School of Industrial and Information Engineering

Department of Electronics, Information and Bioengineering

Master of Science in Computer Science Engineering

POLITECNICO
MILANO 1863

Dynamic Pricing in the Hospitality industry in the presence of data scarcity

Supervisor: **Francesco Trovò Ph.D.**

Master thesis of:

Vinay Krishna Munnaluri, 10780972

Academic Year 2022-23

Contents

List of Figures

List of Tables

Abstract

Over the last few decades, dynamic pricing has become a more prominent topic of study, owing to significant progress in the fields of statistics, economics, and information technology. However, the complex nature of dynamic pricing has prevented it from being widely adopted in business. The traditional approach of setting prices based on cost, profit, and selling price is still widely used in business. Due to the constant changes in data and the variability of prices, this method is no longer effective. Therefore, new methods for dynamic pricing are necessary to maintain a competitive edge. The purpose of this work is to develop an application that will assist businesses and organizations of the hospitality sector in understanding and making better pricing decisions.

Dynamic pricing is an effective approach for businesses to better use their resources. By adjusting the prices of their products based on the demand fluctuations over time, businesses are able to maximize their profits. By utilizing these techniques, businesses can better understand how they have performed in the past, make necessary changes in their pricing strategies to keep up with market demands and take into consideration data that may be volatile. The purpose of this goal-oriented approach is to assist a decision-maker in the organization regarding pricing decisions.

This thesis presents a procedural framework that was implemented for the application of these techniques in the hospitality industry, specifically in resorts. It also outlines the challenges encountered during that time and the methods used to overcome them. The data utilized in this framework demonstrates the efficacy of modelling and decision-making in the face of numerous anomalies and limited data. We show that evolution of new techniques and consistently updating the model provide better results than regular approaches. We also show that occupancy rate improved a lot by adjusting prices in the favour of customers while also balancing the profits.

Sommario

Negli ultimi decenni, la tariffazione dinamica è diventata un argomento di studio cruciale per il business, grazie ai progressi significativi nei campi della statistica, dell'economia e della tecnologia dell'informazione. Tuttavia, la natura complessa della tariffazione dinamica ne ha impedito l'ampia adozione diffusa nel mondo degli affari. L'approccio tradizionale di fissare i prezzi in base al costo, al profitto e al prezzo di vendita è ancora ampiamente utilizzato nel mondo degli affari. A causa dei continui cambiamenti dei dati e della variabilità dei prezzi, questo metodo non è più efficace. Pertanto, sono necessari nuovi metodi di determinazione dinamica dei prezzi per mantenere un vantaggio competitivo. Lo scopo di questo lavoro è sviluppare un'applicazione che aiuti le aziende e le organizzazioni del settore alberghiero a comprendere e prendere decisioni migliori sui prezzi.

Il prezzo dinamico è un approccio efficace per le aziende per utilizzare al meglio le proprie risorse. Adeguando i prezzi dei loro prodotti in base alle fluttuazioni della domanda nel tempo, le aziende sono in grado di massimizzare i loro profitti. Utilizzando queste tecniche le aziende possono comprendere meglio come si sono comportate in passato, apportare le modifiche necessarie alle loro strategie di prezzo per stare al passo con le richieste del mercato e prendere in considerazione dati che potrebbero essere non pi validi. Lo scopo di questo approccio orientato agli obiettivi è assistere un revenue manager nell'organizzazione per quanto riguarda le decisioni sui prezzi.

Questa tesi presenta un quadro procedurale che è stato implementato per l'applicazione di queste tecniche nel settore dell'hospitality, in particolare nei resort. Descrive inoltre le sfide incontrate in quel periodo e i metodi utilizzati per superarle. I dati utilizzati in questo quadro dimostrano l'efficacia della modellazione e del processo decisionale a fronte di numerose anomalie e alla disponibilit di dati limitati. Dimostriamo che l'evoluzione di nuove tecniche e l'aggiornamento costante del modello forniscono risultati migliori rispetto agli approcci tradizionali. Mostriamo anche che il tasso di occupazione è migliorato molto modificando i prezzi a favore dei clienti bilanciando anche i profitti.

1. Introduction

Pricing techniques in the hospitality industry have undergone significant changes in recent years in response to the rapidly evolving market. In the past, traditional methods of pricing, such as fixed or seasonal rates, were the norm. However, these methods proved to be insufficient in dealing with the fluctuations in demand and market conditions, leading to the development of modern, dynamic pricing techniques.

Dynamic pricing, which allows businesses to adjust prices in real-time based on market conditions and other factors, has become increasingly popular in the hospitality industry. The goal of dynamic pricing is to optimize revenue and profits by offering prices that are more attractive to customers during times of low demand, and higher prices during periods of high demand. Volatility of prices is a major consideration in dynamic pricing. Prices can fluctuate rapidly based on changes in demand and market conditions, making it challenging for businesses to keep up. To address this issue, businesses use complex algorithms and data analysis to determine the optimal prices for their products and services. These algorithms consider factors such as competitor prices, customer behaviour, and external factors, such as weather and events, to ensure that prices are set at a proper level. Competitor prices also play a critical role in dynamic pricing. Businesses must be aware of what their competitors are charging and adjust their prices accordingly to remain competitive. The right price must consider the cost of goods and services, as well as other expenses, such as marketing and overhead.

Customer behaviour is also an important consideration in dynamic pricing. Businesses must understand the needs and preferences of their customers and adjust their prices accordingly. For instance, customers might display greater price sensitivity during slower periods, such as those characterized by lower demand, such as those experienced during pandemics and periods of inflation making it important for businesses to offer lower prices to attract them. On the other hand, during busy periods, customers may be more willing to pay higher prices, allowing businesses to maximize their revenue.

External factors, such as weather and events, can also have a significant impact on demand and pricing. Businesses must consider these factors when setting prices to ensure that they are responding appropriately to changes in the market. For example, prices may be lower during periods of bad weather, as fewer customers are likely to be traveling, while prices may be higher during peak travel seasons or when major events are taking place. The availability of internet and the complexity of data inclusion have also had a major impact on pricing techniques in the hospitality industry. Today, businesses have access to vast amounts of data and information, allowing them to make more informed decisions about pricing. However, this also makes the process of setting prices more complex, as businesses must consider a wide range of factors, including market trends, customer behaviour, and competitor prices.

