to appear on all three OTAs' websites above. However, the data of PNR as predictors are especially difficult to be obtained directly from either hotels or these online distribution channels, as it is extremely sensitive information regarding privacy and economic. In this research, another method of indirectly obtaining the replacement of these PNR data from public data in the review section on OTAs' websites is used.

The data which are contained in the review section on different OTAs' websites are shown in Table 5. Compared with the PNR data required as predictors above, the data from review section of Tripadvisor and Hotels is deficient in a number of crucial data separately, such as *Room_Type, Check_In_Year, Check_In_Month, Nights.* The data from the review section of Booking is most suitable for the replacement of the PNR data required as predictors. Among the data, *Rating* is the target variable, while *Review* and *Reviewed_Date* cannot be gathered by the hotel when a reservation is placed. Furthermore, Gupta. (2019) states that the real-world data contains irrelevant or meaningless data termed as noise which can significantly affect data analysis tasks of classification. As for *Check_In_Year*, it is a meaningless variable as a predictor in the classification models, since the test data is always later than the train data in real world, which means that the value of *Check_In_Year* is always increasing and maybe a noise of the classification models.

Therefore, the data from the review section of Booking which can be used as the replacement of the PNR data required as predictors is the following:

- Name
- Country
- Room_Type
- Nights
- Check_In_Month
- Travel_Type

Compared with the PNR data that are required when booking a hotel above, *Email* and *Phone* are confidential data only collected by the hotel for their contact purpose. *Payment_Details* is also confidential data that cannot be gathered. These three kinds of data cannot be gathered and the dataset for building the model will exclude these kinds of variables. *Adult_Number*, *Child_Number* and *Room_Number* can be roughly replaced by *Travel_Type*, because the classification of *Travel_Type* in the review section of Booking.com are Solo, Couple, Family and Group. In addition, *Check_In* and *Check_Out* will be roughly replaced by *Check_In_Month* and *Nights*.

In summary, three most popular OTAs' websites are reviewed to identify the possible PNR-data variables that can be used as the predictor variables in the model to be constructed. Furthermore, by combining the public data in the review sections that can be obtained as replacements of PNR-data variables, the variables as predictors are determined. In order to obtain more PNR-data variables as the predictor variables of the model from the public data in the review section of OTAs' websites, Booking.com was chosen as the data source. Furthermore, Booking.com is trustworthy and possesses a notable advantage in terms of the quantity of reviews available, which will yield a more extensive dataset for subsequent training and testing purposes.

3.2 Hotel selection

The objective of this research is to develop predictive models using machine learning techniques to forecast the future ratings of hotel reservations. The aim is to maintain or improve the online reputation of the hotel. However, it is essential to note that the variables and parameters that impact ratings may vary depending on the hotels' unique characteristics, including their location, target market, amenities, and services. Consequently, hotel-specific predictive models must be developed to capture the specific variables and parameters that influence ratings at the hotel under investigation. Moreover, hotel-specific predictive models will avoid the shortcomings associated with general predictive models. The variables and parameters that significantly impact ratings at one hotel may not have the same influence on ratings at another hotel. For this purpose, the Radisson Blu Seaside Hotel in Helsinki is selected as the focal point of this research.

According to Martin-Fuentes (2018), Booking.com is in a dominating position in Europe. The Radisson Blu brand is well-known for its upscale amenities, exceptional service, and prominent chain presence in Europe, which makes it an ideal candidate for this research. Alongside a vibrant harbour, Radisson Blu Seaside Hotel in Helsinki, is a tribute to the modern Finnish way of life, and offers 348 contemporary guest rooms, along with a comprehensive range of modern amenities (Radissonhotels.com, 2023). Its prime location near popular tourist attractions and cultural landmarks makes it a sought-after destination for both business and leisure travellers alike. The hotel's commitment to providing an exceptional guest experience is reflected in its high ratings and positive reviews on Booking.com, which has amassed over 5800 reviews, with an average rating of 8.4 out of 10.

The reputation of hotel Radisson Blu Seaside in Helsinki and its chain brand in Finland is another factor that makes it an ideal candidate for this research. The chain's reputation for providing upscale amenities, exceptional service, and prime locations makes it a popular choice among travellers. In addition, the Radisson Blu chain has a significant presence in Finland through its partnership with S Group (S-ryhma.fi, 2023), with 7 other Radisson Blu hotels located throughout the country (Radissonhotels.com, 2023).

To develop the predictive models, machine learning algorithms that require large datasets will be employed to generate accurate results. The vast number of reviews available of hotel Radisson Blu Seaside in Helsinki on Booking.com will facilitate the collection of sufficient data. The data will be subjected to further filtering and refinement to include solely the data that are pertinent to the research questions.

In conclusion, the Radisson Blu Seaside Hotel in Helsinki is an ideal candidate for this research, given its upscale amenities, exceptional service, prime location, and prominent chain presence in Finland. Its large number of reviews on Booking.com, along with its reputation as a leading hotel brand in Europe, make it a compelling choice for this research. Focusing on the Radisson Blu Seaside Hotel in Helsinki allows for the development of customized models that capture the distinct variables and parameters unique to the hotel. This customized approach will enable hotel operators to gain valuable insights into the maintenance or improvement of the hotel's online reputation, ultimately leading to an increase in revenue.

4 METHODOLOGY

The methodologies used for this research are:

- To obtain all the required data by web scraping from Booking.com.
- To perform data pre-processing to transfer the original data acquired from Booking.com to a format which is machine-readable and will contribute to a higher accuracy of the machine learning model in the future,
- To build different machine learning models to train and test the data.
- To compare the result of the different machine learning models.
- To perform textual data analysis.

Python's improved library support combined with Python's strength in general purpose programming has made Python an excellent choice as a single language for building data-centric applications (McKinney, 2012). Therefore, all the methodologies used in this research are Python-based and will be illustrated in the following part of this chapter. The version of Python used in this research is Python 3.7.6.

4.1 Web scraping

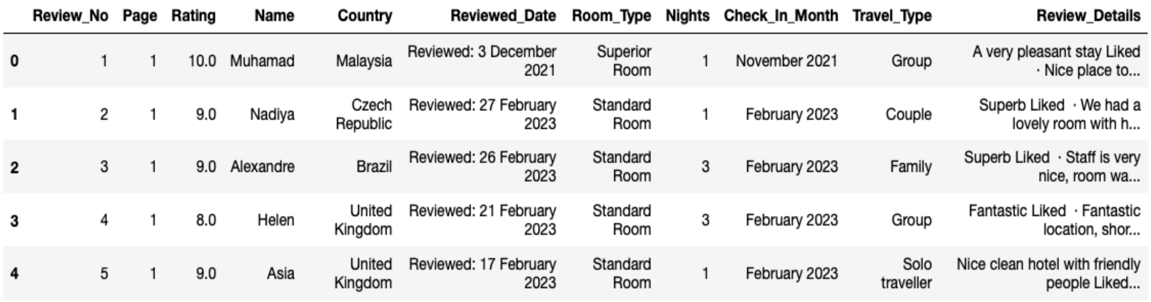
Web scraping is used to obtain all the required data from Booking.com and stored in a machine-friendly way which is a comma-separated values (CSV) file in this research. The Beautiful Soup library in Python is used in this research to achieve this goal. Beautiful Soup is a Python data extraction library developed by Leonard Richardson and other open-source developers, and it works on both Python 2.7+ and Python 3 (Uzun, 2018). The version of Beautiful Soup used in this research to extract data from Booking.com is 4.8.2. Beautiful Soup can extract data from HTML and XML files, and it supports four parsers which are html, lxml's HTML, lxml's XML, and html5lib (Beautiful Soup Documentation, 2023). The parser used in this research to parse the web page of Booking.com is lxml's HTML. According to Beautiful Soup Documentation (2023), this parser is very fast and lenient.

The data variables scraped from the review section of Booking.com are shown in Table 6. Certain variables will still be scraped, despite not being incorporated into the machine learning models. *Review_No* is used to indicate each review. *Page* is used to locate the webpage of the review in case the review will be needed to be found on the website.

Reviewed_Date is also scraped in case there is further use. An empty value is filled in if there is an error raised in the process of scraping a specific data. The white spaces such as the *"\n"* are also be removed during the process of web scraping. The scraped data are stored in a .csv file by using the csv module in python standard library. The first five rows of the original dataset are shown in Figure 3.

Table 6 Data variables scraped from Booking.com

Variables	**Description**
Review_No	The number of the review. (Numeric)
Page	The number of pages the review is on. (Numeric)
Name	The name of the guest. (Textual)
Country	The country of the guest. (Categorial)
Reviewed_Date	The date of the review. (Date)
Room_Type	The room type of the stay. (Categorial)
Nights	The number of nights of the stay. (Numeric)
Check_In_Month	The month in which the guest check in the hotel. (Categorical)
Travel_Type	The type of this travel for the guest. (Categorical)
Rating	The guest rating for this stay. (Numeric)
Review_Details	The guest textual review guest for this stay. (Textual)



	Review_No	Page	Rating	Name	Country	Reviewed_Date	Room_Type	Nights	Check_In_Month	Travel_Type	Review_Details
0	1	1	10.0	Muhamad	Malaysia	Reviewed: 3 December 2021	Superior Room	1	November 2021	Group	A very pleasant stay Liked · Nice place to...
1	2	1	9.0	Nadiya	Czech Republic	Reviewed: 27 February 2023	Standard Room	1	February 2023	Couple	Superb Liked · We had a lovely room with h...
2	3	1	9.0	Alexandre	Brazil	Reviewed: 26 February 2023	Standard Room	3	February 2023	Family	Superb Liked · Staff is very nice, room wa...
3	4	1	8.0	Helen	United Kingdom	Reviewed: 21 February 2023	Standard Room	3	February 2023	Group	Fantastic Liked · Fantastic location, shor...
4	5	1	9.0	Asia	United Kingdom	Reviewed: 17 February 2023	Standard Room	1	February 2023	Solo traveller	Nice clean hotel with friendly people Liked...

Figure 3 The original dataset scraped from Booking.com

4.2 Exploratory data analysis (EDA) and data pre-processing for predicting ratings' categories

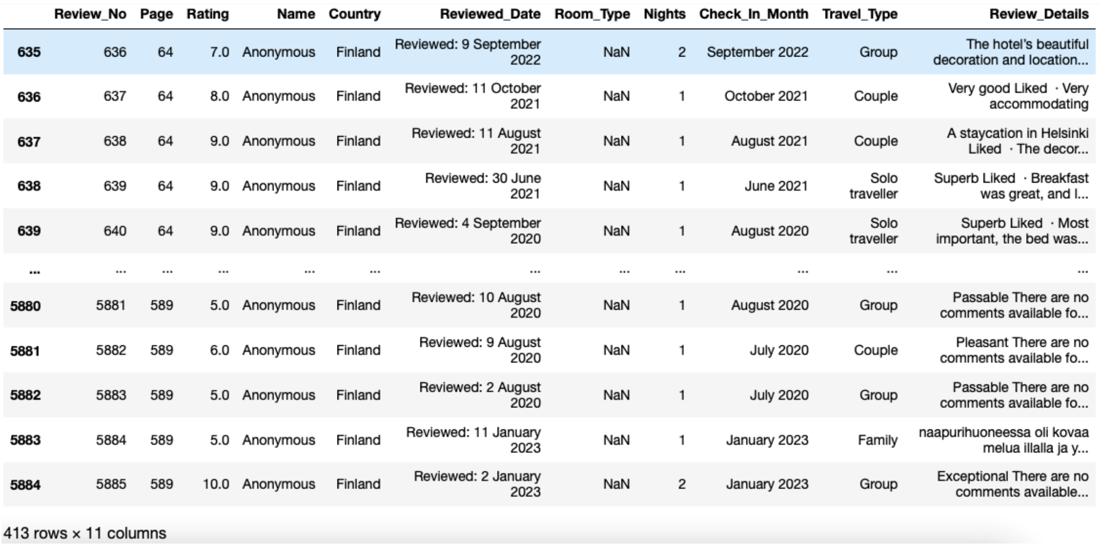
In order to build the machine learning models and achieve better performances of the models, exploratory data analysis and data pre-processing need to be performed. A clear understanding of data distribution and relationships among different variables in the dataset is gained through exploratory data analysis. Based on the understanding of the data and variables, combined with the objective of this research, data pre-processing is performed to transfer the data in the original dataset obtained from Booking.com by web scraping to a format that is machine-readable and contributes to a better performance in the machine learning models. The pandas library, under development since 2008, is intended to provide rich data analysis tools, and meanwhile, it provides integrated, intuitive, and convenient methods for performing common data manipulations (McKinney, 2011). Therefore, the exploratory data analysis and the data pre-processing in this research is based on the pandas library.

The processes of data pre-processing in this research include as follows:

4.2.1 Address missing values.

Real-world datasets often suffer from missing data, which can hinder the analysis and interpretation of results. Various techniques have been developed to handle missing values in a principled and effective manner. In particular, six approaches have been proposed and widely adopted, which can be divided into two categories: remove missing values and fill in the missing values (Han et al., 2022).

The resulting dataset obtained by web scraping from booking.com comprises 5,885 observations. Missing data is present in two variables, namely *Country* and *Room_Type*. Among these, 9 observations are missing *Country* data, and 413 observations are missing *Room_Type* data, representing a relatively small proportion of the dataset. Upon closer examination, it was discovered that all the missing values for *Room_Type* are associated with column *Name*, resulting in null values for this variable. Specifically, as shown in Figure 4, the entries for which the name is 'Anonymous' have a missing *Room_Typ*e (i.e., denoted by 'NaN'). Therefore, in this research, missing values will be addressed using a data cleaning approach that all observations with missing values will be removed from the analysis. 5464 rows are left after removal of all rows that contain missing values. This method is a common practice in data pre-processing when the percentage of missing values is low. However, this approach may result in reduced sample size and potential loss of information, which should be carefully considered in the interpretation of results.



	Review_No	Page	Rating	Name	Country	Reviewed_Date	Room_Type	Nights	Check_In_Month	Travel_Type	Review_Details
635	636	64	7.0	Anonymous	Finland	Reviewed: 9 September 2022	NaN	2	September 2022	Group	The hotel's beautiful decoration and location...
636	637	64	8.0	Anonymous	Finland	Reviewed: 11 October 2021	NaN	1	October 2021	Couple	Very good Liked · Very accommodating
637	638	64	9.0	Anonymous	Finland	Reviewed: 11 August 2021	NaN	1	August 2021	Couple	A staycation in Helsinki Liked · The decor...
638	639	64	9.0	Anonymous	Finland	Reviewed: 30 June 2021	NaN	1	June 2021	Solo traveller	Superb Liked · Breakfast was great, and I...
639	640	64	9.0	Anonymous	Finland	Reviewed: 4 September 2020	NaN	1	August 2020	Solo traveller	Superb Liked · Most important, the bed was...
...	...	...	...	...	...	...	...	...	...	...	...
5880	5881	589	5.0	Anonymous	Finland	Reviewed: 10 August 2020	NaN	1	August 2020	Group	Passable There are no comments available fo...
5881	5882	589	6.0	Anonymous	Finland	Reviewed: 9 August 2020	NaN	1	July 2020	Couple	Pleasant There are no comments available fo...
5882	5883	589	5.0	Anonymous	Finland	Reviewed: 2 August 2020	NaN	1	July 2020	Group	Passable There are no comments available fo...
5883	5884	589	5.0	Anonymous	Finland	Reviewed: 11 January 2023	NaN	1	January 2023	Family	naapurihuoneessa oli kovaa melua illalla ja y...
5884	5885	589	10.0	Anonymous	Finland	Reviewed: 2 January 2023	NaN	2	January 2023	Group	Exceptional There are no comments available...

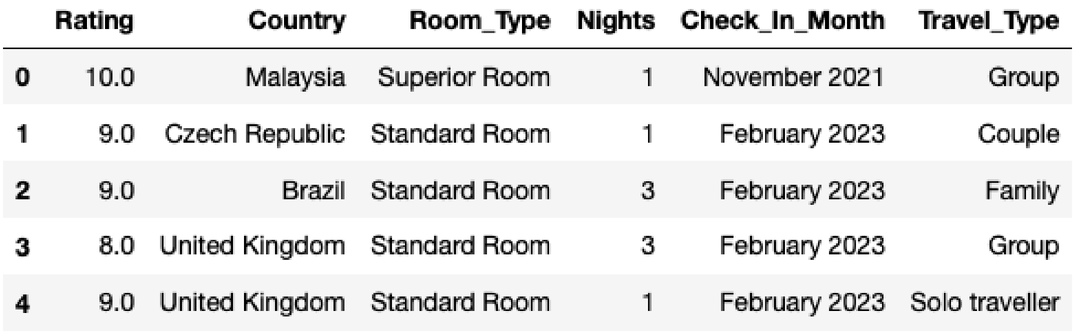
413 rows × 11 columns

Figure 4 Missing values of Room_Type

4.2.2 Feature selection.

In the context of data analysis, a variety of input variables may be collected from the same source or different sources, some of which may be irrelevant to the outcome of interest. While these variables may still have potential value for other purposes, their presence can significantly hinder the performance of machine learning algorithms by distracting or misleading them with irrelevant information. As a result, developing strategies for identifying and removing such variables in order to improve the accuracy and efficiency of predictive models is crucial. Feature selection aims to identify the most relevant input variables for predicting the target variable, which is crucial to prevent distractions or misleading signals from irrelevant variables and develop models with the simplest and most effective performance (Brownlee, 2020).

As noted previously, certain variables that do not contribute to the target variable but may have alternative uses are included in the dataset (i.e., *Review_No, Page, Name, Review_Date, Review_Details*). To mitigate the impact of irrelevant variables for the predict models, this research conducts an initial removal of such columns from the dataset. The first five rows of the dataset presented in Figure 5 has undergone a pre-processing step where irrelevant columns are removed, resulting in a reduced set of six columns. These remaining columns are *Rating, Country, Room_Type, Nights, Check_in_Month, and Travel_Type.* Subsequently, a feature selection process is conducted to further refine the dataset and identify the most pertinent input features that are predictive of the target variable. This process will be discussed in the results part, and it entails assessing the contribution of each input-variable subset to the model performance and selecting those that provide the highest predictive power and identify the most parsimonious, effective, and accurate model.