To date, however, pricing techniques in the hospitality industry have evolved significantly in recent years, moving from traditional, fixed rate methods to dynamic pricing, which takes into account a wide range of factors, including volatility of prices, competitor prices, customer behaviour, external factors, finding the right price, revenue optimization, meeting market demands, costs and profits, traditional methods, modern techniques, internet availability, and the complexity of data inclusion. By taking these factors into account, this thesis focus on all these aspects and how they are incorporated into a pricing strategy while also considering business rules.

1.1 Motivation

In this thesis, we focus on a specific case of dynamic pricing problem for the hospitality sector, more specifically for hotel and resorts business where bookings for various types of rooms occurring from multiple channels and OTA(Online Travel Agencies). Some bookings even pop up from client’s website and offline books. People used to make bookings either by visiting the hotel or through a travel agency which has been an old-fashioned way for so many decades. Now with the availability of internet and age of mobile phones people spend significant amount of time to do the research before making decisions [1]. So, the modern way of bookings is done mostly through OTAs which we also refer to as channels in this study. Furthermore, a single room in a hotel is now being published on multiple sources(i.e., more than 10) at the same time and also with contracts and agreements between tour operators and agencies the number keeps growing. On the one hand, people who wish to book online have access to a vast amount of information about the rooms available and can make a decision at their convenience. On the other hand, neither the OTAs nor the owners have any insight into what their customers are considering when choosing or rejecting a particular room. This disparity in the amount of information available to each party creates a "black box" that forces the OTAs to keep their prices as low as possible and explore all possible options to understand customer behaviour. However, the profit from a sale of single booking depends not only on the cost and selling price but also on the commission that need to be hand out due to agreements between agents, OTAs and owners, hence having higher price can result higher profits. Furthermore, if the whole process of booking is only about the room price, then it would have been easier but some customers prefer to make bookings along with some additional benefits like breakfast, room and spa, room services and many other bonus services which make the study even more complex. At the time of this thesis, there are multiple OTAs involved in the bookings for the same hotel and also data has been collected for few years. Some information about properties of a requested booking includes, e.g. channel(OTA), booking_date, booking_from , booking_to , number of persons staying, family or individual, type of room, bonus services, number of nights of stay. Moreover, there has been a pandemic COVID 19 for a year or two during which several hotels have been closed and this impacted heavily on bookings. One way to do this is to exploit possible machine learning techniques while adding all the available internal and external features mentioned above along with COVID-19 pandemic and predicting the price that will maximize the revenue.

1.2 Proposed Solution

The price model in hospitality sector is not a continuous but rather it is a discrete because each room has a set of prices named as BARs(Best Available Rates) and the price of room is picked from a set and adjusted to one BAR higher or lower from time to time depending on seasonality and trends. In the existing model, the price for a given type of room has 7 different BAR prices which are used for the whole season and all other OTAs have certain formulae to calculate sale price from the BAR price defined by the owners which is finally shown to the customers . This way both the clients and OTAs benefit from the profits and prices are less volatile. Since the prices are not continuous and the price of room cannot change more than level at a time which would make customers lose interest in the room with the sudden spike. For example, the price of room is 50$ which is in level 5 and we cannot jump to level 3 or level 7 suddenly which would represent 150$ or 10$ huge difference among prices. The price can only jump to level 4 or level 6 which makes either 100$ or 40$ to be consistent. Finally, after considering all these limitations, we decided to build a model to determine when we can perform the change in price level. For this purpose, numerous features are considered and various models are tried but primary feature that we decided to use was occupancy rate which is a continuous value. The data collected related to bookings daily and occupancy rate was calculated for several periods like daily, weekly, monthly, yearly and many others depending upon the goals and also definition of the model. In our opinion, it is very difficult to build a single model which can fit all the features and business rules. Therefore, our idea is to find a meaningful patterns in the data and predict the occupancy rate for that period. Furthermore, the data is a time series and so the seasonality and trends are bound to appear and represent auto-correlated characteristics. Thus, we decided to use ridge regression after trying out many multi variate machine learning models. We first apply standard ETL (extract, transform, load) procedure and perform feature selection among the data and insert external features. Then the selected data is trained with the models and result is passed through a set of functions determining business rules where the final price level is decided. We observe that our model is adaptable to existing business model and adjusts the price autonomously with right price at a correct time returning higher returns than previous year.

1.3 Thesis Structure

- In Chapter 2 we present the literature survey of related works for dynamic pricing techniques and time series forecasting applied in context of bookings.

- In Chapter 3 we provide an overview of theoretical background of techniques we use in our implementation.

- In Chapter 4 we describe the problem of ORP (Occupancy Rate Prediction) and existing pricing strategy in detail.

- In Chapter 5 we explain details of proposed solution for occupancy rate prediction and adjusting price level from it.

- In Chapter 6 we present the experiments and results comparison with existing solution and additional details

- In Chapter 7 we summary the solution, comment overall performance and provide ideas for future developments.

2. Related Works

Within this Chapter, we offer an overview of a variety of articles and reports that were employed during the research phase. These resources provide a comprehensive analysis of various subjects and furthermore illustrate the application of certain methodologies utilized in these thesis, though under differing data sets or scenarios.

2.1 Discrete Dynamic Pricing

Discrete dynamic pricing refers to a pricing strategy that adjusts the prices of goods or services in real-time, considering market conditions, consumer behaviour, and other factors. The pricing of a product is dynamically adjusted based on real-time market data, consumer behaviour, and other relevant factors. The approach allows companies to set prices that are competitive and responsive to market changes, without sacrificing profitability. The main objective of this strategy is to maximize revenue by setting prices that reflect changes in demand over time [2]. Hotels face many challenges due to several economical products like rooms are intangible, heterogenous offerings, over occupied and under occupied situations, fixed capacity adds constraints on pricing [3].

The history of discrete dynamic pricing goes back to the 1950s, when it was first implemented by airlines to fluctuate ticket costs depending on the demand. Other industries, like hospitality and retail, soon followed suit and began adopting this pricing strategy. Airlines, on the other hand, typically offer two product classes, such as business/first and coach, while hotels can often have a variety of room types, accompanied by different bed configurations, thus, leading to various product offerings [4]. Hotels initially had a difficult issue when establishing the prices of those products, and then to add to the complexity of the situation – they must also address the dynamic pricing problem, where the prices of those offerings may change as the reservation date draws nearer.

The range of possibilities when it comes to discrete dynamic pricing is immense and continues to expand. It is important to note that a substantial portion of sales within the hospitality sector are still done through customary contracts, while dynamic pricing is mainly employed when trying to reach the online market [5]. Companies are progressively adapting this pricing model to remain competitive and to be able to respond to ever-changing market conditions. The use of sophisticated algorithms and large datasets has made it simpler for businesses to implement dynamic pricing, which is expected to be more and more popular in the future. Discrete dynamic pricing is an adjustable and versatile pricing approach that can be implemented to an array of products and services, making it a desirable option for businesses that seek to improve their pricing strategies.

2.2 Statistical approaches

Pricing has been a topic of great interest for many researchers and statisticians over the years. One of the early theories was presented by Zheng Gu in 1997, who believed that room pricing was simply determined by the construction costs of a hotel room. He proposed the use of Hubbart formula, which considered only the construction costs of a room, the desired profit, and the expected number of rooms to be sold, ignoring the concept of market demand [6]. However, this model was criticized for its limitations in that it only considered the construction costs and did not guarantee that it covered all the operating costs and profits actually achieved.

Armstrong (2006) then introduced the concept of price discrimination and the benefits of model markets. He argued that the platform (OTA) serves particular genders better and that the price is determined by cost, market power, commissions, and offers. Armstrong looked at the entire process as a monopoly and believed that focusing on a particular group would attract other groups to perform better on any OTAs [7]. He proposed the concepts of single homing and multi-homing, where platforms are benefited through per-transaction charges rather than fixed fees.

Costa (2013) took a different approach to room pricing by proposing that several factors should be considered when determining the price of a room. He considered factors such as services offered along with the room, the seaside view, bed and breakfast, size of room, and geographical location proximity to the centre or any tourist spots [3]. He created a linear regression system with 75 dichotomous variables as independent or explanatory variables of the 379 prices considered. He found that the internet was suffering from various high-level errors for gathering data, as hotels tend to offer several undescribed benefits and offers to families and bulk transactions. He used regression analysis to study the impact of these variables on room prices.

In conclusion, the pricing of hotel rooms has been a topic of great interest for many years, with various theories and models being proposed. From Zheng Gu's construction cost-based model, Armstrong's price discrimination and model markets theory, to Costa's multiple factor approach, each has made a significant contribution to the understanding of room pricing. However, the limitations of these models have also been highlighted, and further research is needed to develop a more comprehensive and accurate model that considers all the factors that influence room pricing.

2.3 Theoretical approaches

The pricing of rooms in the hotel industry is a crucial factor in determining the success of a hotel business. A proper pricing strategy is essential for attracting customers and ensuring profitability. The pricing strategy that hotels adopt depends on various elements, including market demand, competition, seasonal trends, occupancy levels, and the information available on the internet. The internet has played a significant role in the marketing process and has helped customers to determine the price of a room.

Recently, researchers have been focused on constructing theoretical approaches to setting room prices within the hotel industry. One such researcher, Yelkur (2001), stated that the internet is an essential component of the marketing process and the availability of data on the web about a hotel can help customers determine the value of the price [8]. Yelkur also studied the concept of differentiating pricing on the web and found that having a good relationship with customers helps to maintain steady prices.

To gain a better understanding of customer behaviour when it comes to making purchases, hotels have multiple products and facilities that they offer. By categorizing customers into segments, hotels can increase their booking count.

For instance, the price of a room in California would differ from that in Arizona, and those customers who are there for business would typically pay higher than individual customers. This segmentation and price discrimination would result in increased profits.

Croes and Semrad (2012) focused their investigation on the impact of fluctuating discount rates on the financial performance of hotels. They analysed the price elasticity and seasonality, and found that during some times of the year, hotels would maintain their regular rack rates, while during other times, they would adjust the price according to demand [9].

Kefela (2014) conducted a study on hedonic pricing, which accounted for the external and internal factors that affect the good and its qualities when evaluating the price. He discovered the various techniques employed by hotels to attract customers, such as offering free parking spots, obtaining a higher star rating, taking advantage of its geographical superiority, or providing free breakfast, with the intention of increasing bookings [10]. These additional benefits raise the cost of a room, but they also make it difficult for customers to reject when booking.

In conclusion, the pricing of rooms in the hotel industry is a crucial element in the success of a hotel business. A proper pricing strategy can help attract customers and ensure profitability. Researchers have studied various approaches to setting room prices, including the impact of fluctuating discount rates, the use of the internet in the marketing process, and hedonic pricing. The internet has played a significant role in the marketing process, and the information available on the web has helped customers determine the worth of the price. The hotel industry should also focus on categorizing customers into segments and providing additional benefits to increase bookings and profits.

2.4 evolution of information

One of the key factors in hotel pricing is the consideration of external and internal variables. These can include holidays, events, location, additional services offered, discrimination, competition, and other discrete and quantitative features that affect room rate pricing. Holidays, for example, can lead to an increase in demand for hotel rooms, which in turn drives up prices. Conversely, events can decrease demand and lead to lower prices. The location of the hotel can also play a significant role in determining room rates, with hotels in popular tourist destinations often commanding higher prices.

The hotel industry has also experimented with various pricing strategies to maximize revenue. Intertemporal price discrimination is one such strategy, which involves setting different prices for similar products based on the time of purchase. This allows hoteliers to target customers who are more price-sensitive and adjust prices to reflect demand. For example, a hotel might offer lower prices during weekdays and higher prices on weekends, depending on demand [10]. This strategy has been found to be effective in maximizing revenue for the hotel, as it allows for a more dynamic pricing structure that adjusts based on demand.

In recent years, the hotel industry has also started to utilize advanced data analytics to better understand customer behaviour and develop more effective pricing strategies. This involves the use of big data, machine learning algorithms, and other advanced technologies to gather and analyse vast amounts of data on customer preferences, behaviour, and purchasing patterns. With this information, hoteliers can develop more informed pricing strategies that consider factors such as customer demographics, travel patterns, and purchasing behaviour.

Abrate et al. (2019) conducted comprehensive research on the subject of setting the right price at the right time. They focused on exploring the impact of frequent price changes on maximizing the revenue of the hotel. The study was cantered on the concept of inter temporal price discrimination [11]. This refers to the practice of setting different prices for similar products based on the time of purchase but ensuring that the prices remain consistent for all customers at any given time.

The researchers aimed to find out if this pricing strategy could bring about an improvement in the revenue generation for the hotel. In order to accomplish this, the team carried out extensive research, gathered data and analysed it to arrive at their conclusions. They found that inter temporal price discrimination can be a very effective way of maximizing the revenue of a hotel. By charging different prices for similar products at different times, the hotel can target different segments of customers and tap into their varying levels of willingness to pay.