	Rating	Country	Room_Type	Nights	Check_In_Month	Travel_Type
0	10.0	Malaysia	Superior Room	1	November 2021	Group
1	9.0	Czech Republic	Standard Room	1	February 2023	Couple
2	9.0	Brazil	Standard Room	3	February 2023	Family
3	8.0	United Kingdom	Standard Room	3	February 2023	Group
4	9.0	United Kingdom	Standard Room	1	February 2023	Solo traveller

Figure 5 The dataset after removal of irrelevant columns

4.2.3 Categorize numerical values.

The categorize-numerical-values method used in a binary classification problem is a technique used to transform a numerical variable into a categorical variable with two or more categories. This method involves selecting a threshold value that separates the numerical values into two or more groups. For example, in the case of transforming a numerical variable into a categorical variable with two categories, any value below the threshold is assigned to one category, while any value at or above the threshold is assigned to the other category.

To apply this method, the first step is to select an appropriate threshold value. This can be done through various methods, such as visual inspection of a histogram or density plot of the numerical variable, or through statistical techniques such as clustering or decision tree analysis.

Once the threshold value has been selected, the numerical values are then categorized into two groups based on their relationship to the threshold value. For example, if the threshold value is set at 50, all values below 50 may be categorized as “low” or 0 and all values at or above 50 may be categorized as “high” or 1.

This method is useful when analysing data where the distinction between two categories is more relevant than the precise numerical value of the variable. It allows for easier interpretation of the data and can be particularly helpful in situations where the relationship between the numerical variable and the outcome variable is nonlinear.

This research employed the categorize-numerical-values method to transform the numerical values in the *Rating* column into categorical values with two categories, which were subsequently utilized as the target variable in the predict models. A threshold value of 8.4, which corresponds to the general rating of Radisson Blu Seaside Hotel in Helsinki on Booking.com, was selected to categorize the values into two groups labeled as 0 and 1. For reservations classified as category 1, the hotel need not be overly concerned since they indicate a rating higher than the general rating of the hotel and, therefore, have no adverse effect on the hotel's overall rating. Conversely, reservations belonging to category 0 may require greater attention, as they could potentially lower the general rating of the hotel and damage its reputation.

4.2.4 Numerize categorical values.

The process of numerizing categorical variables is a crucial technique in data pre-processing for machine learning applications. Categorical variables are variables that take on a limited number of discrete values, such as *Country*, *Room_Type*, *Travel_Type* or *Check_In_Month*. These variables pose a challenge for machine learning algorithms, which typically require numerical inputs for training and prediction purposes. In order to transform categorical variables into a suitable numerical representation, one common method is to use the “get_dummies” function, which is a widely used feature engineering technique in Python.

The “get_dummies” function is a way to convert a categorical variable into several binary columns with 0 and 1 values. This function creates a new column for each unique value in the original categorical variable and assigns a value of 1 or 0 to indicate the presence or absence of that value in the original column. For example, if we have a categorical variable *Travel_Type* with possible values “Group”, “Couple”, “Family” and "Solo traveller", the “get_dummies” function with the parameter “drop_first=True” will create three new columns *Travel_Type_Couple*, *Travel_Type_Family*, and *Travel_Type_ Solo_traveller*. If a data point had the value “Couple” for the *Travel_Type* variable, it would have a 1 in the *Travel_Type_Couple* column and 0s in the other two columns, and if a data point had the value “Group” for the *Travel_Type* variable, it would have all 0s in the three new-generated columns.

However, when dealing with categorical variables that have a large number of unique values, utilizing the “get_dummies” function to represent each value as a separate column can lead to an increase in the number of dimensions in the input space. This phenomenon is commonly referred to as the curse of dimensionality, where the performance of machine learning models deteriorates as the number of dimensions increases. Therefore, several steps are taken in this research to reduce the number of dimensions or the unique values of categorical variables to maintain optimal model performance. By reducing the dimensionality or the unique values of categorical variables, the effects of the curse of dimensionality can be mitigated and the performance of machine learning models can be improved. These will be further discussed and illustrated in the results section of this research.

4.3 Scikit-learn and algorithms for predicting ratings’ categories.

Scikit-learn (Sklearn) is a powerful machine learning library in Python that is widely used for developing and implementing various predictive models. Scikit-learn is an open-source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data pre-processing, model selection, model evaluation, and many other utilities (Scikit-learn, 2023). Scikit-learn is built on top of other popular scientific libraries in Python, such as NumPy, SciPy, and matplotlib, which provide support for numerical operations, scientific computing, and data visualization. Its user-friendly interface and well-documented API make it a popular choice among researchers and practitioners for developing and testing machine learning models.

Scikit-learn provides an efficient and easy-to-use platform for data analysis and offers a wide range of supervised and unsupervised learning algorithms. In this thesis, Scikit-learn is used extensively for building, evaluating, and optimizing different predictive models based on a variety of algorithms, including logistic regression, random forest, XGBoost and Artificial Neural Networks (ANNs).

4.3.1 Logistic regression.

Logistic regression is a statistical method that aims to model the relationship between one or more independent variables and a binary or categorical dependent variable (Wang et al., 2019). Logistic regression has emerged as a widely adopted statistical method for analysing binary response data, and has become a key tool in the arsenal of statisticians and researchers (Hilbe, 2009). It is a powerful tool for analysing and predicting the probability of a particular outcome based on a set of input variables. Logistic regression is particularly efficient and powerful when the dependent variable is binary, i.e., has only two possible outcomes, such as success or failure, yes or no, or true or false (Stoltzfu, 2011). The logistic regression model estimates the probability of the binary outcome as a function of the input independent variables, and it can be used to predict the probability of a particular outcome for new observations. In this research, logistic regression is used to analyse and model the relationship between a binary dependent variable (*Rating*) and a set of independent variables.

4.3.2 Random forest.

Random forest is a powerful and widely used machine learning algorithm that belongs to the family of ensemble learning methods. Breiman. (2001) first introduces the random forest algorithm, which constructs a multitude of decision trees at training time and combines them through a voting mechanism to determine the final prediction. Random forests change how the classification trees are constructed (Liaw et al., 2002). Each decision tree is constructed using a random subset of the available training data and a random subset of the available features, which helps to reduce overfitting and increase the generalization performance of the model. Random forests are a robust method for handling high-dimensional predictor variables in the context of complex interactions, and empirical studies have demonstrated the high prediction accuracy of random forests in such applications (Strobl et al., 2009). In this research, random forest is used to analyse and model the relationship between a binary dependent variable (*Rating*) and all possible independent variables.

4.3.3 XGBoost.

XGBoost is a popular and powerful machine learning algorithm that has gained widespread attention in recent years for its ability to achieve state-of-the-art results in a variety of tasks, including classification, regression, and ranking. XGBoost was first introduced by Chen and Guestrin (2016) as an optimized implementation of gradient boosting, which is a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning (Chen et al., 2016). XGBoost uses a tree-based model and employs a variety of optimizations to improve the accuracy, scalability, and efficiency of the algorithm. As an illustration, XGBoost applies a regularization technique to facilitate the smoothing of the final learnt weights and ultimately prevent overfitting, thus improving the generalization performance of the model (Chen et al., 2016). XGBoost has been shown to outperform other popular machine learning algorithms, such as random forest, in many applications (Chen et al., 2016; Zamani et al., 2019). In this research, XGBoost is used to analyse and model the relationship between a binary dependent variable (*Rating*) and a set of independent variables.

4.3.4 Artificial Neural Networks (ANNs).

Researchers from diverse scientific disciplines are developing artificial neural networks (ANNs) to tackle problems in pattern recognition, prediction, and optimization, with some of these approaches inspired by biological neural networks (Jain et al., 1996). The concept of ANNs was first introduced by Warren McCulloch and Walter Pitts in 1943 (McCulloch et al., 1943). The resurgence of ANNs, which occurred in the early 1980s, was attributed to significant developments such as Hopfield's energy approach in 1982 and the back-propagation learning algorithm for multilayer perceptrons (multilayer feedforward networks) introduced by Werbos and popularized by Rumelhart et al. in 1986 (Jain et al., 1996). ANNs are inspired by the structure and function of the human brain, which consists of interconnected neurons that process and transmit information. ANNs employ a layered structure of artificial neurons, called perceptrons, to capture complex non-linear relationships between inputs and outputs. Each neuron processes its input signals and transmits the result to the next layer of neurons until the final output is produced. According to Wu et al. (2017), ANNs is the method most frequently employed in contemporary literature concerning the hotel industry. ANNs have demonstrated remarkable performance in various applications and have been used in several studies for modelling and prediction in the hotel industry (Sánchez-Medina et al., 2020; Azarmi et al., 2018; Burger et al., 2001). In this research, ANNs is used to analyse and model the relationship between a binary dependent variable (*Rating*) and a set of independent variables.

4.4 Performance evaluation for predicting ratings' categories.

Performance evaluation of machine learning models is a crucial aspect of any data analysis project. As highlighted by Jain et al. (2000), it is essential to measure the performance of a machine learning model accurately to assess its effectiveness in a given task. Confusion matrix is a commonly used technique to evaluate the performance of a binary classification machine learning models (Sokolova et al., 2009). The evaluation of the accuracy of a classification model can be conducted by calculating the number of correctly identified instances of the target class (true positives), accurately classified non-target class instances (true negatives), incorrectly classified non-target class instances (false positives), and non-identified target class instances (false negatives) (Sokolova et al., 2009). These four quantities together comprise a confusion matrix, as illustrated in Table 7 for the scenario of binary classification in Python. The confusion matrix provides a comprehensive summary of the performance of a classification algorithm by summarizing the predicted class and actual class in a table format (Goutte, 2005).

Table 7 Confusion matrix of binary classification in Python

Actual data class	Predicted as 0	Predicted as 1
0	True negative (TN)	False positive (FP)
1	False negative (FN)	True positive (TP)

Several performance metrics such as precision, recall, and F1-score can be calculated from the confusion matrix. To illustrate, precision is defined as the quotient of the number of true positives by the total number of instances that the classifier predicted as positive, and it reflects the proportion of correctly predicted positive instances among all predicted positive instances. Recall is another performance metric used in machine learning and statistics to measure the effectiveness of a classifier. It is the ratio of true positives to the total number of instances that are actually positive. In other words, it measures the proportion of actual positive instances that are correctly predicted as positive by the classifier. F1 score is a measure of the overall performance of a classifier that takes into account both precision and recall. It is the harmonic mean of precision and recall, and provides a balance between the two metrics. The equation of precision, recall and F1 is shown as bellows:

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 = 2 * (Precision * Recall) / (Precision + Recall)

In this research, class 0 represents the ratings of future reservations is below the general rating of the hotel, and class 1 represents the ratings of future reservations is equal to and above the general rating of the hotel.

Another thing needs to mention is that a classification report provides two values for precision, recall, and F1 score. This information is typically presented in a tabular format, as exemplified by Table 8. The corresponding confusion matrix is shown in Table 9. The reason why there is two values for precision, recall, and F1 score can be understanded as each class represents a different category that the model is trying to predict, and can be illustrated with the example of recall as follows:

Recall for class 0: TP0 / (TP0 + FN0)

Recall for class 1: TP1 / (TP1 + FN1)

From the confusion matrix, the following information can be gained:

TP0 = 144 (the number of true positives for class 0)

FN0 = 194 (the number of false negatives for class 0)

TP1 = 340 (the number of true positives for class 1)

FN1 = 142 (the number of false negatives for class 1)

Therefore:

Recall for class 0 = TP0 / (TP0 + FN0) = 144 / (144 + 194) = 0.43

Recall for class 1 = TP1 / (TP1 + FN1) = 340 / (340 + 142) = 0.71

Table 8 An example of classification report

Class	Precision	Recall	F1-score	Support
0	0.50	0.43	0.46	338
1	0.64	0.71	0.67	482
Accuracy			0.59	820
Macro avg	0.57	0.57	0.57	820
Weighted avg	0.58	0.59	0.58	820

Table 9 Corresponding confusion matrix

Confusion matrix:			
TN:	144	FP:	194
FN:	142	TP:	340

This research places significant emphasis on the accuracy of class 0 prediction, as well as the ratio of true positive class 0 instances to all predicted class 0 instances. This holds particular importance for the hotel, which aims to focus on correctly identifying class 0 instances, while also minimizing the false identification of class 1 instances as class 0. The objective is to take appropriate measures to improve the hotel's overall rating by addressing class 0 instances, while also optimizing costs by maximizing the proportion of actual class 0 and predicted class 0 instances. This is because an increase in the number of identified class 0 instances leads to a higher cost associated with addressing them. In other words, this research will focus on the more the recall for class 0 the better, and simultaneously focus on the more the ratio of TN and FN the better. Therefore, a new measurement of performance (TN_score) should be generated as follows for this research.

Recall_0 = TN / (TN+FP)

TN_FN_Ratio = TN / FN

TN_score = Recall_0* TN_FN_Ratio

4.5 Data analysis for gaining insights from text reviews.

To extract insights for hotel managers from customers' (text) evaluations, a textual data analysis is conducted on the *Review_Details* column of the dataset scraped from Booking.com. The original dataset contains 5885 customer reviews. After removing missing values in other columns to maintain consistency with the previous analysis of the rating category prediction, the dataset utilizes for textual data analysis contained 5464 textual reviews.

Pre-processing of the text reviews is carried out using tokenization, stop word removal, and lemmatization. The CountVectorizer technique is then employed to conduct word frequency analysis. Textual data analysis is a multifaceted field that encompasses diverse techniques and methods for extracting insights from textual data. A commonly used technique in textual data analysis is natural language processing (NLP), which employs algorithms and computational methods to analyse and comprehend human language.

The present textual analysis is executed using Python and various libraries, including nltk, langdetect, googletrans, and sklearn. By employing these resources, this research aimes to uncover insights relevant to hotel managers from customers’ textual evaluations.

Textual data analysis is a process that involves utilizing various techniques to extract insights from unstructured textual data. Comprehending the meaning of textual data requires a comprehensive understanding of fundamental concepts and the use of specialized techniques for pre-processing textual data. The concepts and techniques commonly used in textual data analysis include tokenization, lemmatization, stop word removal, and CountVectorizer.

4.5.1 Tokenization.

The first step in textual data analysis involves identifying tokens, which are basic units that do not require further decomposition during subsequent processing (Webster et al., 1992). Tokenization is the process of breaking down a sentence or paragraph into individual words or tokens. By implementing tokenization, analysts can analyse individual words rather than treating a sentence as a whole. Following tokenization, textual data can be analysed for various features, including sentiment, word frequency, and other relevant characteristics.

4.5.2 Lemmatization.

Lemmatization is a textual data analysis technique that involves identifying the normalized or base form of a word by applying a transformation to it (Plisson et al., 2004). Lemmatization is a critical process in textual data analysis that involves reducing words to their base or root form, typically through the application of linguistic rules and algorithms. This technique facilitates the simplification of complex textual data by transforming inflected or variant word forms into their canonical or base form. For example, the words "running," "ran," and "run" can be lemmatized into their base form "run," thereby streamlining the data and mitigating the impact of sparsity on analysis outcomes. Moreover, lemmatization also helps to increase the accuracy of analysis results by treating variant forms of a word as a single entity. By reducing the dimensionality of the data, lemmatization can significantly improve the efficiency and effectiveness of textual data analysis, especially when dealing with large datasets. The resulting reduction in computational complexity enables analysts to focus on relevant information, thereby enhancing the quality of insights gleaned from the data.

4.5.3 Stop word removal.

Stop word removal is a technique in textual data analysis that entails the elimination of frequently used words, such as "the," "and," "a," and "is," which are not semantically meaningful and are unlikely to offer valuable insights during the analysis. The removal of stop words is a crucial technique in textual data analysis that helps to refine the focus of the analysis. Additionally, eliminating terms that are irrelevant to a particular topic or research context can enhance the accuracy and quality of the textual data analysis. For example, in the context of this research, tokens such as "review," "hotel," and "nights" could be excluded to strengthen the significance and relevance of the textual data analysis. Through this approach, the analysis can concentrate on the significant and pertinent terms in the text.

4.5.4 CountVectorizer.

CountVectorizer is a widely used tool of the Scikit-learn library for Python in the fields of textual data analysis and machine learning that enables the transformation of textual data into numerical feature vectors. This tool is instrumental in the analysis of large volumes of unstructured textual data, as it allows such data to be represented in a format that can be more easily analysed and processed by machine learning algorithms.

In practical terms, CountVectorizer operates by converting a collection of textual data into a matrix of token counts. This matrix represents the frequency of words within the textual data, which enables the creation of a numerical representation of the data. By doing so, CountVectorizer enables the identification of patterns, trends, and insights within the data that may otherwise be difficult to discern. Furthermore, CountVectorizer is a highly customizable tool, and can be configured to use different parameters and settings depending on the specific needs and requirements of the analysis.

4.5.5 NLTK

Natural Language Toolkit (NLTK) is a widely used open-source software library for the Python programming language, designed to facilitate the exploration and analysis of natural language data. In the context of this research, NLTK is employed as a key component of the methodology to enable the processing, cleaning, and analysis of textual data. To be more specific, the toolkit provides a range of functionalities and techniques for the textual data analysis of this research, including tokenization, lemmatization, and stop word removal, which are essential for the effective analysis of large volumes of unstructured textual data.

Moreover, NLTK offers a range of pre-built corpora, including the Brown Corpus, which can be leveraged to improve the accuracy and efficiency of textual data analysis. Overall, the use of NLTK in this research helped to facilitate the identification of patterns, trends, and insights within the textual data, thereby contributing to the overall rigor and validity of the research findings.

5 RESULTS

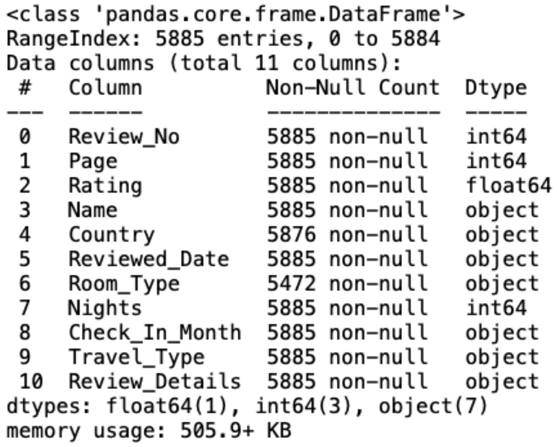
5.1 Exploratory data analysis

According to Wongsuphasawat et al. (2019), exploratory data analysis can be divided into two main goals: profiling and discovery. Profiling involves gaining an understanding of the data and determining whether it is suitable for further analysis. This can be accomplished by broadly examining the data and its visualizations to identify its distribution, shape, and quality issues, such as missing data, extreme values, or inconsistent data types. By conducting a thorough profiling of the data, analysts can establish a solid foundation for subsequent analytical tasks. The second goal of exploratory analysis, as described by Wongsuphasawat et al. (2019), is discovery, which involves uncovering new insights or generating hypotheses through data exploration. Analysts may develop intuitive ideas about how to answer questions or create models by examining potential relationships between variables or identifying the importance of different features through techniques such as variable ranking. This process of discovery can help guide further analytical and hypothesis-testing activities.

In this research, the following exploratory data analysis and data pre-processing setps have been conducted:

5.1.1 Check missing values.

Based on the information presented in Figure 6, it can be determined that the dataset contains a total of 5885 reviews (instances). Additionally, it was observed that the *Country* and *Room_Type* columns contain missing values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5885 entries, 0 to 5884
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Review_No       5885 non-null   int64
 1   Page            5885 non-null   int64
 2   Rating          5885 non-null   float64
 3   Name            5885 non-null   object
 4   Country         5876 non-null   object
 5   Reviewed_Date   5885 non-null   object
 6   Room_Type       5472 non-null   object
 7   Nights          5885 non-null   int64
 8   Check_In_Month  5885 non-null   object
 9   Travel_Type     5885 non-null   object
 10  Review_Details  5885 non-null   object
dtypes: float64(1), int64(3), object(7)
memory usage: 505.9+ KB
```

Figure 6 Information of the dataset

5.1.2 Exclude missing values.

As the presence of missing values may negatively impact the accuracy of subsequent predictive modelling, it was deemed necessary to remove any rows containing such values in this research. Therefore, any row in the dataset that contained missing values was excluded from further analysis to ensure the integrity of the predictive modelling results. Upon removal of all rows containing missing values, there are 5464 reviews (instances) left.

5.1.3 Check the statistical analysis of the numerical columns.

Table 10 provides a summary of the statistical characteristics of two variables, *Rating* and *Nights*. *Rating* ranges from 1 to 10. The mean rating is 8.4, indicating that the dataset contains mostly positive ratings. The dataset also includes information about the number of nights stayed, with a range from 1 to 17. The mean number of nights stayed is 1.69 and 75% of the instances in the *Nights* column have a value of 2 or less. This indicates a tendency for shorter stays, with the majority of guests staying no longer than 2 nights. These findings are corroborated by Figures 7 and 8, which illustrate the distributions of *Rating* and *Nights* respectively. The visualizations indicate that most review ratings fall in the range of 8, 9, and 10, while the majority of guests stayed for only 1 or 2 nights. Overall, this statistical analysis provides a quick summary of the central tendency, variability, and distribution of the dataset, which can be used to guide further analysis.

Table 10 Statistical information of numerical columns

	Rating	Nights
Count	5464	5464
Mean	8.41	1.69
Std	1.59	1.13
Min	1	1
25%	8	1
50%	9	1
75%	10	2
Max	10	17

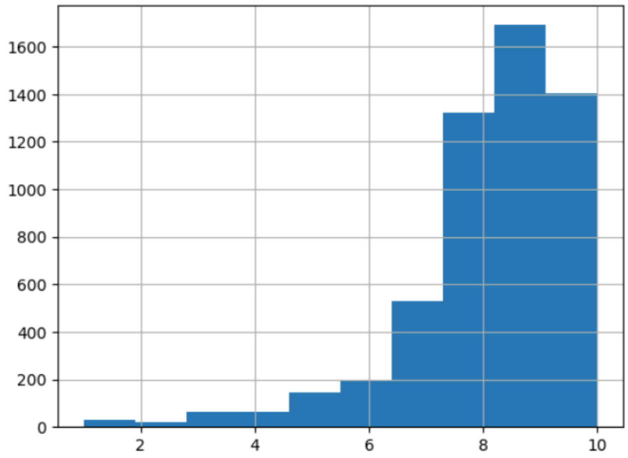


Figure 7 The distribution of Rating

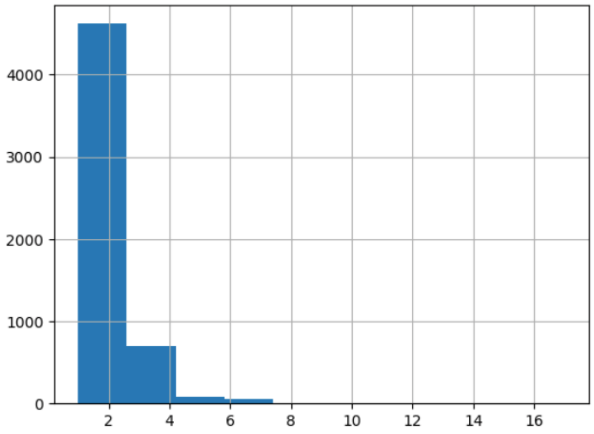


Figure 8 The distribution of Nights

5.1.4 Country distribution and ratings.

An analysis of *Country* revealed that guests arrived from a total of 91 countries. Figure 9 displays the top 10 countries with the highest number of guests. As indicated by the figure, Finland was the country with the most guests, followed by Germany and Estonia. Instances of Finland have an absolute numerical advantage accounting for 51%, about 8 times higher than the second-ranked Germany.

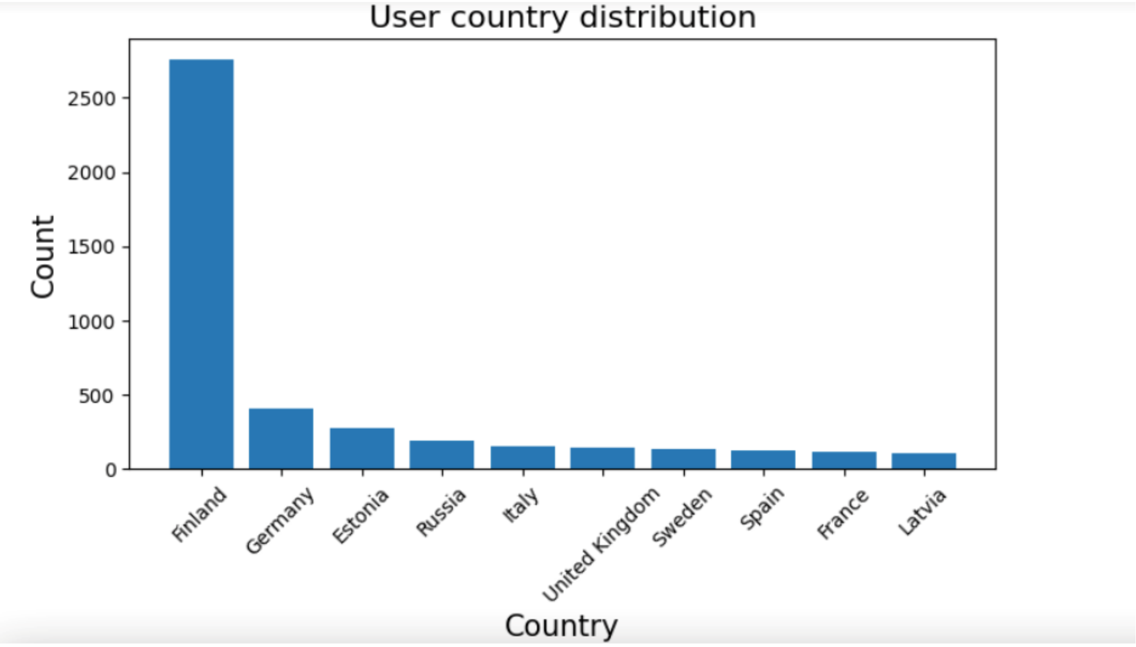


Figure 9 The top 10 countries with the most guests

The findings presented in Figure 10 indicate that among the top 10 countries with the most guests, the highest average rating for the hotel was awarded by guests from Latvia, whereas guests from Germany provided the lowest average rating. Notably, guests from Finland provided an average rating that slightly exceeded 8.4, a value slightly higher than the overall rating of the hotel.

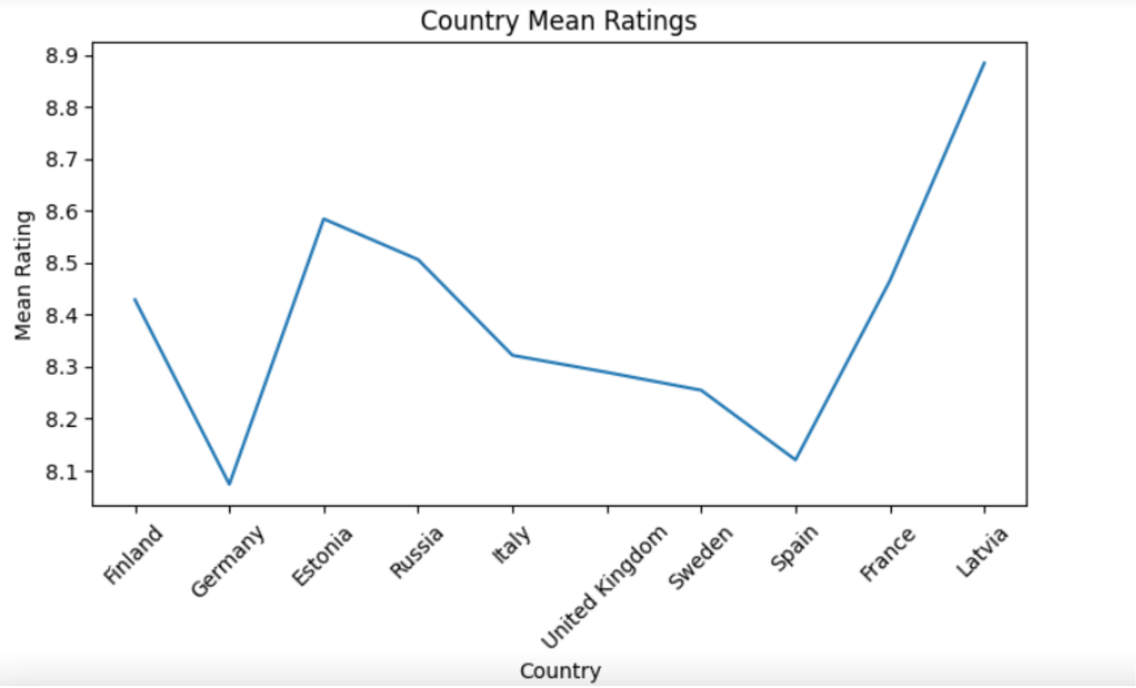


Figure 10 The mean rating of top 10 countries

5.1.5 Room type distribution and ratings.

The hotel under research offers five distinct types of rooms, namely standard room, superior room, business room, one-bedroom suite, and executive suite. The prevalence of these room types in prior reservations is illustrated in Figure 11. Standard Room and Superior Room are the most common room types, followed by Business Room, One-Bedroom Suite, and Executive Suite, which is the least frequently booked room type. A substantial disparity between the number of instances of the most common room type, Standard Room (3478), and the least common room type, Executive Suite (3), exists in the dataset and deserves attention. Moreover, a noteworthy observation is that the combined number of instances for the top two ranked room types, standard room and superior room, constitutes over 96% of the total instances in the dataset.

In Figure 12, the mean rating varies across different room types. Executive Suite has the highest mean rating of 10.0, followed by Superior Room at 8.67, One-Bedroom Suite at 8.5, Business Room at 8.42, and Standard Room at 8.27. The combination of these two pieces of information suggests that the most prevalent room types, Standard Room has relatively lower mean ratings than other room types. Simultaneously, this room type is the sole category that receives a rating lower than the overall rating of the hotel. This finding may indicate that the more exclusive or higher-end room types are more likely to receive higher ratings. However, further analysis on more data would be necessary to validate this hypothesis.

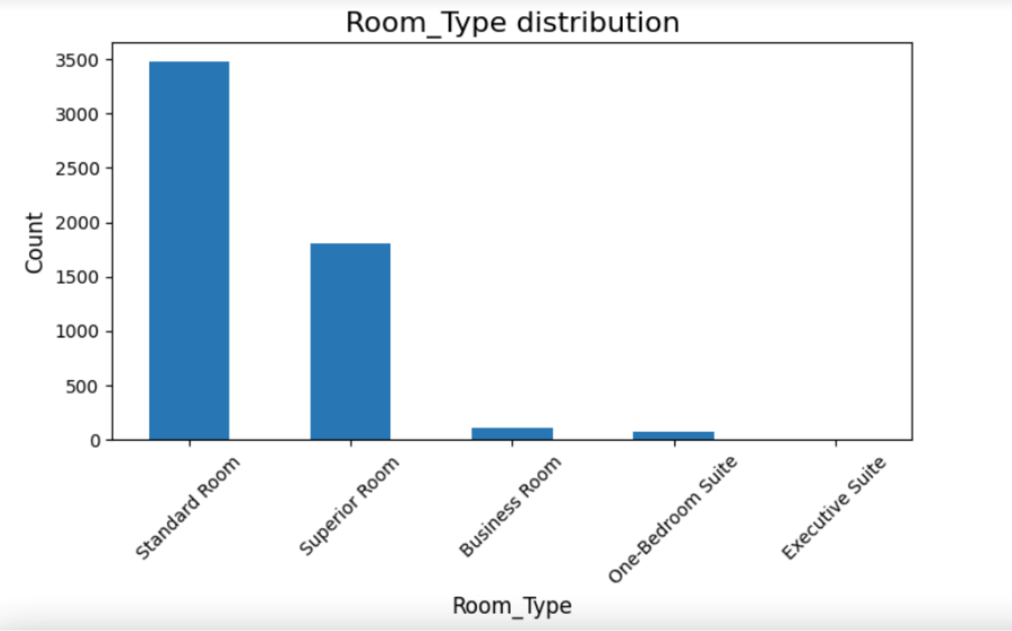


Figure 11 Room_Type distribution

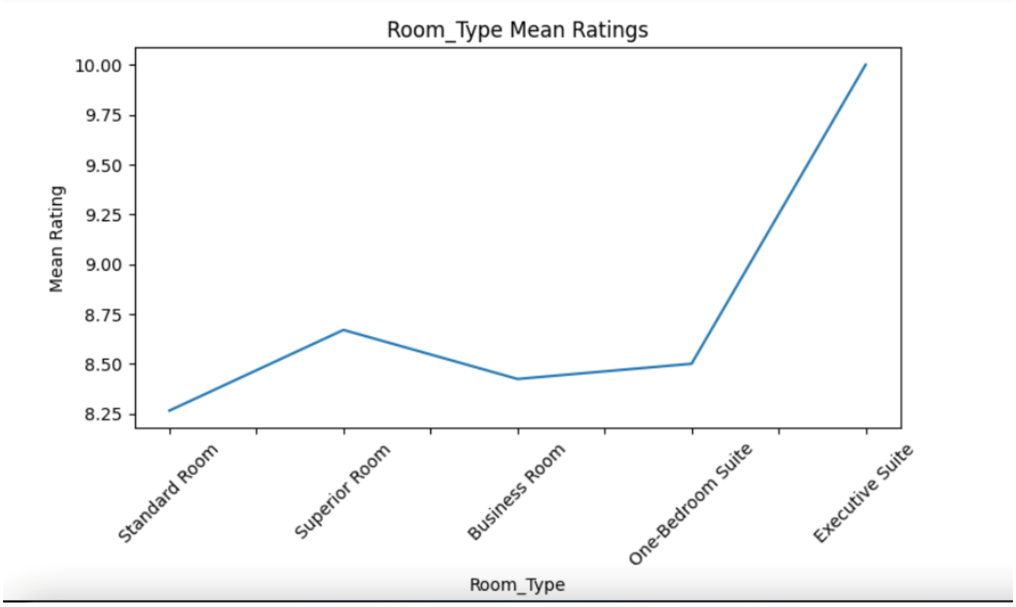


Figure 12 The mean rating of different Room_Type

5.1.6 Travel type distribution and ratings.

Figure 13 presents a breakdown of the types of travellers who have made reservations at the hotel. The four categories identified are families, couples, groups, and solo travellers. The data indicate that families (2148) and couples (1700) represent the majority of the total reservations (5464), followed by groups (810) and solo travellers (806). These proportions translate into percentages of 39.3% for families, 31.1% for couples, 14.8% for groups, and 14.7% for solo travellers.

Figure 14 depicts the mean ratings given by different types of travellers to the hotel. The overall rating of the hotel is 8.4. The mean ratings for family and group travelers are higher than the overall rating, with family travellers giving the highest mean rating of 8.55, followed by group travellers with a mean rating of 8.47. In contrast, the mean ratings for couples (8.25) and solo travellers (8.26) are slightly lower than the overall rating.

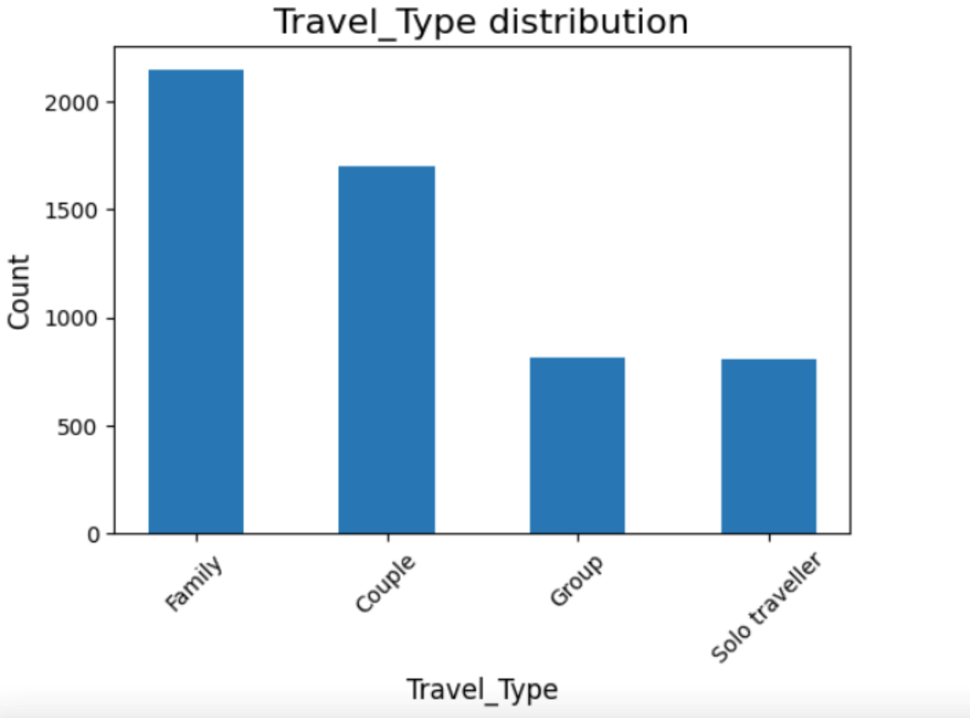


Figure 13 Travel_Type distribution

These findings suggest that families and groups are more satisfied with their hotel experience than couples and solo travellers. However, given that families and couples constitute the majority of travellers, it appears that the hotel caters primarily to these groups. The lower ratings given by couples and solo travellers imply that the hotel may need to make some improvements to better serve their needs and improve their satisfaction. By addressing the needs of couples and solo travelers and improving their level of satisfaction, the hotel may attract a larger share of these traveler types in the future. Furthermore, such efforts may contribute to enhancing the hotel's online reputation, increasing its number of reservations, and ultimately improving its overall rating.

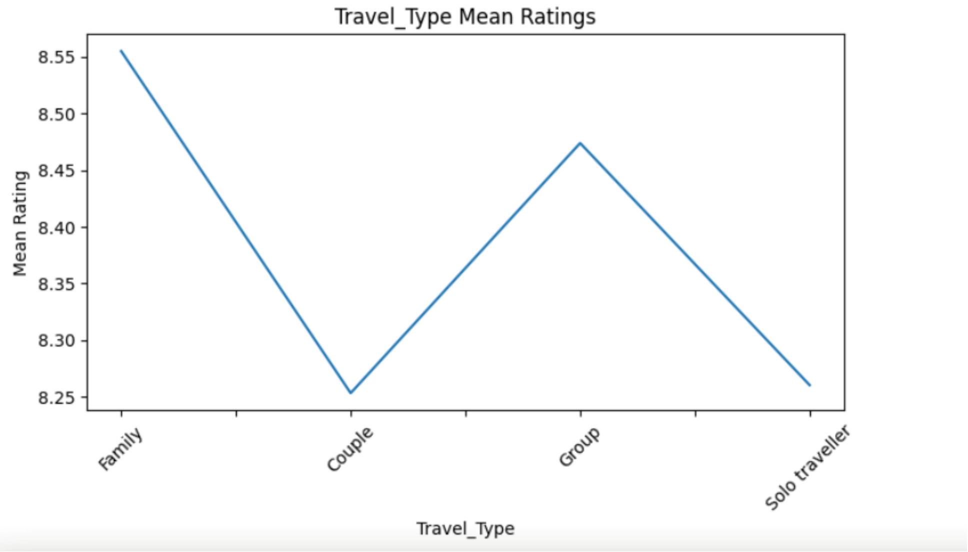
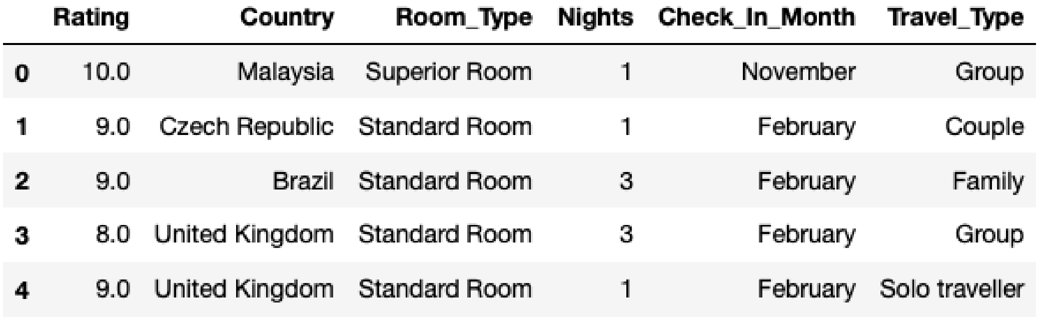


Figure 14 The mean rating of different Travel_Type

5.1.7 Month distribution and ratings.

Prior to conducting an analysis of the distribution of months and its correlation with ratings, undertaking certain measures with respect to the column of *Check_in_Month* is imperative. The original dataset presents the column of *Check_in_Month* in a format that incorporates both the month and year components. However, the year component is expected to increase monotonically over time due to future reservations having larger year values. Failure to remove the year component of the column may introduce biases and distortions in our analysis and projections. Hence, extracting and retaining only the month component of the column of *Check_in_Month* is essential. Following this pre-processing step, the modified dataset assumes the form depicted in Figure 15.



	Rating	Country	Room_Type	Nights	Check_In_Month	Travel_Type
0	10.0	Malaysia	Superior Room	1	November	Group
1	9.0	Czech Republic	Standard Room	1	February	Couple
2	9.0	Brazil	Standard Room	3	February	Family
3	8.0	United Kingdom	Standard Room	3	February	Group
4	9.0	United Kingdom	Standard Room	1	February	Solo traveller

Figure 15 Datasets after remove year in the column of Check_in_Month

Analysis of Figure 16 provides insight into the monthly distribution of the number of reservations within the dataset. The data indicates that July exhibited the highest number of the number of reservations, with a count of 1045, followed by August with 940 reservations. In contrast, the months of April and March recorded the lowest number of reservations, with respective counts of 259 and 283. The mean value of reservations across all months is 455. Seven months, including January, February, March, April, May, November, and December, fall below the mean value of reservations across all months. Additionally, July and August display an amount of absolute advantage.

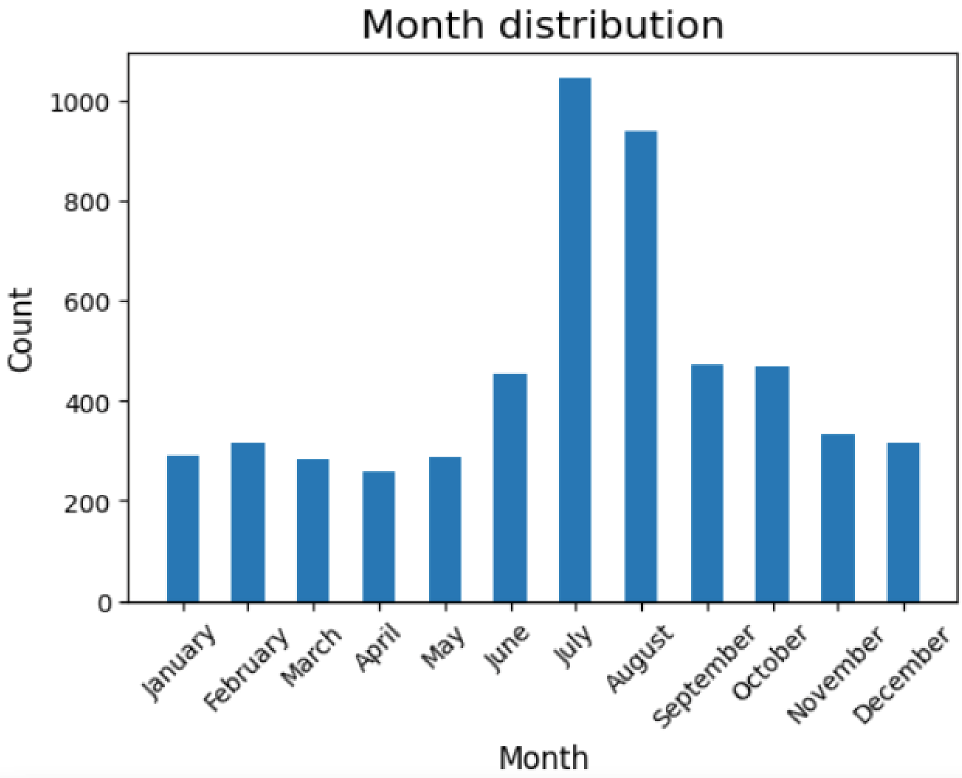


Figure 16 Check_in_Month distribution

As shown in Figure 17, March has the highest average rating, followed by January and February, while October and July have the lowest. The lower average ratings during these months suggest that guests are generally less satisfied with their experience at the hotel. Furthermore, the data shows that the average ratings in the first half of the year are higher than those in the second half.

When considering the number of reservations in each month, July and August have the highest number of reservations but also have a relatively lower average rating. This indicates that the hotel may struggle to maintain the same level of quality and satisfaction during peak periods. In contrast to other months, January, February, and March exhibit a relatively higher mean rating, yet fewer reservations, indicating that the hotel staff can concentrate more on each guest or reservation resulting in higher satisfaction. However, this also implies that the hotel may need to intensify its promotional efforts to attract more guests during these months to increase its popularity.

The analysis of months and ratings highlights the need for the hotel to implement strategies to maintain a consistent level of quality and satisfaction during peak periods, such as hiring additional staff or increasing operational efficiency. Additionally, the hotel may benefit from marketing and promotional activities to increase its popularity and attract more reservations during the lower peak months of the year, such as January, February, and March.

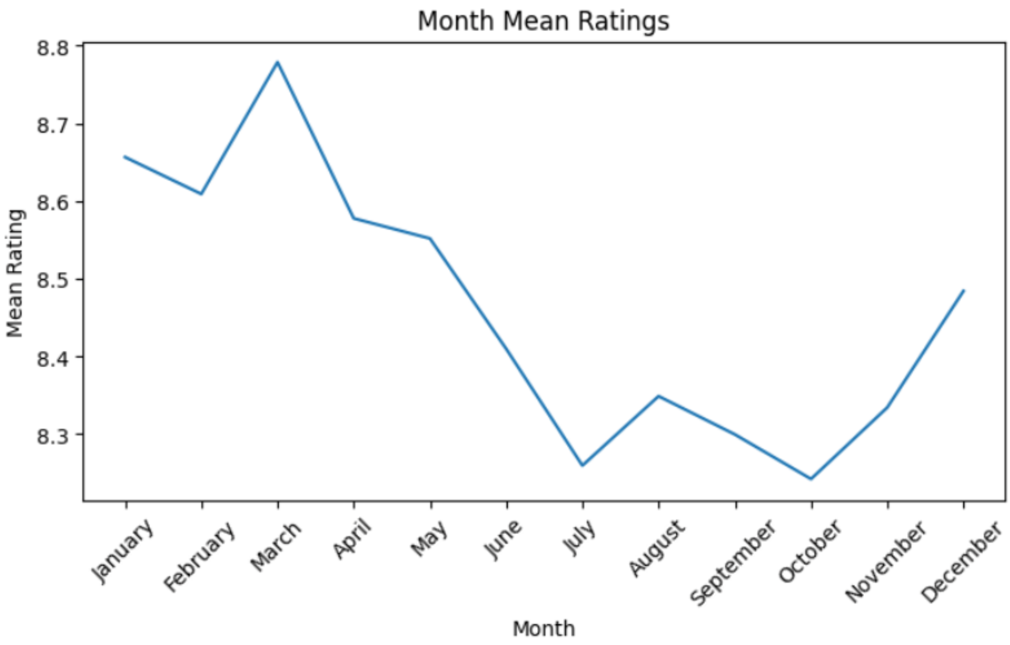


Figure 17 The mean rating of different Check_in_Month

5.1.8 Nights distribution and ratings.

The *Nights* column in the dataset contains 13 distinct values, ranging from 1 to 17, indicating the range of durations for which guests typically stayed at the hotel. Specifically, these 13 unique values represent the number of nights for which guests stayed at the hotel for all reservations included in the dataset.

As shown in Figure 18, among the 13 unique values, the most frequent stay duration was 1 night, with a count of 3206 reservations, accounting for 59%, followed by 2 nights with 1413 reservations, accounting for 26%. The number of reservations for stays longer than 5 nights decreases rapidly, with only 29 reservations for stays of 6 nights. Notably, a significant difference has been observed between the number of reservations for stays of 1 or 2 nights compared to those of longer durations. Specifically, 85% of all reservations in the dataset were for stays of 1 or 2 nights, while only 15% of reservations were for stays longer than 2 nights. This indicates that guests tend to prefer shorter stays at the hotel, with a significant majority of guests staying for one or two nights.

The data indicates that the customers who stayed at the hotel for a duration of 9 nights provided the highest average rating of 9.0 out of 10, although this is based on a small sample size of only 2 reservations. Following closely, those who stayed for 1 or 2 nights gave the next highest ratings, with averages of 8.46 and 8.40 respectively, which is either equal to or higher than the overall rating of the hotel (8.4). However, customers who stayed for 11 or 16 nights gave the lowest average ratings of 7.0 out of 10, as illustrated in Figure 19. These findings suggest that the length of a customer's stay may have a significant influence on their perception of the hotel's quality.

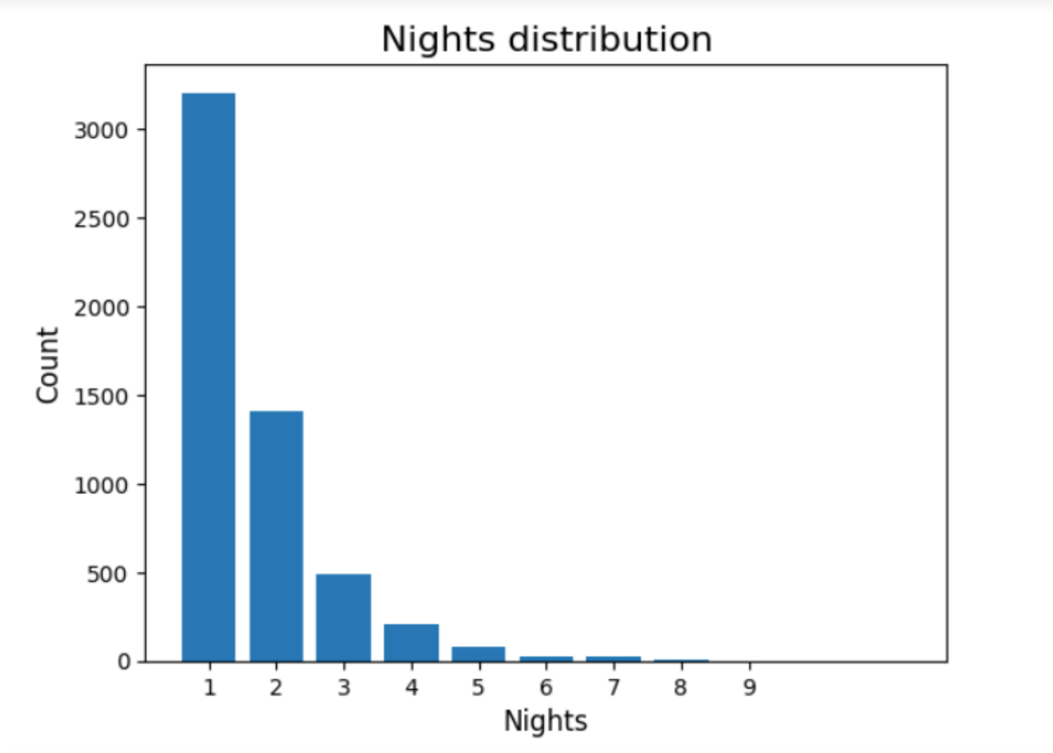


Figure 18 Nights distribution (Top 9)

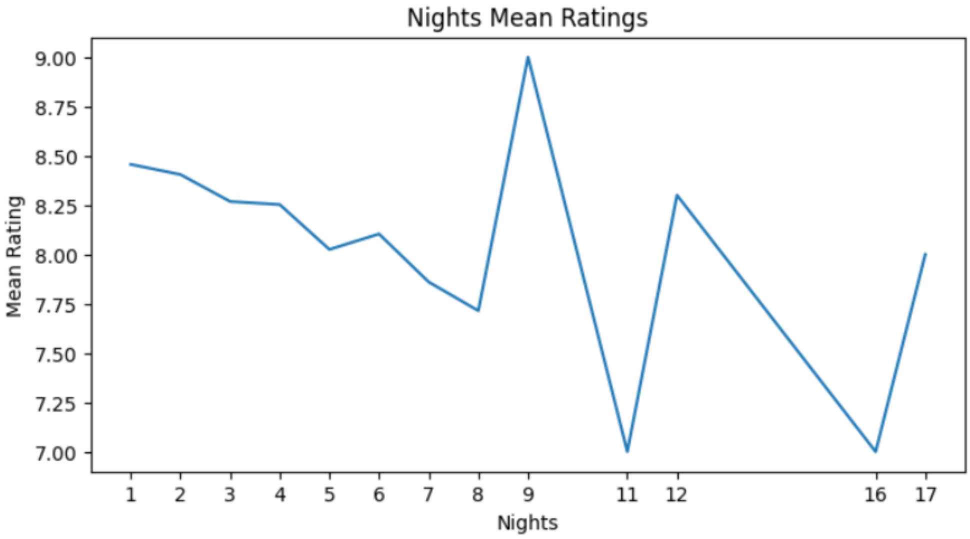


Figure 19 The mean rating of different Nights

5.1.9 Summarize for exploratory data analysis.

The exploratory data analysis of the dataset included information about the total number of instances, the presence of missing values. The dataset contains 5885 reviews, but after removing rows with missing values, 5464 reviews remained for further analysis. The analysis also provided a summary of the central tendency, variability, and distribution of the dataset for two numerical variables, *Rating* and *Nights*. The mean rating was 8.4, indicating mostly positive ratings, while the mean number of nights stayed was 1.69, with a tendency for shorter stays.