This helps to ensure that the hotel is making the most out of its available resources and is generating the maximum amount of revenue.

The findings of Abrate et al. (2019) are of great significance for the hospitality industry, as it highlights the importance of adopting dynamic pricing strategies. By doing so, hotels can optimize their pricing policies, take advantage of fluctuations in demand, and improve their revenue generation. The results of this research are highly relevant to managers, researchers, and academics in the hospitality industry, who are looking to explore new and innovative pricing strategies.

In addition to utilizing advanced data analytics, the hotel industry has also started to experiment with dynamic pricing. This involves using real-time data to adjust prices in response to changes in demand. This can include factors such as weather, economic conditions, and other external factors that affect demand for hotel rooms. By using dynamic pricing, hotels can ensure that they are always charging a fair price for their rooms, which in turn maximizes revenue and customer satisfaction.

By staying up to date with the latest research and best practices in the field, hotels can continue to develop effective pricing strategies that meet the changing needs of their customers and maximize their profits.

2.5 Assumptions

In the hospitality industry, the discrete dynamic pricing approach has been applied to increase revenue. However, certain assumptions must be considered for it to be successful.

1. It is assumed that customer behaviour is predictable and that they will respond to price changes in a certain way, enabling accurate forecasting of demand and revenue.
2. The approach is most effective in stable market conditions where supply and demand can be estimated with precision.
3. The effectiveness of the approach is reduced in highly competitive markets where prices are influenced by multiple factors.
4. The approach assumes that price changes occur gradually over time and are infrequent, allowing customers to adapt to the new prices.

5. The approach requires access to data on supply, demand, and market conditions to make informed pricing decisions.

However, these assumptions were made for an ideal data under normal circumstances. The current thesis focuses on the development of models during abnormal situations, such as data scarcity due to a pandemic, lack of market trend and seasonality, and inability to use available data due to inconsistencies. Thus, the following chapters will verify whether these assumptions hold in such situations and propose solutions to deal with circumstances where these assumptions cannot be applied to the data.

2.6 Data Driven approaches

The Hospitality industry has seen a surge in the adoption of data-driven approaches for discrete dynamic pricing in recent times, owing to the technological advancements and the availability of data. This approach leverages the power of data to make informed pricing decisions and optimize prices based on the current market conditions [2]. The data used in data-driven pricing approaches in the hospitality industry can be sourced from various channels such as reservations, website traffic, social media, customer profiles, and competitor data [1]. The key elements used in these data driven approaches include historical booking data, customer demographic information, and market-level data [12].

The type of data used in data-driven pricing approaches in the hospitality industry can be categorized into several categories, including the type of season (Season), the day of the week (Day), the length of stay (Length), the length of the time period between the reservation and the check-in time (Before), and the tariff class (Tariff) [13]. Additionally, data on the type of booking, whether it is a group booking or an individual booking, and the inclusion of breakfast (B&B) can also be used in these approaches.

One of the key advantages of data-driven approaches for dynamic pricing is that they provide more accurate and real-time pricing information. With the increasing availability of big data, businesses can collect and analyse vast amounts of pricing data to make informed pricing decisions [6]. This data can help identify patterns and trends that would otherwise be hidden and can be used to improve pricing strategies and increase profitability. By using predictive models, data mining, and machine learning algorithms, businesses can make better use of their pricing data to identify market trends and consumer behaviour, which can be used to optimize pricing strategies.

In 2013, Abd El-Moniem Bayoumi et al. conducted a study exploring the use of Monte-Carlo simulation in pricing techniques. The authors argued that dynamic pricing, as opposed to quantity-based pricing, will become the dominant approach in the future. They introduced a new method in which the price of a room is first determined by the managers and then adjusted using various multipliers, such as seasonality, room inventory, time until arrival, and occupancy rate [14]. These variables are transformed into decimal numbers ranging from 0 to 1.0 to represent the amount of change to be made to the price. The final price is determined by multiplying the existing price with all the multipliers. The researchers modelled the reservations and cancellations using Bernoulli trials and noted that there is a distinct pattern in the curve. Additionally, they utilized exponential smoothing for price forecasting.

P. Talon-Ballestero et al. (2022) conducted research on the utilization of data in price optimization and customer-centric pricing [15]. The study viewed data from four dimensions: data, optimization, information processing systems, and key performance indicators (KPIs). The authors proposed an open pricing concept, which is a sophisticated pricing discrimination technique that offers real-time price recommendations without rate changes and optimizes pricing on an individual night basis. This concept represents a significant departure from previous pricing methods, which typically made changes on a per-room basis rather than a per-night basis.

Despite the widespread use of data-driven approaches for discrete dynamic pricing, there are several problems associated with the data used in data driven approaches. One of the main challenges is data accuracy and reliability, as the data used to make pricing decisions can be outdated or incorrect. Another challenge is data privacy and security, as the use of customer data for pricing purposes may be perceived as intrusive by some customers. As we all know COVID-19 pandemic has led to complete shutdown of the world for almost two years which created a tremendous imbalance in hospitality industry [16]. Several last-minute bookings and sudden last minute surcharges/discounts applied led to inconsistencies in data compared to the previous years. This led us to study data with short term periods more frequent rather than historical data more accurate.

Several methods and algorithms have been used over the years to implement data-driven approaches for discrete dynamic pricing. These include regression analysis, time series analysis, and simulation models like Monte Carlo where numerous random paths for the price of an underlying asset are generated, each having an associated payoff [17] [14], LSTM [18], Random Forest [19] and many others.

3. Context

In this chapter we discuss about the background and theoretical concepts of the industry and also the introduction and explanation of most relevant techniques and algorithms used in this thesis.

3.1 history of pricing in hospitality industry

The hospitality industry is one of the most dynamic and rapidly evolving sectors of the global economy. The dynamics of the hospitality industry have a significant impact on room pricing and revenue management. These dynamics are shaped by various factors such as consumer behaviour, sociodemographic factors, psychological factors, barriers, competitors, market demand, volatility, constraints, developments, internet, availability of data, consumer research, human judgment, discounting, offers, choices made by hotel managers, market equilibrium, and customer segmentation. This thesis will explore these factors in-depth and examine their impact on pricing and revenue management in the hospitality industry [1] [8] [3].

Consumer behaviour is a crucial factor that affects room pricing and revenue management. Customers' preferences are influenced by factors such as value for money, quality of service, convenience, and personalization. As such, hospitality providers need to be aware of customers' evolving needs and adjust their pricing strategies accordingly [15] [14]. For example, hotels may adjust their room pricing based on seasonal demand, customer preferences, and availability.