A thorough analysis of the dataset's *Country* column revealed the presence of guests originating from 91 countries. Among these countries, Finland accounted for the highest proportion of guests, representing 51% of the total instances, followed by Germany and Estonia. The top 10 countries with the most guests were further examined with respect to their average rating. Guests from Latvia had the highest average rating, whereas guests from Germany had the lowest. The *Room_Type* column was also scrutinized, and the standard room and superior room were the predominant room types, comprising over 96% of the total instances in the dataset. The Executive Suite emerged as the highest-rated room type, followed by Superior Room, One-Bedroom Suite, Business Room, and Standard Room. The types of travellers were analysed, and families and couples were found to be the most prevalent, followed by groups and solo travellers. Families and groups were more satisfied with the hotel experience compared to couples and solo travellers. Furthermore, an investigation into the monthly distribution of reservations and its relationship with ratings revealed that July and August had the highest number of reservations, while April and March had the lowest. However, March exhibited the highest average rating, followed by January and February, whereas October and July displayed the lowest. The *Nights* column represents the length of stay for guests, with a minimum of 1 night and a maximum of 17 nights. The mean number of nights stayed is 1.69, indicating that guests typically had short stays. Specifically, 85% of the *Nights* column instances have a value of 2 or less, suggesting that the majority of guests stayed for only 1 or 2 nights.

5.2 Data Pre-processing

According to Kotsiantis et al. (2006), the representation and quality of instance data are primary factors that influence the success of machine learning tasks. In the previous analysis, certain data pre-processing steps were taken, such as the removal of missing values and the adjustment of *Check_in_Month* values. However, handling categorical predictors with a large number of levels or categories can pose a challenge in data analysis (Criscuolo et al., 2023). In this research, all predictors were treated as categorical, thereby presenting a challenge in dealing with the categorical values within each predictor. Nevertheless, various data pre-processing algorithms should be employed based on the characteristics of the original dataset. In this regard, an attempt was made to address the issue using sound judgment and experience. To improve model performance, several categorical predictors were processed after several attempts in the following manner:

5.2.1 Country.

An examination of the *Country* column in the dataset revealed the existence of guests hailing from 91 different countries. Notably, guests from Finland constituted the highest proportion, accounting for 51% of the total instances, while the remaining 90 countries accounted for the remaining 49%. Upon further investigation, among the top 10 countries with the highest number of guests, Germany had the lowest average rating for the hotel. As a result, the value of the *Country* column was transformed from its original 91 categories to three categories: Finland, Germany, and Others. This modified dataset was subsequently employed to train and test the machine learning models.

5.2.2 Nights.

The range of values within the *Nights* column provides valuable insights into guest behaviour and preferences, and can be expected to be valuable for determine ratings. However, due to a substantial disparity in the proportion of reservations between those for 1 or 2 nights and those for longer durations, the following research will treat the *Nights* column as categorical column rather than numerical column when building prediction models. Consequently, the *Nights* column will be transformed from 13 unique numerical values into three categorical value columns, representing 1 night, 2 nights, and more than 2 nights. This approach aims to account for the marked difference in the distribution of reservations based on the number of nights stayed and enhance the accuracy of the predictive models.

Upon completing the pre-processing steps, the resulting modifications to the dataset will produce a refined and standardized format that aligns with the desired structure. This updated dataset will provide a more precise and accurate representation of the underlying data, enabling more informed decision-making and reliable conclusions to be drawn from the analysis. The expected format of the column *Nights* in the processed dataset contains value of 1 night, 2 nights and more than 2 nights.

5.2.3 Rating (Target variable).

The purpose of this research is to develop machine learning models for a binary classification problem focused on predicting rating levels of future reservations. To accomplish this objective, a pre-processing stage is required for the original dataset's target variable, namely *Rating*. This variable is a numerical representation of guests' ratings on a scale from 1 to 10, indicating their level of satisfaction with their hotel experience. However, to predict whether a reservation would result in a rating above or below a certain threshold, enabling the hotel to take proactive measures to prevent lower guest ratings and maintain a favourable online reputation, a binary classification approach is necessary.

In this research, the threshold for the binary classification problem is set as 8.4, which corresponds to the overall rating of the hotel on the Booking.com platform. Any reservation resulting in a rating below this threshold is classified as class 0, indicating a lower rating level, whereas any reservation resulting in a rating of 8.4 or higher is classified as class 1, indicating a higher rating level. The conversion of the original numerical *Rating* variable to a binary target variable enable to develop an accurate and reliable machine learning models for this binary classification problem.

By training the model on this binary classification problem, the hotel could anticipate which reservations are more likely to result in a lower rating and take proactive measures to prevent this lower rating of the future reservation. This could, in turn, enhance the hotel's online reputation, a critical factor in the competitiveness of the hotel industry. Thus, the data pre-processing approach played a crucial role in preparing the dataset for machine learning, leading to the development of an accurate and reliable prediction model for this binary classification problem.

5.3 Models and performances

The present research endeavours to utilize four distinct machine learning algorithms, namely logistic regression, random forest, XGBoost, and ANN, to predict the rating category (*rating_cat*) of future reservations based on several features including the country of origin (*Country*), room type (*Room_Type*), travel type (*Travel_Type*), check-in month (*Check_in_Month*), and length of stay (*Nights*). However, due to the limited availability of training data, the inclusion of an excessive number of features may lead to a significant slowdown in the learning process. Moreover, irrelevant or redundant features can potentially mislead the learning algorithms and lead to over-fitting of the training data (Yu et al., 2004). Consequently, a feature selection approach to identify a concise subset of informative variables that minimizes the data measurement, storage, and transmission costs while maximizing model performance is adopted. Ultimately, this can result in more concise classification models with better generalization capabilities (Hu et al., 2008).

Figure 20 depicts a traditional framework of feature selection through subset evaluation (Liu et al., 1998), which involves generating all possible candidate feature subsets. In this research, feature selection through subset evaluation is conducted by assessing each possible subset using a consistent measure, namely the higher TN_score. The performances of the resulting models generated from all possible subsets are recorded for identification of the optimal subset and subsequent comparison.

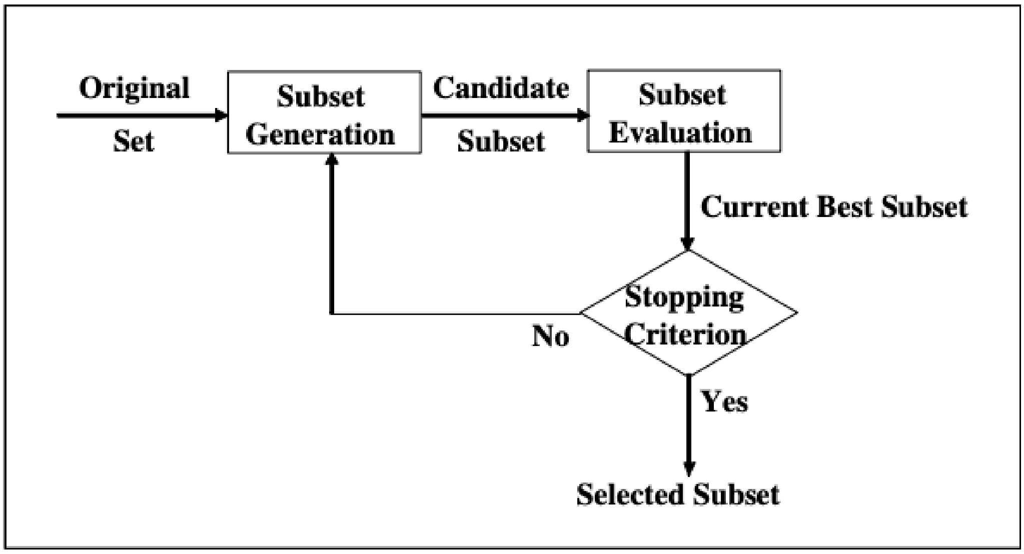


Figure 20 A traditional framework of feature selection

Furthermore, a crucial aspect of machine learning, namely model optimization, is conducted in this research. Model optimization is recognized as an essential component of machine learning as it aims to improve the performance of the machine learning models (Sun et al., 2019). Specifically, the optimization process in this research focused on fine-tuning the hyperparameters of each algorithm to identify the most optimal hyperparameters for achieving the highest TN_score, a metric defined in this research.

Hyperparameter tuning was performed using the GridSearchCV function provided by the Scikit-learn library. This approach involves searching over a grid of hyperparameters to identify the combination that resulted in the best cross-validation score (Pedregosa et al., 2011), and GridSearchCV has been widely used in various machine learning applications. For each algorithm, a range of hyperparameters were selected for tuning, based on their known impact on model performance. For example, in this research, these included the regularization strength and solver algorithm for logistic regression (Hastie, Tibshirani, & Friedman, 2009), the number of trees, maximum depth of the trees, and minimum number of samples required to split a node for random forest (Breiman, 2001), the learning rate, maximum depth of the trees, and number of trees for XGBoost (Chen et al., 2016), and the number of hidden layers, number of neurons per layer, activation function, and learning rate for ANN (Goodfellow et al., 2016).

In this research, each possible subset performed a GridSearchCV optimization. The combination of subset feature selection and GridSearchCV optimization enabled the identification of the most relevant features and the optimal hyperparameters of each algorithm. As a result, the models achieved better performances and generalization capabilities, which are critical for real-world applications of predictive modelling.

After the processes of the combination of subset feature selection and GridSearchCV optimization, the optimal hyperparameters for each subset of each algorithm are identified based on the highest mean cross-validation score. The performance of each subset of each algorithm is then evaluated using the TN_score metric and the subset with optimal hyperparameters that achieved the highest TN_score of each algorithm are identified and recorded for subsequent comparison.

5.3.1 Logistic regression

The logistic regression model is a commonly used method in machine learning for binary classification tasks. In this research, logistic regression is the first machine learning model employed to predict the ratings' category of future reservations for the hotel. The purpose of this section is to present and compare the results of the logistic regression model with all possible feature subsets.

The dataset in this research comprises six categorical variables, namely *Country*, *Room_Type, Travel_Type, Check_In_Month, Nights*, and *rating_cat*. The total number of features is five since the target variable *rating_cat* is not included in this count. According to the theory of power sets, the total number of all the possible subsets that can be formed from a set of n elements is given by the equation 2^n. In the context of this research, since the value of n is equal to 5, there are 32 possible subsets in the power set of the dataset. Each subset represents a unique combination of features that can be included or excluded from the original set. This technique can be applied in various fields, such as combinatorics, probability theory, and computer science, to enumerate all possible combinations of a set. Therefore, a total of 31 possible feature subsets can be used to build the logistic regression model excluding the empty subset. Furthermore, several hyperparameters are used to optimize each logistic regression model with different feature subset including "C", "solver", and "max_iter". "C" is the regularization strength hyperparameter in logistic regression, and "C" is used to control overfitting by penalizing large coefficient values. A smaller value of “C” corresponds to stronger regularization, and a larger value of “C” corresponds to weaker regularization. Hyperparameter “solver” specifies the algorithm used to optimize the logistic regression model. Different solvers are available in scikit-learn, such as ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’, and ‘saga’. Each solver has its own strengths and weaknesses, and the choice of solver depends on the dataset size and complexity. Hyperparameter “max_iter” controls the maximum number of iterations for the solver to converge. If the solver has not converged after this many iterations, it stops and returns the current solution. A higher value of max_iter may improve the model's accuracy, but it also increases the computation time.

Hyperparameters in logistic regression are crucial for improving the performance of the model. In order to determine the optimal hyperparameters for each subset, various values of the hyperparameters are tested. The corresponding values tested for each hyperparameter are presented in Table 11. The ultimate aim is to identify the best-performing model for each subset, with the optimal hyperparameters. The results of these models, along with the best TN_score and corresponding confusion matrix, are summarized in Table 12.

Table 11 The values tested of each hyperparameter of logistic regression

Hyperparameters	Values tested
C	[0.001, 0.01, 0.1, 1, 10, 100]