Sociodemographic factors such as age, gender, income, education, and lifestyle also affect room pricing and revenue management. Hotel providers need to understand their customers' sociodemographic profiles to offer relevant packages and promotions. For instance, luxury hotels may offer premium packages to high-income customers, while budget hotels may offer discounts to students or backpackers [10].

Psychological factors such as attitudes, perceptions, and emotions also impact room pricing and revenue management. Customers' perceptions of the quality of service, ambiance, and level of personalization can affect their willingness to pay for a room. Moreover, customers may choose to pay more for a room that offers an immersive experience that aligns with their emotions and moods, such as relaxation or excitement [20].

Barriers such as language, cultural differences, and legal regulations also affect room pricing and revenue management. For instance, hotels operating in countries with strict licensing and taxation regulations may have higher room prices to offset these costs. Additionally, hotels may need to offer multilingual services to cater to customers from different countries.

Competitors are a significant factor that affects room pricing and revenue management. Hotels need to differentiate themselves from their competitors by offering unique experiences, exceptional service, and competitive pricing [21] [22]. Moreover, hotels need to be aware of competitors' pricing strategies and adjust their room rates accordingly.

Market demand and volatility also affect room pricing and revenue management. Hotels need to be aware of market trends and adjust their pricing strategies to meet demand. Moreover, external factors such as natural disasters or pandemics can affect room pricing and revenue management. Hotels may need to adjust their room pricing during periods of low demand to attract customers.

The availability of data, internet, and consumer research is a significant factor that affects room pricing and revenue management. Hotels can leverage data and consumer research to develop more targeted pricing strategies [1]. Moreover, hotels can use the internet to offer online booking services and promotions to attract customers.

Human judgment, discounting, and offers are factors that affect room pricing and revenue management [23]. Hotel managers need to make informed decisions when setting room prices and offering discounts and promotions. Moreover, hotels may offer discounted rates during periods of low demand to attract customers and maximize occupancy rates.

Choices made by hotel managers, market equilibrium, and customer segmentation also affect room pricing and revenue management. Hotel managers need to understand the market equilibrium to set competitive room prices [22]. Additionally, hotels may segment their customers based on their sociodemographic profiles and adjust their pricing strategies accordingly.

In conclusion, the dynamics of the hospitality industry have a significant impact on room pricing and revenue management. Hospitality providers need to be aware of the various factors that affect pricing and revenue management and adjust their strategies accordingly. By leveraging data, consumer research, and technology, hotels can develop more targeted pricing strategies that meet customers' needs and maximize revenue. Ultimately, hotels need to be adaptive and responsive to changing market trends to remain competitive in the hospitality industry.

3.2 Type of data

Here are some common types of data found in the hotel industry, along with a brief description of each:

- Transactional Data: Data related to transactions, such as room bookings, food and beverage purchases, and spa services. This data is typically stored in a property management system (PMS) and can provide valuable insights into customer behaviour and spending patterns. [24]
- Customer Data: Data related to guests, such as their contact information, booking history, and preferences. This data is typically stored in a customer relationship management (CRM) system and can be used to personalize the guest experience and improve customer satisfaction. [12] [24] [20]
- Operational Data: Data related to the operation of a hotel, such as employee schedules, inventory levels [21], and maintenance records. This data is typically stored in an enterprise resource planning (ERP) system and can be used to optimize operations and reduce costs.

- Social Media Data: Data related to guest sentiment and feedback, such as online reviews and social media posts. This data is typically collected from social media platforms and can provide insights into guest satisfaction and areas for improvement [25].
- Competitive Data: Data related to competitors, such as pricing, promotions, and market share [22]. This data is typically collected from industry reports and can be used to benchmark performance and inform pricing strategies.
- Web Analytics Data: Data related to website traffic, such as page views, bounce rates, and click-through rates. This data is typically collected using web analytics tools and can be used to optimize website performance and increase online bookings [26].
- Financial Data: Data related to financial performance, such as revenue, expenses, and profit margins. This data is typically stored in an accounting system and can be used to track performance and inform financial decisions [13] [5].

These types of data can be used to gain insights into various aspects of a hotel's operations and customer behaviour and can be analysed to make informed decisions that improve performance and enhance the guest experience.

Our focus in this research is on data that we obtain and utilize for time series analysis which is transactional, customer and competitor data.

3.2.1 ETL process

In the hospitality industry, hotels receive transactional data from a variety of sources, including online travel agencies (OTAs) [27]. This data is stored in cloud data warehouses in the form of files, which are often large in volume and size. To ensure that this data can be used for analysis, it is important to go through an ETL (Extract, Transform, Load) process that brings all the data to the same structure and removes any inconsistencies [28].

The first step in the ETL process is the extraction of the data from its source. This involves retrieving the data from the various information systems, such as

OTAs, and transferring it to the cloud data warehouse. Once the data is in the warehouse, the transformation process begins.

During the transformation process, various formulas, aggregations, and validations are applied to the data to obtain fact and dimension tables. This is a crucial step as it ensures that the data is in the correct format and can be easily used for analysis. Any redundant, duplicate, or inconsistent data is eliminated, and data entry errors are corrected [27].

The final step in the ETL process is the loading of the data into the data warehouse. The data is loaded into a database with multiple indexed columns that enable the hotel to perform in-depth analysis and generate insights.

In our case client stored all the raw data in form excel files in Filesystem which is accessible via SSH. So we utilized the sftp package in python to read the files and load them into data ware house. We used database connectors provided as python [29] package pymysql [30] to store output of ETL process into the MySQL database. As for transformations and all the pre-processing we used pandas [31] library which is a python package built for data processing.

In summary, the ETL process is critical to the hospitality industry as it streamlines transactional data from various sources and ensures that it is in the correct format for analysis. By removing any inconsistencies and correcting errors, hotels can generate accurate insights that can inform their decision-making and drive business success.

3.2.2 Data Wrangling

The hospitality industry generates immense amounts of data, with thousands of rows being generated in mere seconds. Once the ETL process is complete, the next step is data wrangling. This process involves cleaning, organizing, and transforming raw data into a desired format that analysts can use for prompt decision-making. One of the primary objectives of data wrangling in the hospitality industry is to cleanse the data of any noise, flaws, or missing elements. This helps to improve data usability by ensuring that the data is accurate and relevant. Data wrangling also converts the data into a compatible format for the end system, making it easier for analysts to work with.