solver	['liblinear', 'saga', 'lbfgs', 'newton-cg']
max_iter	[100, 500, 1000]

Table 12 Performances of 31 logistic regression models

No.	Feature subset	C	solver	max_iter	TN_score	Confusion matrix
1	['Country']	0.1	'liblinear'	100	0.12	[[30 307] [22 461]]
2	['Room_Type']	0.001	'liblinear'	100	Nan	[[0 337] [0 483]]

3	['Travel_Type']	0.001	'liblinear'	100	Nan	[[0 337] [0 483]]
4	['Check_In_Month']	0.001	'liblinear'	100	Nan	[[0 337] [0 483]]
5	['Nights']	0.1	'liblinear'	100	0.14	[[57 280] [67 416]]
6	['Country', 'Room_Type']	0.01	'liblinear'	100	0.09	[[23 314] [18 465]]
7	['Country', 'Travel_Type']	10	'liblinear'	100	0.12	[[30 307] [22 461]]
8	['Country', 'Check_In_Month']	10	'liblinear'	100	0.13	[[26 311] [16 467]]
9	['Country', 'Nights']	1	'liblinear'	100	0.07	[[24 313] [26 457]]
10	['Room_Type', 'Travel_Type']	0.001	'liblinear'	100	Nan	[[0 337] [0 483]]
11	['Room_Type', 'Check_In_Month']	1	'liblinear'	100	0.42	[[143 194] [143 340]]
12	['Room_Type', 'Nights']	0.1	'saga'	100	0.09	[[34 303] [40 443]]
13	['Travel_Type', 'Check_In_Month']	10	'liblinear'	100	0.33	[[100 237] [90 393]]
14	['Travel_Type', 'Nights']	1	'liblinear'	100	0.12	[[36 301] [33 450]]
15	['Check_In_Month', 'Nights']	1	'liblinear'	100	0.15	[[46 291] [42 441]]
16	['Country', 'Room_Type', 'Travel_Type']	1	'liblinear'	100	0.18	[[57 280] [54 429]]
17	['Country', 'Room_Type', 'Check_In_Month']	10	'saga'	100	0.50	[[139 198] [115 368]]

18	['Country', 'Room_Type', 'Nights']	0.1	'liblinear'	100	0.14	[[51 286] [54 429]]
19	['Country', 'Travel_Type', 'Check_In_Month']	10	'saga'	100	0.32	[[87 250] [71 412]]
20	['Country', 'Travel_Type', 'Nights']	1	'liblinear'	100	0.15	[[43 294] [36 447]]
21	['Country', 'Check_In_Month', 'Nights']	100	'liblinear'	100	0.17	[[53 284] [49 434]]
22	['Room_Type', 'Travel_Type', 'Check_In_Month']	10	'saga'	100	0.44	[[139 198] [130 353]]
23	['Room_Type', 'Travel_Type', 'Nights']	10	'liblinear'	100	0.18	[[66 271] [73 410]]
24	['Room_Type', 'Check_In_Month', 'Nights']	100	'liblinear'	100	0.36	[[115 222] [108 375]]
25	['Travel_Type', 'Check_In_Month', 'Nights']	10	'liblinear'	100	0.31	[[97 240] [89 394]]
26	['Country', 'Room_Type', 'Travel_Type', 'Check_In_Month']	1	'liblinear'	100	0.43	[[127 210] [112 371]]
27	['Country', 'Room_Type',	100	'liblinear'	100	0.21	[[78 259] [84 399]]

	'Travel_Type', 'Nights']					
28	['Country', 'Room_Type', 'Check_In_Month', 'Nights']	10	'saga'	100	0.38	[[117 220] [108 375]]
29	['Country', 'Travel_Type', 'Check_In_Month', 'Nights']	10	'liblinear'	100	0.31	[[94 243] [85 398]]
30	['Room_Type', 'Travel_Type', 'Check_In_Month', 'Nights']	10	'liblinear'	100	0.43	[[129 208] [115 368]]
31	['Country', 'Room_Type', 'Travel_Type', 'Check_In_Month', 'Nights']	1	'saga'	100	0.45	[[122 215] [98 385]]

Analysis of the results presented in Table 12 indicates that the 17th model, with the feature subset of *Country*, *Room_Type*, *Check_In_Month* and the optimized hyperparameters {“C”:10, “solver”: “saga”, “max_iter”: 100}, achieves the best performance, with a TN_score of 0.5. However, despite this seemingly best outcome, further examination reveals that the model's overall performance falls short of expectations. Specifically, the confusion matrix demonstrates that out of the total of 337 actual negative instances, only 41% (139 instances) were accurately predicted as negative by the model. Meanwhile, of the 254 instances predicted as negative, 45% (115 instances) were in fact positive, indicating that the model is still failing to accurately classify a significant proportion of instances.

5.3.2 Random forest

Logistic regression has some limitations, such as its linear nature and its reliance on a specific set of assumptions. Therefore, another algorithm random forest is tested in this research. Random forest is a popular ensemble learning algorithm used in machine learning for classification and regression tasks. Furthermore, random forest is a more flexible and robust machine learning algorithm that can capture non-linear relationships and interactions between variables.

This section presents and compares the results of the random forest models used to predict the categories of future reservations' rating categories for the hotel. Models with 31 possible subsets of the five features are tested in this research. Furthermore, random forest also has several hyperparameters that can be adjusted to optimize the model's performance. In this research, to find the best-performing models with optimal hyperparameters, a range of hyperparameters are tested, including n_estimators, max_depth, min_samples_split, and criterion. The n_estimators hyperparameter determines the number of decision trees to be used in the forest. Increasing the number of trees can improve model performance, but it also increases computation time. The max_depth hyperparameter determines the maximum depth of each decision tree in the forest. A higher max_depth can lead to overfitting, while a lower max_depth can lead to underfitting. The min_samples_split hyperparameter sets the minimum number of samples required to split a node. The min_samples_split hyperparameter can help prevent overfitting of the model to the training data. Finally, the criterion hyperparameter determines the function to measure the quality of a split. The most commonly used criterion is Gini, but the criterion entropy is also an option. Overall, optimizing the hyperparameters of a random forest model can significantly improve its performance. The corresponding values tested for each hyperparameter are presented in Table 13. By optimizing these hyperparameters, the performance of the random forest models can be improved, providing more accurate predictions of the categories of future reservations' rating for the hotel. The best results and the optimal hyperparameters of the random forest model are displayed in Table 14.

Table 13 The values tested of each hyperparameter of random forest

Hyperparameters	Values tested

n_estimators	[300,600,1000]
max_depth	[4,6,10]
min_samples_split	[10, 100,200]
criterion	['gini', 'entropy']

Table 14 The best results and the optimal hyperparameters of the random forest model

No.	Feature subset	n_estimators	max_depth	min_samples_split	criterion	TN_score	Confusion matrix
1	['Country']	300	20	2	gini	0.12	[[30 308] [22 460]]
2	['Room_Type']	300	20	2	gini	nan	[[0 338] [0 482]]
3	['Travel_Type']	300	20	2	gini	nan	[[0 338] [0 482]]
4	['Check_In_Month']	300	20	2	gini	nan	[[0 338] [0 482]]
5	['Nights']	300	20	2	gini	0.15	[[58 280] [66 416]]
6	['Country', 'Room_Type']	300	20	100	gini	0.09	[[23 315] [18 464]]
7	['Country', 'Travel_Type']	300	20	2	gini	0.12	[[30 308] [22 460]]
8	['Country', 'Check_In_Month']	300	20	2	gini	0.23	[[76 262] [74 408]]
9	['Country', 'Nights']	300	20	2	gini	0.19	[[63 275] [63 419]]
10	['Room_Type', 'Travel_Type']	300	20	2	gini	0.00	[[1 337] [3 479]]

11	['Room_Type', 'Check_In_Month']	600	20	2	gini	0.36	[[107 231] [93 389]]
12	['Room_Type', 'Nights']	600	20	2	gini	0.09	[[35 303] [39 443]]
13	['Travel_Type', 'Check_In_Month']	1000	20	2	gini	0.21	[[64 274] [57 425]]
14	['Travel_Type', 'Nights']	300	20	2	gini	0.13	[[37 301] [32 450]]
15	['Check_In_Month', 'Nights']	300	20	2	gini	0.14	[[55 283] [62 420]]
16	['Country', 'Room_Type', 'Travel_Type']	1000	20	10	gini	0.14	[[43 295] [40 442]]
17	['Country', 'Room_Type', 'Check_In_Month']	600	20	2	entropy	0.39	[[118 220] [106 376]]
18	['Country', 'Room_Type', 'Nights']	300	20	100	gini	0.22	[[78 260] [82 400]]
19	['Country', 'Travel_Type', 'Check_In_Month']	1000	20	2	gini	0.31	[[113 225] [121 361]]
20	['Country', 'Travel_Type', 'Nights']	600	20	2	gini	0.20	[[56 282] [46 436]]

21	['Country', 'Check_In_Month', 'Nights']	600	20	2	gini	0.22	[[75 263] [75 407]]
22	['Room_Type', 'Travel_Type', 'Check_In_Month']	1000	20	10	entropy	0.34	[[127 211] [140 342]]
23	['Room_Type', 'Travel_Type', 'Nights']	1000	20	10	gini	0.16	[[58 280] [61 421]]
24	['Room_Type', 'Check_In_Month', 'Nights']	300	20	2	gini	0.37	[[122 216] [120 362]]
25	['Travel_Type', 'Check_In_Month', 'Nights']	300	20	10	entropy	0.36	[[113 225] [104 378]]
26	['Country', 'Room_Type', 'Travel_Type', 'Check_In_Month']	1000	50	2	gini	0.34	[[121 217] [128 354]]
27	['Country', 'Room_Type', 'Travel_Type', 'Nights']	600	20	2	gini	0.21	[[73 265] [76 406]]
28	['Country', 'Room_Type', 'Check_In_Month', 'Nights']	1000	20	10	entropy	0.44	[[134 204] [122 360]]
29	['Country', 'Travel_Type',	1000	20	2	entropy	0.34	[[124 214] [132 350]]

	'Check_In_Month', 'Nights']						
30	['Room_Type', 'Travel_Type', 'Check_In_Month', 'Nights']	600	20	2	gini	0.45	[[146 192] [141 341]]
31	['Country', 'Room_Type', 'Travel_Type', 'Check_In_Month', 'Nights']	1000	20	10	Entropy	0.35	[[127 211] [136 346]]

Based on the results presented in Table 14, the best performing Random Forest model achieves a TN_score of 0.45 when trained with the following hyperparameters: {'criterion': 'gini', 'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 600}. Although this model predicts 7 more true negative instances than the best performing Logistic Regression model, it also predicts much more false negative instances. Specifically, the confusion matrix shows that out of a total of 338 actual negative instances, only 43% are correctly classified as negative, while 57% are incorrectly predicted as positive instances. While this performance is an improvement over the best Logistic Regression model, the best Random Forest model still predicts a higher number of false negative instances (141), which constitutes 29% of actual positive instances. As a consequence, the hotel would need to pay closer attention to predicted-negative cases that are actually positive, which could result in unnecessary costs or time spent on addressing issues that do not significantly impact the overall rating of the hotel.

5.3.3 XGBoost

Given the research objective of predicting future reservations’ rating categories, another algorithm which is XGBoost is tested due to its popularity in predictive modelling tasks. XGBoost is a gradient boosting algorithm that iteratively builds decision trees to optimize the prediction accuracy of the model. Compared to other machine learning algorithms, XGBoost is known for its fast computation speed and high accuracy.

In a manner similar to the aforementioned logistic regression analysis, 31 XGBoost models based on the 31 possible feature subsets are constructed. Furthermore, to improve the model's performance, a range of hyperparameters are optimized including “max_depth”, “n_estimators”, “learning_rate”, “subsample”, “colsample_bytree”, and “objective”. The various values utilized for the optimization of the aforementioned hyperparameters are presented in Table 15. The hyperparameter “max_depth” is the maximum depth of the tree and controls the complexity of the tree model. The hyperparameter “n_estimators” is the number of trees in the model. Increasing this hyperparameter can improve the model's performance, but also increase the training time. The hyperparameter “learning_rate” is the step size shrinkage used to prevent overfitting and scales the contribution of each tree in the model. A smaller learning rate requires more trees to be added to the model, but can result in better performance. The hyperparameter “subsample” is the fraction of observations used to train each tree and can be used to prevent overfitting by introducing randomness into the model. The hyperparameter “colsample_bytree” is the fraction of features used to train each tree and can be used to prevent overfitting by introducing randomness into the model. Finally, the hyperparameter “objective” is the loss function to be minimized. For this research, 'binary:logistic' is used, which is the binary classification objective for logistic regression.

Table 15 The values tested of each hyperparameter of XGBoost

Hyperparameters	Values tested
max_depth:	[10,20,30]
n_estimators	[100, 200]
learning_rate	[0.2, 0.4, 0.6]
subsample	[0.5, 0.9]
colsample_bytree	[0.8, 1.0]
objective	['binary:logistic']

Table 16 The best results and the optimal hyperparameters of XGBoost models

No.	Feature subset	max_depth	n_estimators	learning_rate	colsample_bytree	subsample	TN_score	Confusion matrix
1	['Country']	10	100	0.2	0.8	0.5	0.12	[[30 307] [22 461]]
2	['Room_Type']	10	100	0.2	0.8	0.5	Nan	[[0 337] [0 483]]
3	['Travel_Type']	10	100	0.2	0.8	0.5	Nan	[[0 337] [0 483]]
4	['Check_In_Month']	10	100	0.2	0.8	0.5	Nan	[[0 337] [0 483]]
5	['Nights']	10	100	0.2	0.8	0.5	0.14	[[57 280] [67 416]]
6	['Country', 'Room_Type']	10	100	0.4	1.0	0.5	0.09	[[23 314] [18 465]]
7	['Country', 'Travel_Type']	10	100	0.4	1.0	0.5	0.12	[[30 307] [22 461]]
8	['Country', 'Check_In_Month']	10	100	0.6	0.8	0.5	0.27	[[92 245] [92 391]]
9	['Country', 'Nights']	10	200	0.6	1.0	0.9	0.19	[[63 274] [63 420]]
10	['Room_Type', 'Travel_Type']	10	100	0.6	0.8	0.9	0.001	[[1 336] [4 479]]
11	['Room_Type', 'Check_In_Month']	20	200	0.6	0.8	0.5	0.34	[[112 225] [108 375]]
12	['Room_Type', 'Nights']	10	200	0.6	1.0	0.9	0.10	[[38 299] [41 442]]
13	['Travel_Type', 'Check_In_Month']	10	100	0.6	0.8	0.5	0.35	[[93 244] [73 410]]
14	['Travel_Type', 'Nights']	10	200	0.6	1.0	0.5	0.14	[[51 286] [57 426]]

15	['Check_In_Month', 'Nights']	20	100	0.6	1.0	0.5	0.19	[[57 280] [51 432]]
16	['Country', 'Room_Type', 'Travel_Type']	10	100	0.2	1.0	0.5	0.09	[[26 311] [22 461]]
17	['Country', 'Room_Type', 'Check_In_Month']	10	100	0.2	0.8	0.5	0.29	[[90 247] [82 401]]
18	['Country', 'Room_Type', 'Nights']	10	200	0.4	1.0	0.9	0.22	[[78 259] [83 400]]
19	['Country', 'Travel_Type', 'Check_In_Month']	20	100	0.4	1.0	0.5	0.35	[[124 213] [130 353]]
20	['Country', 'Travel_Type', 'Nights']	10	100	0.6	0.8	0.9	0.18	[[49 288] [39 444]]
21	['Country', 'Check_In_Month', 'Nights']	10	100	0.6	1.0	0.5	0.26	[[94 243] [100 383]]
22	['Room_Type', 'Travel_Type', 'Check_In_Month']	20	200	0.4	1.0	0.5	0.31	[[124 213] [145 338]]
23	['Room_Type', 'Travel_Type', 'Nights']	10	100	0.6	1.0	0.9	0.16	[[59 278] [65 418]]
24	['Room_Type', 'Check_In_Month', 'Nights']	10	200	0.6	1.0	0.5	0.33	[[125 212] [142 341]]

25	['Travel_Type', 'Check_In_Month', 'Nights']	20	100	0.6	0.8	0.5	0.33	[[119 218] [128 355]]
26	['Country', 'Room_Type', 'Travel_Type', 'Check_In_Month']	10	100	0.4	1.0	0.5	0.31	[[120 217] [140 343]]
27	['Country', 'Room_Type', 'Travel_Type', 'Nights']	10	100	0.6	1.0	0.5	0.35	[[120 217] [125 358]]
28	['Country', 'Room_Type', 'Check_In_Month', 'Nights']	10	100	0.6	1.0	0.9	0.44	[[135 202] [124 359]]
29	['Country', 'Travel_Type', 'Check_In_Month', 'Nights']	10	100	0.6	0.8	0.5	0.31	[[121 216] [139 344]]
30	['Room_Type', 'Travel_Type', 'Check_In_Month', 'Nights']	10	200	0.4	1.0	0.9	0.41	[[138 199] [138 345]]
31	['Country', 'Room_Type', 'Travel_Type', 'Check_In_Month', 'Nights']	20	100	0.6	1.0	0.5	0.41	[[148 189] [160 323]]

Analysis of the results presented in Table 16 indicates that the 28th model, with the feature subset of *Country*, *Room_Type*, *Check_In_Month, Nights* and the optimized hyperparameters {'colsample_bytree': 1.0, 'learning_rate': 0.6, 'max_depth': 10, 'n_estimators': 100, 'objective': 'binary:logistic', 'subsample': 0.9}, achieves the best performance, with a TN_score of 0.44. The model accurately predictes 135 true negative instances out of a total of 337 actual negative instances, resulting in a true negative rate of 40%. However, the model also classifies 124 positive instances as negative, meaning that 48% of all negative predictions are actually positive instances. This suggests that the hotel should exercise caution when interpreting the model's predictions, as nearly half proportion of the predicted negative cases may not necessarily lead to a decrease in the overall rating.

Another important aspect to consider is the performance of the 31st model, which incorporated a feature subset comprising *Country, Room_Type, Travel_Type, Check_In_Month, and Nights*. This model yields a higher number of true negative predictions compared to the 28th model, correctly predicting 13 more negative instances. However, this improvement was accompanied by a substantial increase in false negative predictions, with 36 more positive instances incorrectly classified as negative. Similar to the issue encountered in the best performing random forest model, an excessive number of false negative predictions can result in unnecessary costs or efforts spent in addressing issues that may not significantly affect the overall rating of the hotel.

5.3.4 ANN

The artificial neural network (ANN) is a powerful machine learning technique that has been successfully used in various applications, including the prediction in the hotel industry. The advantage of using ANN models for the prediction of the hotel rating categories is that they can capture complex, nonlinear relationships between the input features and the output variable (i.e., *rating_cat*).

In this research, ANN models are utilized to predict the hotel’s rating categories based on the 31 feasible subsets of the five features. Consequently, 31 ANN models are developed and evaluated, akin to the earlier logistic regression and XGBoost models. This section presents and compares the results of 31 artificial neural network (ANN) models, each optimized for a variety of hyperparameters, such as batch size, number of epochs, number of hidden layers, and number of nodes per layer. To illustrate, when training an ANN model, various hyperparameters need to be set to optimize the model's performance. Four of the most important hyperparameters are batch size, number of epochs, number of hidden layers, and number of nodes per layer. Batch size refers to the number of samples that are processed by the model at once during training. Larger batch sizes may lead to faster training times, but may also lead to less accurate results. The number of epochs, on the other hand, refers to the number of times the entire training dataset is presented to the model during training. Increasing the number of epochs may improve the model's accuracy, but may also increase the risk of overfitting, where the model performs well on the training data but poorly on new data. The number of hidden layers and nodes per layer determine the complexity and capacity of the model. Increasing the number of layers and nodes may improve the model's ability to capture complex patterns in the data, but may also increase the risk of overfitting and result in slower training times. Therefore, careful tuning of these hyperparameters is crucial to achieving optimal performance of an ANN model.

In Table 17, the values used for optimizing the hyperparameters of the ANN models are listed. The performance of all 31 ANN models is presented in Table 18, where an analysis of their respective results allowed the identification of the best performing model as well as the optimal combination of hyperparameters for each model.

Table 17 The values tested of each hyperparameter of ANN

Hyperparameters	Values tested
batch_size	[30, 50]
epochs	[100, 200]
num_hidden_layers	[2, 6]
num_nodes	[2, 6]

Table 18 The best results and the optimal hyperparameters of ANN models

No.	Feature subset	batch_size	epochs	num_hidden_layers	num_nodes	TN_score	Confusion matrix
1	['Country']	30	100	2	6	0.12	[[30 307] [22 461]]
2	['Room_Type']	30	100	2	2	Nan	[[0 338] [0 482]]
3	['Travel_Type']	30	100	2	2	Nan	[[0 337] [0 483]]
4	['Check_In_Month']	30	100	2	2	Nan	[[0 337] [0 483]]
5	['Nights']	30	100	6	6	0.15	[[58 280] [66 416]]
6	['Country', 'Room_Type']	30	100	2	6	0.09	[[23 315] [18 464]]
7	['Country', 'Travel_Type']	50	100	6	6	0.12	[[30 308] [22 460]]
8	['Country', 'Check_In_Month']	30	200	6	6	0.26	[[88 250] [87 395]]
9	['Country', 'Nights']	30	100	6	6	Nan	[[0 338] [0 482]]
10	['Room_Type', 'Travel_Type']	30	100	2	2	Nan	[[0 338] [0 482]]
11	['Room_Type', 'Check_In_Month']	50	200	6	6	0.36	[[104 234] [88 394]]
12	['Room_Type', 'Nights']	30	100	6	6	0.11	[[39 299] [40 442]]

13	['Travel_Type', 'Check_In_Month']	50	100	2	6	0.23	[[70 268] [64 418]]
14	['Travel_Type', 'Nights']	50	100	6	6	0.13	[[37 301] [32 450]]
15	['Check_In_Month', 'Nights']	30	200	2	6	0.13	[[50 288] [58 424]]
16	['Country', 'Room_Type', 'Travel_Type']	50	100	6	6	0.13	[[42 296] [39 443]]
17	['Country', 'Room_Type', 'Check_In_Month']	30	100	6	6	0.39	[[115 223] [101 381]]
18	['Country', 'Room_Type', 'Nights']	30	100	2	6	0.23	[[82 256] [87 395]]
19	['Country', 'Travel_Type', 'Check_In_Month']	50	200	6	6	0.44	[[123 215] [102 380]]
20	['Country', 'Travel_Type', 'Nights']	30	200	6	6	0.21	[[68 270] [64 418]]
21	['Country', 'Check_In_Month', 'Nights']	30	200	2	6	0.23	[[75 263] [72 410]]
22	['Room_Type', 'Travel_Type', 'Check_In_Month']	30	100	6	6	0.31	[[106 232] [107 375]]

23	['Room_Type', 'Travel_Type', 'Nights']	30	100	6	6	0.14	[[48 290] [49 433]]
24	['Room_Type', 'Check_In_Month', 'Nights']	50	100	2	2	Nan	[[0 338] [0 482]]
25	['Travel_Type', 'Check_In_Month', 'Nights']	30	100	6	6	Nan	[[0 338] [0 482]]
26	['Country', 'Room_Type', 'Travel_Type', 'Check_In_Month']	30	200	6	6	Nan	[[146 192] [158 324]]
27	['Country', 'Room_Type', 'Travel_Type', 'Nights']	50	100	6	6	0.24	[[82 256] [82 400]]
28	['Country', 'Room_Type', 'Check_In_Month', 'Nights']	50	100	6	6	0.48	[[136 202] [115 367]]
29	['Country', 'Travel_Type', 'Check_In_Month', 'Nights']	50	100	6	6	0.45	[[125 213] [103 379]]
30	['Room_Type', 'Travel_Type', 'Check_In_Month', 'Nights']	30	100	6	6	0.55	[[204 134] [224 258]]
31	['Country', 'Room_Type', 'Travel_Type',	30	200	2	6	0.37	[[151 187] [183 299]]

	'Check_In_Month', 'Nights']						

Based on the results presented in Table 18, the 30th model that utilizes the feature subset of *Room_Type, Travel_Type, Check_In_Month, Nights* and the optimized hyperparameters {' batch_size ': 30, ' epochs ': 100, ' num_hidden_layers ': 6, ' num_nodes ': 6} demonstrates best performance, achieving a TN_score of 0.55. The aforementioned model effectively predicted 204 true negative instances out of a total of 338 actual negative instances, which results in a true negative rate of 60%. However, the model's classification of 224 positive instances as negative implies that 52% of all negative predictions correspond to positive instances. Consequently, the hotel should also exercise caution when interpreting the model's predictions, as more than half of the predicted negative cases may not necessarily result in a decrease in the overall rating.

5.3.5 Comparison of different models

In this research, four distinct classification algorithms are employed to predict the rating categories (*rating_cat*) for the hotel. Specifically, a total of 31 different models are constructed for each of the four classification algorithms using unique subsets of features, and distinct hyperparameters are also optimized for each model. The results of the evaluation are presented in the tables above. Overall, the models generated by the ANN algorithm exhibit superior performance in terms of TN_score, as demonstrated by Figure 21. To be more specific, among all models produced by the four algorithms, the 30th model generated by the ANN algorithm attains the highest TN_score of 0.55, surpassing the performance of all other models generated by the four algorithms. This particular model is constructed using a feature subset comprising of *Room_Type, Travel_Type, Check_In_Month, Nights*, and optimized hyperparameters that included 'batch_size' of 30, 'epochs' of 100, 'num_hidden_layers' of 6, and 'num_nodes' of 6.

However, the results of this research indicate that an increase in the number of correctly predicted true negative instances is accompanied by an increase in the number of false negative instances in all models. Careful consideration of the relationship between the number of true negative instances and false negative instances is crucial when utilizing these models for predicting rating categories of the hotel. The optimal strategy for the hotel is to enhance the number of true negative instances while minimizing the number of false negative instances to improve the prediction accuracy. Consequently, the processing of the least negative cases can be conducted with minimal cost, leading to an improvement in the hotel's overall rating in the future.

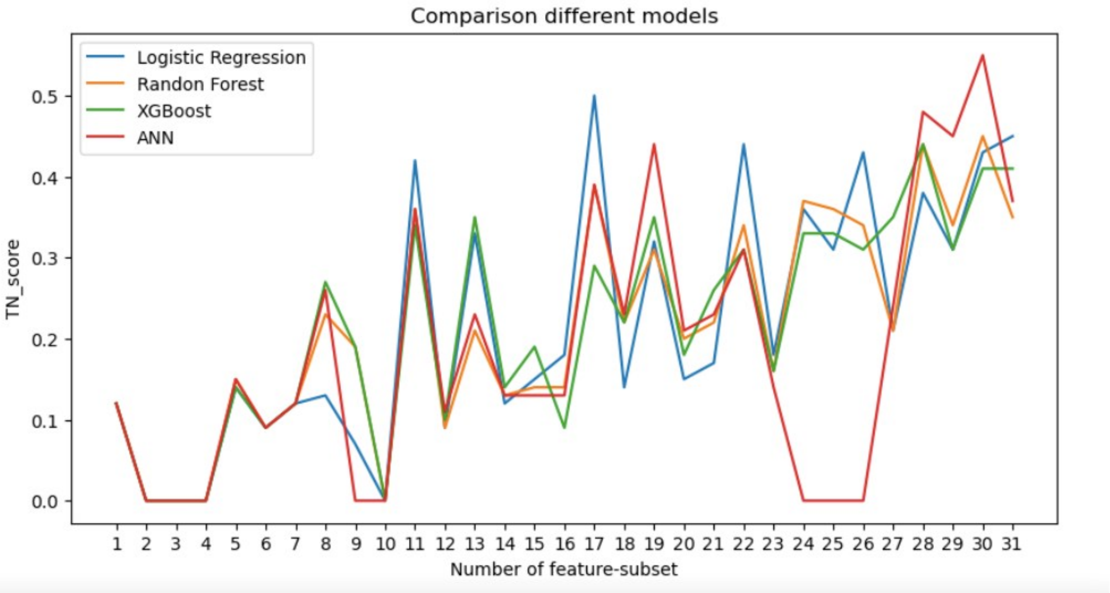


Figure 21 comparison of different models

5.4 Textual data analysis

In order to gain insights for managers of the hotel into the factors that contribute to guest satisfaction, a textual analysis of *Review_Details* column in the dataset is conducted.

5.4.1 Check and translate the languages of all reviews.

Initially, the languages employed in all reviews are subjected to verification by means of the "langdetect" library, a tool in Python specifically designed for language detection. This library operates through a probabilistic language detection algorithm, utilizing the n-gram approach, and supports 55 languages (langdetect, 2023). After the verification process, a total of 33 languages that are employed in the reviews are determined. In order to ensure consistency and accuracy in the analysis, as well as to facilitate a comprehensive understanding of the data under scrutiny, the text in other languages employed in the reviews are translated to English prior to undertaking any form of textual analysis. This step is deemed essential given the research's focus on textual data analysis in the English language.

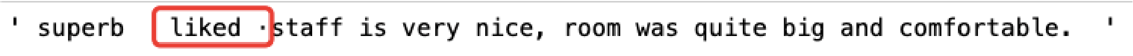
The process of translating the texts in other languages used in the reviews to English is executed by using “deep_translator”, a Python library that provides a free and accessible method for language translation. With “deep_translator”, the translation process can be carried out automatically, saving time and resources that would otherwise have been expended on manual translation. The translated texts are then subjected to the same analytical procedures as the English texts, facilitating a comprehensive and unified approach to textual data analysis.

5.4.2 Textual data transformation

To enable further textual data analysis, certain transformations must be performed on the data. The first transformation involves converting all translated English reviews to lowercase. This approach can significantly enhance the efficiency and accuracy of the analysis by standardizing the text and eliminating the distinction between uppercase and lowercase forms of words. By doing so, the text can be analysed more consistently, facilitating the extraction of meaningful insights and patterns from the data. In addition, the lowercase transformation simplifies the task of matching words during the analysis process, promoting the identification of relevant terms and phrases within the text. Ultimately, this transformation can enhance the quality and relevance of textual data analyses. The second transformation entails performing lemmatization on the text after its conversion to lowercase. Lemmatization involves reducing words to their base or root form, enabling the identification of common word forms and the elimination of variations that could potentially distort analysis results. By performing lemmatization after the conversion to lowercase, the analysis is immune to case-based variations in word forms, ensuring that the focus is solely on the fundamental meaning of words in the text. This approach can yield more accurate and insightful results, enhancing the overall quality and relevance of the textual data analysis.

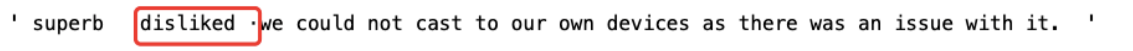
5.4.3 Reviews categories.

Upon conducting a more in-depth analysis, specific terms, namely "liked" and "disliked", are found to be prevalent in the reviews under scrutiny. This observation is corroborated by the graphical representations depicted in Figures 22 and 23, which provide illustrative examples of the aforementioned terms as they occur in the analysed reviews.