Furthermore, data wrangling in the hospitality industry helps to quickly build data flows within an intuitive user interface. This makes it easy to schedule and automate the data-flow process, which can save time and improve efficiency. It also integrates various types of information and their sources, including databases, web services, and files.

Finally, data wrangling in the hospitality industry allows users to process very large volumes of data easily and to share data-flow techniques. This can be extremely beneficial for organizations that rely heavily on data to make informed decisions [32].

3.2.3 Feature Selection

Feature selection is a crucial process in the hospitality industry, where businesses need to analyse vast amounts of data to determine the optimal pricing, predict demand, and optimize sales. The process of feature selection helps businesses identify and select the most relevant features that have a correlation with price, demand, or sales, thereby improving the accuracy of their predictions, save time, and reduce costs.

There are several methods used in feature selection, including correlation analysis, feature importance ranking, and many other machine learning algorithms [33]. These methods help businesses identify the most significant features by removing irrelevant or redundant ones. Despite its benefits, feature selection in the hospitality industry also presents several challenges, such as data quality issues, data complexity, and the need for domain expertise. It is essential to have experienced data analysts with a deep understanding of the industry to extract the best features from the data. In our study we utilize this concept and spend most of the time here figuring out which columns are useful as input to machine learning model to get better results and minimal errors.

3.3 Data Analysis

The processing and storage of data from multiple data warehouses in a central database, followed by analysis, is a crucial step in deriving valuable insights.

This is especially important considering the presence of numerous products and the possibility that some products may have more transactions than others, which can significantly impact the final model. Moreover, this approach aids in the identification of outliers, trends, seasonality, and other valuable insights. For instance, it is possible for a customer to perform a booking of a hundred bookings in a single transaction, which could affect daily sales. Similarly, multiple customers may book the same room for the same stay night in a single day, leading to unbalanced weights. Through the analysis of such data, it is possible to identify these anomalies and adjust the final model accordingly.

3.3.1 Kurtosis and skewness

The frequency of bookings received on a daily, weekly, monthly, quarterly, and yearly basis is not consistent. Therefore, it is crucial to analyse and identify the period during which the majority of bookings are made. By doing so, hoteliers can prioritize their efforts and resources to maximize their profits. In order to obtain such insights, various univariate and multivariate normality tests are conducted. These tests assess kurtosis and skewness of different data columns with respect to time periods. Furthermore, the presence of skewness provides an opportunity for hoteliers to create premiums or discounts in favour of the side with greater skewness [34] . Since test is a generic one and we have several date columns in the data we could utilize it for various purposes. For example, we can use this skewness to check which part of the year has more bookings if we take booking_date along x-axis and daily bookings along y-axis then you could see that more bookings occur during 2nd quarter because it is the time most tourists visit the country and so we can provide more discounts during that time.

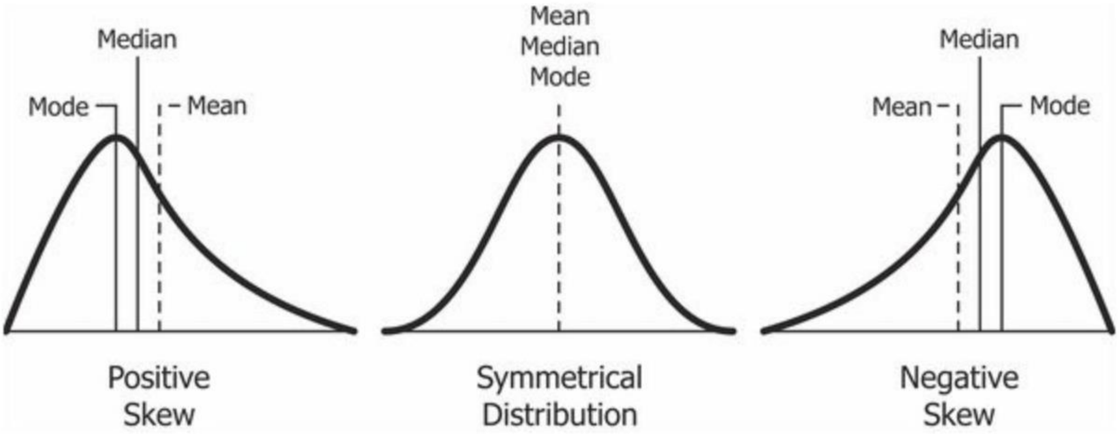


Figure 3.1 Different skewness distributions by Diva Jain CC BY SA.

In statistical analysis, measures of skewness and kurtosis are used to describe the shape of a probability distribution or data set. Skewness is a measure of symmetry, which determines the extent to which the distribution is symmetrical or asymmetrical. Specifically, it assesses the degree of deviation from the symmetry of a distribution around its central point. From the above example, we can say that skewness is a property of distribution with some abuse of significance and our data clearly proves this explained with more details in Chapter 5.

A data set is said to be symmetric if it is balanced, such that the observations on one side of the central point are a mirror image of those on the other side. The presence of skewness indicates that the data is skewed, that is, it is not symmetric, and one tail of the distribution is longer than the other [35].

On the other hand, kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. It assesses the extent to which the distribution has a heavy tail or a light tail, indicating the presence or absence of outliers. A normal distributed data with have bell-shaped curve when projected. Data sets with high kurtosis have heavy tails, which means that they have extreme values that occur more frequently than in a normal distribution. Conversely, data sets with low kurtosis have light tails, indicating that they have fewer outliers than a normal distribution. The uniform distribution represents the extreme case of low kurtosis [36].

3.3.2 Scaling

Scaling of variables is a standard pre-processing step utilized in function approximation and time series analysis. The primary goal of scaling is to assign weights to input variables based on their relevance to the estimated output. In simple words you convert the values to be specific range. For example, if you have 10 observations out of 8 observations have values in range between 1 to 10 and last two values are 89 and 95 and so if you try to plot these values in graph the last would be shown as outliers and y-axis in graph would be in range of 1 to 100 which makes the shapes of first observations very small and you can differentiate this because you can observe what if it is a scatter plot with 1000 observations and you wish to find outliers. In those cases scaling would help to identify outliers by transforming them using scaling techniques like min-max scalers gives you differentiation between outlier and normal values and so the model can learn from that. This technique is utilized to identify any redundant inputs and assign low weights to them to reduce their impact on the learning process [37]. As a result, variable scaling is a critical step in optimizing the learning process and enhancing the accuracy of output estimates.