```
' superb   liked ·staff is very nice, room was quite big and comfortable.   '
```

Figure 22 an example of key word "like" in the reviews



```
' superb   disliked ·we could not cast to our own devices as there was an issue with it.   '
```

Figure 23 an example of key word "dislike" in the reviews

Following a thorough investigation, the investigation reveals that 2359 reviews express a preference with "liked" terms, whereas 165 reviews convey a negative sentiment with "disliked" terms. The remaining 2941 reviews, devoid of either "liked" or "disliked" terms, are categorized as "unknown" in this research. Further inspection discloses that within the subset of 2941 "unknown" reviews, seven contain the message "this review has been hidden because it doesn't meet our guidelines," while the remaining 2934 reviews feature the message "there are no comments available for this review". Given that the messages found in the subset of 2941 "unknown" reviews do not provide meaningful insights for the textual data analysis, focus is redirected to the 2359 reviews displaying "liked" terms and the 165 reviews featuring "disliked" terms, respectively.

5.4.4 Stopwords removal and most frequent words generation

To facilitate accurate textual data analysis, distinct stopwords removal procedures were applied to reviews featuring "liked" and "disliked" terms. Following this, the 25 most frequently occurring words in each category, along with their corresponding frequencies and proportion of all reviews, were computed and depicted in Table 19 and 20, respectively.

As illustrated in Table 19 and 20, a number of words frequently appear in the textual data, yet lack relevance to the goal of the textual data analysis, which is to provide insights for the hotel managers. These common words, including "like", "dislike", "hotel", "get", "one", and "great" can be considered as stopwords for reviews containing "liked" and "disliked" terms, respectively. Therefore, these words should also be removed in order to get better insights. Subsequently, a secondary round of stopwords removal is conducted, resulting in the final top 10 most frequent words for reviews featuring “liked” and “disliked” terms, as depicted in Table 21 and 22.

As shown in Table 21 and 22, the top 10 most frequent words for reviews with “liked” terms are “breakfast”, “room”, “location”, “staff”, “bed”, “comfortable”, “clean”, “friendly”, “view”, and “service”, which provide insights into the most appreciated aspects of the hotel by guests. The top 10 most frequent words for reviews with “disliked” terms were “room”, “breakfast”, “bed”, “bathroom”, “reception”, “clean”, “stay”, “water”, “long”, and “staff”, which suggest that issues related to the room, breakfast, and staff may be the main reasons for dissatisfaction among the guests. Some of the words that appear in the reviews with "liked" terms and reviews with "disliked" terms have overlapped, which is worth noting, such as “room”, “breakfast”, and “bed”. For example, based on the frequency, the words “room” and “breakfast” are mentioned much more frequently in the reviews with “liked” terms compared to those reviews with “disliked” terms. This may suggest that guests generally have positive experiences with the room and breakfast at the hotel, however, some guests may not have derived as much enjoyment from these aspects.

Table 19 The top 25 most frequent words and their frequency and proportion of reviews with “liked” terms

Keywords	**Frequency**	**Proportion**
Like	2421	44.31%
Breakfast	1375	25.16%
Good	1170	21.41%
Room	1125	20.59%
Location	657	12.02%
Hotel	564	10.32%
Staff	553	10.12%

Great	521	9.54%
Bed	484	8.86%
Nice	461	8.44%
Comfortable	383	7.01%
Clean	344	6.30%
Friendly	308	5.64%
Really	294	5.38%
Excellent	251	4.60%
View	205	3.75%
Service	190	3.48%
Well	173	3.17%
Everything	167	3.06%
Big	166	3.04%
Beds	156	2.86%
Spacious	151	2.76%
Stay	143	2.62%
Also	136	2.49%
Helpful	129	2.36%

Table 20 The top 25 most frequent words and their frequency and proportion of reviews with "disliked" terms

Keywords	Frequency	Proportion
Disliked	165	3.02%
Room	102	1.87%
Hotel	37	0.66%
Breakfast	36	0.66%
Good	25	0.46%
Get	25	0.46%
One	19	0.35%
Bed	19	0.35%
Book	17	0.31%
Time	17	0.31%
Reception	16	0.29%
Day	16	0.29%
Bad	16	0.29%
Bathroom	16	0.29%
Even	16	0.29%
Could	15	0.27%
Check	15	0.27%
Clean	15	0.27%
Stay	14	0.26%
Also	12	0.22%

Staff	12	0.22%
Water	12	0.22%
Long	12	0.22%
Night	12	0.22%
Ask	12	0.22%

Table 21 The top 10 most frequent words and their frequency and proportion of reviews with "liked" terms after remove meaningless words

Keywords	**Frequency**	**Proportion**
Breakfast	1375	25.16%
Room	1125	20.59%
Location	657	12.02%
Staff	553	10.12%
Bed	484	8.86%
Comfortable	383	7.01%
Clean	344	6.30%
Friendly	308	5.64%
View	205	3.75%
Service	190	3.48%

Table 22 The top 10 most frequent words and their frequency and proportion of reviews with "disliked" terms after remove meaningless words

Keywords	**Frequency**	**Proportion**

Room	102	1.87%
Breakfast	36	0.66%
Bed	19	0.35%
Bathroom	16	0.29%
Reception	16	0.29%
Clean	15	0.27%
Stay	14	0.26%
Water	12	0.22%
Long	12	0.22%
Staff	12	0.22%

6 DISCUSSION AND LIMITATION

The aim of this research is to develop a predictive model for the future rating categories of the Radisson Blu Seaside Hotel in Helsinki, based on publicly available data from the guest review section of Booking.com. Specifically, the study seeks to utilize the data provided by guests during the reservation process to predict their future rating categories for their reservations. Additionally, the research aims to identify the most frequently used negative textual words in the reviews of the hotel, which may serve as a factor contributing to negative ratings.

By leveraging the data provided by guests during the reservation process, such as their country of origin and intended check-in month, the hotel can employ predictive analytics to forecast the guests’ future rating category of the accommodation experience. In the event of a prediction indicating a probable negative rating, the hotel can proactively take measures to enhance the guests’ satisfaction prior to their arrival, thereby mitigating the risk of receiving unfavourable feedback after their stay. This approach not only allows the hotel to improve their overall ratings on Booking.com, but also fosters greater guest satisfaction and loyalty, ultimately leading to increased competitiveness within the marketplace. Therefore, by achieving these objectives, the research aims to provide useful insights for the hotel management to improve the hotel’s online reputation and bring more new reservations to improve the hotel’s profitability.

This study utilized publicly available data to train the machine learning models for predicting reservations' rating category based on the information provided by guests during the reservation process. However, it should be noted that there are several limitations to the study. Firstly, the data used for this research are solely obtained from public sources, and therefore, it may not represent the full scope of information that can be gathered by the hotel from their guests when making reservations. Access to additional data, such as passport or payment data, could potentially improve the accuracy of the models.

Furthermore, predicting customer behaviour remains a challenging issue even with the use of advanced machine learning techniques. Predicting customer behaviour is a challenging task due to several reasons, such as personal preferences, cultural background, and economic conditions. Moreover, customer behaviour can be highly dynamic, which further complicates the task of predicting it accurately. The present research’s findings align with the aforementioned limitation, providing partial explanation for the observed results. Specifically, increasing the number of predicted true negative instances does not lead to an improvement in model accuracy, as the simultaneous increase in false negative instances counteracts this effect. Thus, when increasing the number of predicted true negative instances, the model tends to produce more negative instances, thereby including a considerable number of actual positive instances. Therefore, future research should consider incorporating additional data sources, exploring new modelling techniques and deep research into customer behaviour to further improve the accuracy of predicting reservations' rating category.

In order to achieve the objectives of this research, three research questions have been formulated. The ensuing sections provide a concise summary and discussion of these research questions and their corresponding answers:

RQ1: What topics have been covered in literatures on using machine learning to make predictions in the hotel industry?

In order to address this question, a comprehensive review of ten distinct literature pieces published between the years 2015 and 2022, which explore the utilization of machine learning techniques for the purpose of forecasting research within the hotel industry, has been conducted.

The literature review reveals that previous literatures have focused on seven topics in the forecasting research, including predicting the energy consumption in a hotel room, predicting the hotel cancellations, forecasting hotel occupancy, forecasting hotel demand, forecasting hotel room price, and predicting sentiment and rating of tourist reviews. A wide range of algorithms have been utilized in the literatures to address forecasting research in the hotel industry. Some popular algorithms include, but are not limited to, SVM, XGBoost, and ANN.

However, despite the abundance of literature on the application of machine learning in the hotel industry for forecasting purposes, none have specifically addressed the topic of forecasting the likelihood of negative guest ratings prior to their arrival. This research addresses this gap in the literature by proposing a novel approach to predicting negative ratings through the integration of machine learning techniques and the data guests provide at the time of reservation. The aim is to try to provide hotels with an effective means of predicting negative ratings and possible negative reviews, enabling them to take appropriate actions to mitigate any potential issues and ultimately improve the hotel's online reputation by enhancing the guest experience.

RQ2: Which machine learning models offer the best performance in predicting customer's (numeric) evaluations?

To enable a comprehensive comparison of the models' performance across all four algorithms, a novel metric, TN_score, has been introduced. The primary objective of this metric is to facilitate the identification of negative-rated reservations with high accuracy and minimal time and cost. Accordingly, the hotel emphasizes the importance of correctly identifying true negative instances while minimizing the occurrence of false negative instances. The optimization of the prediction model is therefore focused on maximizing the identification of negative-rated reservations while minimizing the number of false negative predictions. False negative predictions refer to instances where positive-rated reservations are misclassified as negative-rated. By prioritizing the accuracy of negative-rated reservations and the minimization of the false negative predictions, the hotel aims to enhance the overall predictive accuracy of the model and subsequently improve service quality and customer satisfaction by dealing with all negative predictions using the minimal time and cost.

Out of the 124 models generated from the four algorithms under consideration, the 30th ANN model with hyperparameters tuned to {'batch_size': 30, 'epochs': 100, 'num_hidden_layers': 6, 'num_nodes': 6}, exhibites the highest TN_score value of 0.55. This model is designed with a specific set of features consisting of *Room_Type, Travel_Type, Check_In_Month, and Nights*. As per the confusion matrix, the model performs with relatively high precision when identifying instances that are actually negative, with a precision rate of 60% among the 338 negative instances. However, the model displays a lower precision rate of 48% when it comes to identifying instances that are positive but are predicted as negative, leading to a considerable number of false negative predictions. This implies that the hotel ought to pay attention to a significant number of reservations predicted to be negative-rating, even though only a small subset of these reservations will have a positive impact on improving the hotel's overall rating. As a consequence, the return on investment of the hotel's efforts including time and other costs to deal with possible negative-rating reservations may be below 50%, indicating that the efficacy of such efforts may be limited.

This result relatively affirms the widely held view that predicting consumer behaviour presents a significant challenge. The complexity of personal preferences, cultural background, and economic conditions makes it difficult to develop accurate predictions of consumer behaviour. Additionally, the dynamic nature of customer behaviour further complicates this task, as it can be influenced by changing market trends, evolving consumer preferences, and technological advancements. Given the subjectivity of individual decision-making processes, which are influenced by a range of psychological, social, and environmental factors, predicting consumer behaviour remains a significant challenge for businesses. Despite these difficulties, the accurate prediction of consumer behaviour is essential for developing effective marketing strategies, enhancing customer satisfaction and loyalty, and driving profitability.

In order to enhance the accuracy of predicting reservations' rating categories, future research may explore the potential benefits of collaborating with relevant platforms, such as Booking.com. This collaboration could involve integrating user data from the platform, analysing the historical data of individual users, and incorporating these data as predictive variables in the predictive models. Such an approach may provide valuable insights and contribute to a more comprehensive understanding of factors affecting reservation ratings. Nonetheless, it is important to acknowledge that such an approach may raise concerns related to privacy and data protection. As a result, a coordinated effort between hotels and the online booking platforms would be necessary to address these potential issues and arrive at a mutually acceptable solution.

RQ3: What insights can be gained for hotel managers from customers' (text) evaluation?

The findings obtained from the analysis of textual data offer significant implications for the hotel managers concerning customers' text evaluations. Specifically, the results revealed that breakfast, room, location, staff, and bed are the most frequently mentioned aspects in the reviews with the “liked” term. These findings imply that the hotel should give priority to these areas in their service offerings and endeavour to uphold a high level of quality to ensure customer satisfaction and loyalty.

In contrast, the most commonly mentioned aspects in the reviews with the "disliked" term are room, breakfast, and bed, which partially overlap with the most frequently mentioned aspects in the reviews with the "liked" term. This may be because while these overlapped areas are frequently mentioned in the reviews with “liked” term, it does not necessarily indicate their flawlessness. While most guests may enjoy these areas overall, some other guests may still find room for improvement. Moreover, guest evaluations of the hotel amenities and services may vary based on their diverse expectations and preferences.

Overall, the textual data analysis suggests that the hotel managers can gain valuable insights from customers' (text) evaluations, and use this information to identify areas of improvement in their service offerings. More importantly, in cases where the predicted rating category is negative, the insights gained from customers' text evaluations can provide more precise guidance for the hotel managers. This enables the managers to address potential issues with a higher probability of success before guests arrive by focusing on these areas, leading to enhanced customer satisfaction and improved online reputation for the hotel. Specifically, the hotel managers that hotel managers should prioritize the quality and cleanliness of their rooms and the comfort of their bedding. Additionally, the managers may need to reconsider their breakfast options by offering a more diverse selection or enhancing the quality of the food provided.

By prioritizing these areas in their service offerings and ensuring their quality, the hotel can not only improve the satisfaction of their current guests but also attract new ones. Additionally, by paying attention to the specific aspects that guests enjoy, the hotel can tailor their services to meet the needs and expectations of their target audience. This can lead to higher customer satisfaction and loyalty, as well as positive WOM recommendations and online reviews, ultimately benefiting the hotel's reputation and contributing to attract new guests.

7 CONCLUSION

The aim of the thesis is to employ machine learning and data analysis techniques to predict the rating categories of forthcoming reservations for a hotel. Additionally, this research intends to utilize textual data analysis to extract insights that would assist the hotel management in dealing with reservations whose predicted ratings fall below the hotel's overall rating. This approach is expected to improve the hotel's online reputation and enhance its profitability.

After considering various data sources for this research, it is determined that public data from the review section of Booking.com is the most suitable due to the availability of a substantial number of predictor variables that could be used to construct machine learning models. Consequently, the Radisson Blu Seaside Hotel in Helsinki is selected as the focal hotel for this study, based on the availability of a significant volume of data for training and testing the machine learning models. To ensure that the data is in a machine-readable state and to allow subsequent models to have better performance, corresponding data pre-processing is carried out. This involved various data cleaning and transforming techniques, such as removing missing values, numeric categorical values and categoric numerical values. Four machine learning algorithms, including logistic regression, random forest, XGBoost, and ANN, are chosen to build the models for predicting the rating categories of future reservations. These algorithms are selected based on their popularity, accuracy, and reliability in the field of machine learning for forecasting research. Logistic regression is a widely used linear model that is known for its interpretability, while random forest and XGBoost are ensemble methods that can handle complex interactions between variables. ANN, on the other hand, is a non-linear model that can capture highly complex relationships between variables. Furthermore, a novel evaluation metric, known as TN_score, has been introduced. This metric places emphasis on the proportion of true negative instances in relation to the total number of the actual negative instances, while also taking into account the proportions of false negative instances in the total predicted-negative instances. In this context, a higher TN_score value indicates superior performance.

The ANN model with hyperparameters tuned to {'batch_size': 30, 'epochs': 100, 'num_hidden_layers': 6, 'num_nodes': 6}, exhibites the highest TN_score value of 0.55.

This model is designed with a specific set of features consisting of *Room_Type, Travel_Type, Check_In_Month, and Nights*. As per the confusion matrix, the model performs with relatively high precision when identifying instances that are actually negative, with a precision rate of 60% among the 338 negative instances. Nonetheless, the model yields a notable number of false negative predictions, whereby 224 instances are erroneously classified as negative despite their actual positive status among the total of 482 positive instances.

The findings of all models demonstrate that increasing the number of predicted true negative instances results in a simultaneous increase in the number of false negative instances. Hotels are adverse to this situation as it diminishes the efficiency of hotel managements, leading to higher costs incurred in addressing the greater number of cases and enhancing the overall rating of the hotel. The unpredictability of consumer behaviour stemming from various factors is a salient reason underlying this situation. Subsequently, future research could examine mitigating this issue by collaborating with multiple booking platforms to develop a database that incorporates the historical data of each guest. The analysis of such data could contribute to enhancing the comprehension of customer behaviour within a limited range. The amalgamation of such guest data and PNR data would result in an improvement in the models' performance upon analysis.

In addition to machine learning models for forecasting research, textual data analysis is also performed in this research to generate insights that would assist the hotel managers in addressing reservations whose predicted ratings fall below the overall rating of the hotel. To begin with, a total of 33 languages that are used in the reviews are identified, and reviews in languages other than English are translated accordingly to enable the analysis of the text data. Next, lemmatization was performed to transform the text data into a standard form, and stop words in English and meaningless words in the context of reviews are removed to improve the accuracy of the analysis. Stop words and meaningless words refer to common words in English and in this context that do not add any meaning to the text, such as “the”, “is”, and “hotel”. Finally, the most frequent words that are associated with positive and negative sentiments towards the hotel are identified. Specifically, the top 10 frequent words that are linked to the term "liked" and the top 10 frequent words that are related to the term "disliked" are generated. These insights could provide the hotel managers with valuable information about the factors that influence guests' perceptions of their stay, and allow them to take appropriate measures to improve their services and the overall guest experience for preventing the happen of negative ratings.

The results of the textual data analysis indicate that breakfast, room, location, staff, and bed are the most commonly mentioned aspects in reviews with a positive sentiment. Conversely, in reviews expressing a negative sentiment, the most frequently mentioned aspects are room, breakfast, and bed, with some overlap with the aspects mentioned in positive reviews. These identified aspects can be considered the primary concerns of guests and, as such, should be the first areas of focus for improving overall ratings. By directing efforts towards these aspects, hotel managers can potentially improve ratings more efficiently and at a reduced cost.

In conclusion, the thesis has introduced a promising research direction in the utilization of machine learning for predictive analysis within the hotel industry. Although the performance of the prediction models presented in this research is not entirely satisfactory, the research has identified potential areas of improvement and provides recommendations for future research to enhance the models' effectiveness. Moreover, this thesis has demonstrated that the key words extracted from guest reviews can be utilized by hotel managers to strategically enhance the overall rating of their hotels. These findings suggest that machine learning has the potential to significantly contribute to the development of the hotel industry and warrants further investigation.

REFERENCES

Al Shehhi, M., & Karathanasopoulos, A. (2020). Forecasting hotel room prices in selected GCC cities using deep learning. *Journal of Hospitality and Tourism Management*, 42, 40-50.

Alotaibi, E. (2020). Application of machine learning in the hotel industry: a critical review. *Journal of Association of Arab Universities for Tourism and Hospitality*, 18(3), 78-96.

Antonio, N., de Almeida, A., & Nunes, L. (2017, December). Predicting hotel bookings cancellation with a machine learning classification model. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1049-1054). IEEE.

Antonio, N., de Almeida, A., & Nunes, L. (2019). An automated machine learning based decision support system to predict hotel booking cancellations. An automated machine learning based decision support system to predict hotel booking cancellations, (1), 1-20.

Azarmi, S. L., Oladipo, A. A., Vaziri, R., & Alipour, H. (2018). Comparative modelling and artificial neural network inspired prediction of waste generation rates of hospitality industry: the case of North Cyprus. Sustainability, 10(9), 2965.

Banerjee, A., Bandyopadhyay, T., & Acharya, P. (2013). Data analytics: Hyped up aspirations or true potential?. Vikalpa, 38(4), 1-12.

Beautiful Soup Documentation. (2023). About Us. Retrieved on January 9, 2023, from https://beautiful-soup-4.readthedocs.io/en/latest/#

Beer, D., & Burrows, R. (2010). Consumption, prosumption and participatory web cultures: An introduction. *Journal of consumer culture*, 10(1), 3-12.

Booking.com. (2023). About Booking.com. Retrieved on January 4, 2023, from https://www.booking.com/content/about.en-gb.html

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

Brown, J., Broderick, A. J., & Lee, N. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of interactive marketing*, 21(3), 2-20.

Brownlee, J. (2020). Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery.

Burger, C. J. S. C., Dohnal, M., Kathrada, M., & Law, R. (2001). A practitioners guide to time-series methods for tourism demand forecasting—a case study of Durban, South Africa. Tourism management, 22(4), 403-409.

Calvin, J. & Hobbes, G. (1999). *The Impossibilities of Life*. Jersey: McGraw-Hull. This second reference has not been used in the text but has been included as an example.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Criscuolo, T. L., Assunção, R. M., Loschi, R. H., Meira Jr, W., & Cruz-Reyes, D. (2023). Handling categorical features with many levels using a product partition model. The Annals of Applied Statistics, 17(1), 786-814.

Diebold, F. X. (2012). On the Origin (s) and Development of the Term'Big Data'.

Elgendy, N., & Elragal, A. (2014, July). Big data analytics: a literature review paper. In Industrial conference on data mining (pp. 214-227). Springer, cham.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), 137-144.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. IDC iview, 1142(2011), 1-12.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27 (pp. 345-359). Springer Berlin Heidelberg.

Gudivada, V. N., Irfan, M. T., Fathi, E., & Rao, D. L. (2016). Cognitive analytics: Going beyond big data analytics and machine learning. In Handbook of statistics (Vol. 35, pp. 169-205). Elsevier.

Gupta, S., & Gupta, A. (2019). Dealing with noise problem in machine learning data-sets: A systematic review. Procedia Computer Science, 161, 466-474.

Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

He Xinying. (2019). A Brief Discussion on the Marketing Advantages and Disadvantages of the Hotel Industry in the Internet Era——Taking Internet Hotels as an Example. *Contemporary Tourism* (1), 1. ([1]何欣颖. (2019). 浅论互联网时代下酒店行业的营销优劣——以互联网酒店为例. 当代旅游(1), 1.)

Hilbe, J. M. (2009). Logistic regression models. CRC press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 79(8), 2554-2558.

Hotels.com. (2023). About us. Retrieved on January 4, 2023, from https://uk.hotels.com/lp/b/about_us

Hu, Q., Yu, D., Liu, J., & Wu, C. (2008). Neighborhood rough set based heterogeneous feature subset selection. *Information sciences*, 178(18), 3577-3594.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. IEEE Transactions on pattern analysis and machine intelligence, 22(1), 4-37.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. Computer, 29(3), 31-44.

Jin, S. V., & Phua, J. (2016). Making reservations online: The impact of consumer-written and system-aggregated user-generated content (UGC) in travel booking websites on consumers' behavioral intentions. *Journal of travel & tourism marketing*, 33(1), 101-117.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. International journal of computer science, 1(2), 111-117.

Król, K., & Zdonek, D. (2020). Analytics maturity models: An overview. Information, 11(3), 142.

Kumar, V., & Garg, M. L. (2018). Predictive analytics: a review of trends and techniques. *International Journal of Computer Applications*, 182(1), 31-37.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70), 1.

langdetect (2023). Project description. Retrieved on April 5, 2023, from https://pypi.org/project/langdetect/

Leal, F., Malheiro, B., & Burguillo, J. C. (2019). Analysis and prediction of hotel ratings from crowdsourced data. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery*, 9(2), e1296.

Li, J., Xu, L., Tang, L., Wang, S. and Li, L. (2018), "Big data in tourism research: a literature review", *Tourism Management*, Vol. 68, pp. 301-323.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Line, N. D., Dogru, T., El-Manstrly, D., Buoye, A., Malthouse, E., & Kandampully, J. (2020). Control, use and ownership of big data: A reciprocal view of customer big data value in the hospitality and tourism industry. *Tourism Manage*ment, 80. Advance online publication. https://doi.org/10.1016/j.tourman.2020.104106

Liu, H., & Motoda, H. (Eds.). (1998). Feature extraction, construction and selection: A data mining perspective (Vol. 453). *Springer Science & Business Media.*

Lv, H., Shi, S., & Gursoy, D. (2022). A look back and a leap forward: a review and synthesis of big data and artificial intelligence literature in hospitality and tourism. *Journal of Hospitality Marketing & Management*, 31(2), 145-175.

Madison, W. (2019). “Booking.com Customers Underline Influence of ‘People Powered’ Reviews”, available at: https://www.travolution.com/news/travel-sectors/accommodation/booking-com-customers-underline-influence-of-people-powered-reviews/

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Mariani, M., & Baggio, R. (2021). Big data and analytics in hospitality and tourism: a systematic literature review. International Journal of Contemporary Hospitality Management.

Mariani, M., Baggio, R., Fuchs, M., & Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*.

Martin-Fuentes, E., & Mellinas, J. P. (2018). Hotels that most rely on Booking. com–online travel agencies (OTAs) and hotel distribution channels. Tourism Review.

Martinez-Torres, M. D. R., & Toral, S. L. (2019). A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. *Tourism Management*, 75, 393-403.

Mathew, E. (2019, February). Big Data Analytics in E-procurement of a Chain Hotel. In International Conference on Emerging Internetworking, Data & Web Technologies (pp. 295-308). Springer, Cham.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). Parallel distributed processing (Vol. 2, pp. 20-21). Cambridge, MA: MIT press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5, 115-133.

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), 1-9.

McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.".

Moon, J. W., Jung, S. K., Lee, Y. O., & Choi, S. (2015). Prediction performance of an artificial neural network model for the amount of cooling energy consumption in hotel rooms. Energies, 8(8), 8226-8243.

Murphy, C. (2017), “Report: 78% of all online hotel reviews come from the top four Sites”, available at: https://learn.revinate.com/blog/report-78-of-all-online-hotel-reviews-come-from-the-top-four-sites

Pan, B. and Yang, Y. (2017), “Forecasting destination weekly hotel occupancy with big data”, *Journal of Travel Research*, Vol. 56 No. 7, pp. 957-970.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.

Pereira, L. N., & Cerqueira, V. (2022). Forecasting hotel demand for revenue management using machine learning regression methods. *Current Issues in Tourism*, 25(17), 2733-2750.

Phillips, P., Barnes, S., Zigan, K., & Schegg, R. (2017). Understanding the impact of online reviews on hotel performance: an empirical analysis. *Journal of Travel Research*, 56(2), 235-249.

Plisson, J., Lavrac, N., & Mladenic, D. (2004, October). A rule based approach to word lemmatization. In Proceedings of IS (Vol. 3, pp. 83-86).

Puh, K., & Babac, M. B. (2022). Predicting sentiment and rating of tourist reviews using machine learning. *Journal of Hospitality and Tourism Insights*, (ahead-of-print).

Radissonhotels.com. (2023). Overview & Rooms. Retrieved on March 12, 2023, from https://www.radissonhotels.com/en-us/hotels/radisson-blu-helsinki-seaside

Radissonhotels.com. (2023). Retrieved on March 12, 2023, from https://www.radissonhotels.com/en-us/destination/finland

Rezaei, F., Raeesi Vanani, I., Jafari, A., & Kakavand, S. (2022). Identification of Influential Factors and Improvement of Hotel Online User-Generated Scores: A Prescriptive Analytics Approach. Journal of Quality Assurance in Hospitality & Tourism, 1-40.

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management*, 57, 12–20. https://doi.org/10.1016/j.tourman.2016.04.008

Russom, P. (2011). Big data analytics. TDWI best practices report, fourth quarter, 19(4), 1-34.

S-ryhma.fi. (2023). Retrieved on March 12, 2023, from https://s-ryhma.fi/en/about-us/business-operations

Sánchez-Medina, A. J., & C-Sánchez, E. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management*, 89. Advance online publication. https://doi.org/10.1016/j.ijhm.2020.102546

Sánchez-Medina, A. J., & Eleazar, C. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. International Journal of Hospitality Management, 89, 102546.

Sánchez-Medina, A. J., & Eleazar, C. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. International Journal of Hospitality Management, 89, 102546.

Saputro, P. H., & Nanang, H. (2021). Exploratory data analysis & booking cancelation prediction on hotel booking demands datasets. Journal of Applied Data Sciences, 2(1), 40-56.

Scikit-learn. (2023). Getting Started. Retrieved on March 17, 2023, from https://scikit-learn.org/stable/getting_started.html

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), 427-437.

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. Academic emergency medicine, 18(10), 1099-1104.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological methods, 14(4), 323.

Sun, S., Cao, Z., Zhu, H., & Zhao, J. (2019). A survey of optimization methods from a machine learning perspective. *IEEE transactions on cybernetics*, 50(8), 3668-3681.

Sun, S., Wei, Y., Tsui, K.-L. and Wang, S. (2019), "Forecasting tourist arrivals with machine learning and internet search index", *Tourism Management*, Vol. 70, pp. 1-10.

The Economist (2017), "The world's most valuable resource is no longer oil, but data", 6 May, available at: www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (accessed 5 October 2021).

Tian, X., & Pu, Y. (2008). An artificial neural network approach to hotel employee satisfaction: The case of China. Social Behavior and Personality: an international journal, 36(4), 467-482.

TIEKE (2007). Finnish Information Society Development Centre. *Publications*. Retrieved October 26, 2007 from http://www.tieke.fi/in_english/publications/.

Tripadvisor.com. (2023). About Us. Retrieved on January 4, 2023, from https://tripadvisor.mediaroom.com/us-about-us

Uzun, E., Yerlikaya, T., & Kırat, O. (2018). Comparison of python libraries used for web data extraction. Journal of the Technical University-Sofia Plovdiv Branch, Bulgaria, 24, 87-92.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165, 234-246.

Wang, Q. Q., Yu, S. C., Qi, X., Hu, Y. H., Zheng, W. J., Shi, J. X., & Yao, H. Y. (2019). Overview of logistic regression model analysis and application. Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine], 53(9), 955-960.

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In COLING 1992 volume 4: The 14th international conference on computational linguistics.

Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. PhD thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA.

White, M. (2012). Digital workplaces: Vision and reality. Business information review, 29(4), 205-214.

Wongsuphasawat, K., Liu, Y., & Heer, J. (2019). Goals, process, and challenges of exploratory data analysis: An interview study. arXiv preprint arXiv:1911.00568.

World Committee Tourism Ethics. (2017). Recommendations on the responsible use of ratings and reviews on digital platforms. In 3rd International Congress on Ethics and Tourism.

Wu, D. C., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. International Journal of Contemporary Hospitality Management.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management, 58, 51-65.

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

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5, 1205-1224.

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere, 10(7), 373.

Zhang Hongwu, & Zhang Jishun. (2019). My opinion on the application of big data in hotel marketing. *Market Weekly Theory Edition* (17), 2. ([1]张宏武, & 张继舜. (2019). 大数据在酒店营销中的应用之我见. 市场周刊•理论版(17), 2.)

Zhang, D. (2018, October). Big data security and privacy protection. In 8th international conference on management and computer science (ICMCS 2018) (Vol. 77, pp. 275-278). Atlantis Press.