3.4 Regression models

Supervised learning is a type of machine learning methodology in which an individual or system supplies input/output pairs that assist in recognizing the association between the related training data. Regression is a statistical processes for estimating the relation between a dependent variable(output) and one or more independent variables (input variables or features) [38]. The regression problem strives to ascertain a mathematical formula that accurately approximates the hidden function demonstrating the input/output relationship [39]. Regression helps in finding relationship between demand and period [40].

3.4.1 Ridge Regression Model

The solution to the issue of highly correlated regressors and a way to stabilize the linear regression problem was proposed by Hoerl and Kennard in 1970 - ridge regression. Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated [38]. What makes this technique so attractive is its easy to compute analytic solution, which allows the coefficients associated with the least relevant predictors to be shrunk towards zero without them reaching exactly zero [41]. Let us explain this with an example, the hospitality industry has a variety of characteristics, and researchers and data scientists are attempting to determine the correlation between various independent columns of the data and the cost and revenue of a given room [10]. In other words, since we do not the exact behaviour of the customer we try to relate the price of the room with other columns in data a linear relationship. It is sometime easy to build relationships between two columns of numbers and other time it can be difficult but there can always a relation be formed but if it is a real or fake is hard to determine. Nevertheless, there may be associations between these independent variables that can lead to multicollinearity among them, thus decreasing the accuracy of predictions [37]. Usually, the columns are omitted to bypass this issue but this results in the suppression of important data while modelling. Hence, ridge regression assists us in finding best coefficients with help of a concept called damping factor λ .

We use generally MSE (Mean Square Error) score for calculating the error between original and predicted values but this does not tell us error is due to which input features. In case of several independent features it is possible to obtain a biased result due to the overfitting of some features diminishing the impact of other features. So the MSE can be modified in a way such that the scores can demonstrate which features are causing the bias in the predictions. The modification is what differentiates ridge regression from linear regression.

Regular MSE (Mean Square Error) = $\sum(y - \hat{y_i})^2$ is modified into

Penalized MSE = $\sum(y - \hat{y_i})^2 + \lambda \sum_j^p \beta_j^2$

Where y is the actual value

$\hat{y_i}$ is the prediction value

λ is the damping factor

β_j^2 is the regression coefficients of the feature variables.

The damping factor is also known as the shrinkage coefficient and its job is to penalize the coefficients and verify their contribution to the outcome and if it finds the results are similar or better after making some of the coefficients to zero then those variables can be deemed as less relevant features. We have to set the value for damping factor and right value can help us determine best features to keep and discarded indirectly during the process [42] [41].

But we have to be careful because of the effects of lambda can be two ways.

- The bias increases as λ increases.
- The variance decreases as λ increases.

Reduction in variance is good thing but if we keep increasing the λ we may run into increased bias so we have find the optimal value which can reduce variance as well as not increase the bias. To do this we apply GridSearchCV [43] to find best lambda value by looping through a range of values and testing out each value and comparing results obtained from those. Therefore, we can use cross-validation techniques like GridSearchCV(which iterates through a loop of provided parameters for a model and returns a model with best set of value after testing it through a scoring metric)   to select the best value of λ that gives the best trade-off between model complexity and performance on the test set. GridSearchCV is explained in detail in later chapters with examples.

3.5 Random Forest Models

Random Forest was first introduced by Breiman [44] is an ensemble of randomized decision trees which makes decisions based on aggregation of results from these individual trees in the forest. Random forest uses the bootstrap aggregating or bagging technique to generate multiple decision trees. In bagging, a random sample of the training data is taken with replacement, and a decision tree is trained on each sample. This generates multiple trees with different decision boundaries. Once the individual decision trees are trained, their predictions are combined to generate the final prediction. In regression tasks, the predictions of the individual trees are averaged to generate the final output [45]. Random forest has several hyperparameters that can be tuned to optimize its performance. These include the number of trees, the depth of the trees, the size of the bootstrap sample, and the number of features considered at each node. The Random Forest performs well even under uncertain conditions which makes it easy to pick and model. The performance of the model depends on parameters we chose for hypertuning and they can be looped through and also with application of techniques like GridSearchCV we can estimate the best parameters for the model and utilize those. It has only few parameters to tune compared to that of models described in the next section like LSTM . Even though the increase in depth of trees in Random Forest improves accuracy it is not advisable in most of the scenarios because it with every level of depth increase the computation power increases drastically and time to run the model also increases [19]. We can also define number of estimators to be used during sampling as one of the parameters during definition of the model and by right combination of all these parameters we can construct a good split between the trees nodes.

Overall, random forest regressor is a powerful and versatile algorithm for regression tasks, particularly when dealing with high-dimensional data or noisy data with complex interactions between features. Its ability to generate multiple decision trees with different decision boundaries and to combine their predictions makes it a robust and reliable algorithm for a wide range of applications

3.6 LSTM Models

The Long Short Term Memory (LSTM) network is a type of Recurrent Neural Network (RNN) that is particularly useful for learning long-term dependencies in sequential time series forecasting. First introduced by Sepp Hochreiter and Jurgen Schmidhuber in Neural Computing, LSTM models have the ability to selectively remember or forget information from previous time steps, which makes them well-suited to problems that involve time step lags [18].

In time series forecasting, historical data patterns are analysed to create a collinearity that can be used to predict future values. RNNs are effective at this task, but they often encounter the problem of vanishing gradient, which makes it difficult to propagate output back to the model layer of input. LSTM memory cells can be used to overcome this problem, as they allow for differential memory that makes small modifications to data through the addition or multiplication of information during the flow of information through cell states.

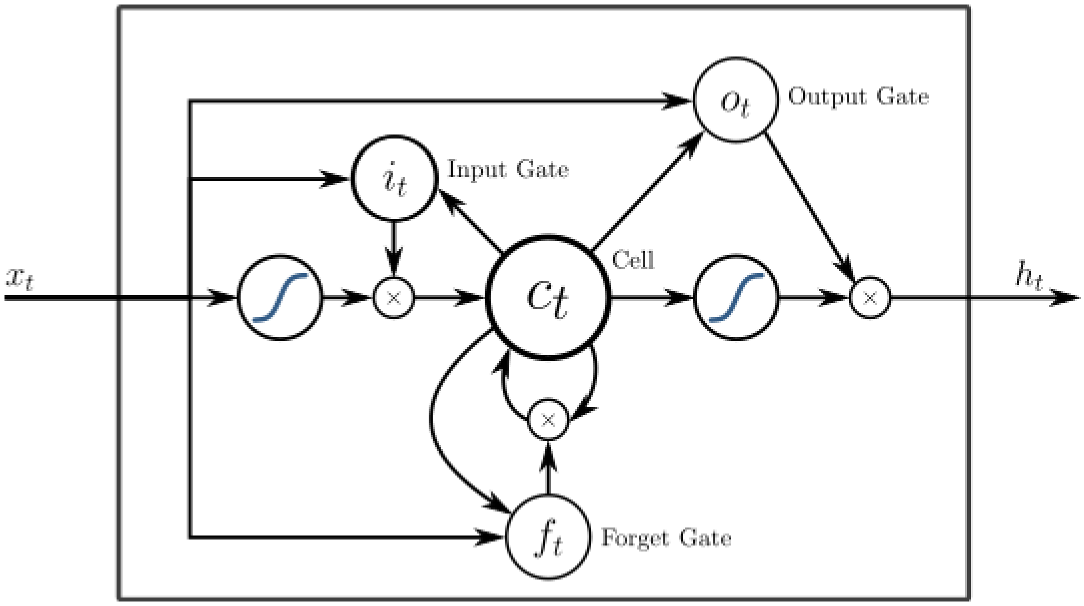


Figure 3.2 Long Short-Term Memory cell by Alex Graves under CC BY SA.

The LSTM architecture is particularly useful for time series analysis, where past observations from multiple time steps are used as inputs to the model. The process of incorporating these past observations into the model is known as time step lags [46]. For example, to predict the temperature at time t, temperature readings from the previous 5 time steps (t-5, t-4, t-3, t-2, t-1) might be used as inputs to the LSTM.

As shown in Figure 3.2 the LSTM model reads input(i_t) and passes the information through different cell states and can selectively remember or forget(f_t) information from previous time state based on the current input (C_t). This is achieved through the use of memory cells that can store information over multiple time steps and gates that control the flow of information into and out of the cells. To incorporate time step lags into an LSTM model, additional features are typically added to the input vector to represent past observations. In the temperature prediction example, the input vector at time t would include the temperature readings at time t-5, t-4, t-3, t-2, t-1, as well as any other relevant features such as time of day or day of the week.

The architecture of the LSTM model will depend on the specific problem and data at hand. Some common approaches include using multiple LSTM layers, adding additional regularization techniques such as dropout or weight decay, and using techniques such as teacher forcing to improve training stability.

3.7 Forecasting

Forecasting is a fundamental concept that constitutes the primary objective of this thesis. It entails comprehending the examination of historical data, making certain assumptions, and establishing the possible future values under similar conditions as those of the past. Time series forecasting is a distinct type of forecast in which the approach is through sequential periods. For instance, after analysing the sales data of the previous year, patterns and conclusions can be drawn, and presumptions can be made to determine the correctness of the thinking. This process facilitates decision-making regarding marketing, sales, and adjustments of prices for the current year, thereby enhancing the outcome. It is preferable to base decisions on thorough analysis rather than relying on intuition or gut feeling. Additionally, forecasting aids in comprehending the rationale behind any deviations from expected results, allowing for re-evaluation and better judgment in future decision-making. The forecasting process involves various techniques and steps that aim to improve the accuracy of results. All the preceding and subsequent sections of this thesis aim to enhance forecasting and improve the precision of the outcomes.

3.8 validation techniques

The validation of time series forecasting is a necessary step, as the accuracy of the model may be impacted by biases, overfitting, or underfitting [47]. In such cases, the true value of the forecast may differ significantly from the reality, rendering it unusable. However, the evaluation of time series forecasts is not a straightforward process, as it often requires significant amounts of time and memory and is particularly challenging due to the temporal nature of the data. In this thesis, we will explore various validation techniques, including cross-validation and repeated k-fold, to address these challenges. Cross-validation is a widely used technique for model validation in machine learning, and it can also be used for time series forecasting. In cross-validation, the data is split into multiple segments, or folds, and the model is trained on a subset of the data and tested on the remaining data. The process is repeated multiple times, with each fold acting as a test set once. The results are then averaged to provide an overall performance estimate.

Algorithm 1 Repeated K fold cross validation

Input: *D*: time series data set , *K*: number of iteration folds

Output: *O*: Model performance

```
function cross_validation(D,K)
    P ⟵ {}
    scores ⟵ {}
    for i in 1 to K do
        Randomly split D into K parts namely {$D_1$,$D_2$,….$D_K$}
        for d in D do
            $D_{test}$ ⟵ d
            $D_{train}$ ⟵ D-d  //all the remaining subsets
            model.fit($D_{train}$[features], $D_{train}$[target])
            Y ⟵ model.predit($D_{test}$[features])
            Save the Y to P
        end for
        Compare the P={ $Y_{1,}$ $Y_{2,}$$Y_{3,}$….$Y_k$} with orginal datset
        prediction
        Save the performance score of MSE or RMSE to scores
    end for
    return average(scores)
```

As defined in Algorithm 1 Repeated K-fold is a technique used for model validation that involves dividing the data into K subsets and repeating the process multiple times. In each iteration, K-1 subsets are used for training the model, and the remaining subset is used for testing the model. This process is repeated multiple times to obtain a more robust estimate of the model's performance. In time series forecasting, repeated K-fold can be used in a similar way to cross-validation. However, instead of using a rolling window, the time series data is divided into K subsets of equal size. The model is trained on K-1 subsets and tested on the remaining subset. This process is repeated multiple times, with the subsets being shuffled each time to provide a more robust estimate of the model's performance.

4. Problem Formulation

The present chapter tries to explicate the intricacies pertaining to the data that are the subject of examination in this thesis. The scope of this examination ranges from the raw data procured from diverse sources to the output data employed in the following chapters as inputs for model application. Furthermore, we also describe the pre-existing model utilized for pricing the room, along with a discussion of the need for a novel solution.

4.1 Problem Description

The hospitality industry is comprised of various resources, customers, roles, products, benefits, and services - both digital and physical. Each of these elements significantly influences the price of a room. This study focuses on hotels and resorts and how they determine pricing for their various rooms. The bookings for rooms in these establishments are received through several channels, including online travel agencies (OTAs) such as Booking.com and Expedia, client websites, direct check-ins, physical tour operators, and various other physical handbooks. All data is recorded in Excel files and stored in warehouses [21].

While OTAs have their individual analytics platforms and dashboards, which help in understanding the pricing and its effects, it does not provide entire picture of all the bookings for hotel. This is because the price of a room depends on all the bookings from the total, and not just one OTA. Therefore, a central analytics platform is necessary to process all the bookings that come through. The Figure 4.1 explains the structure of the data collection flow.

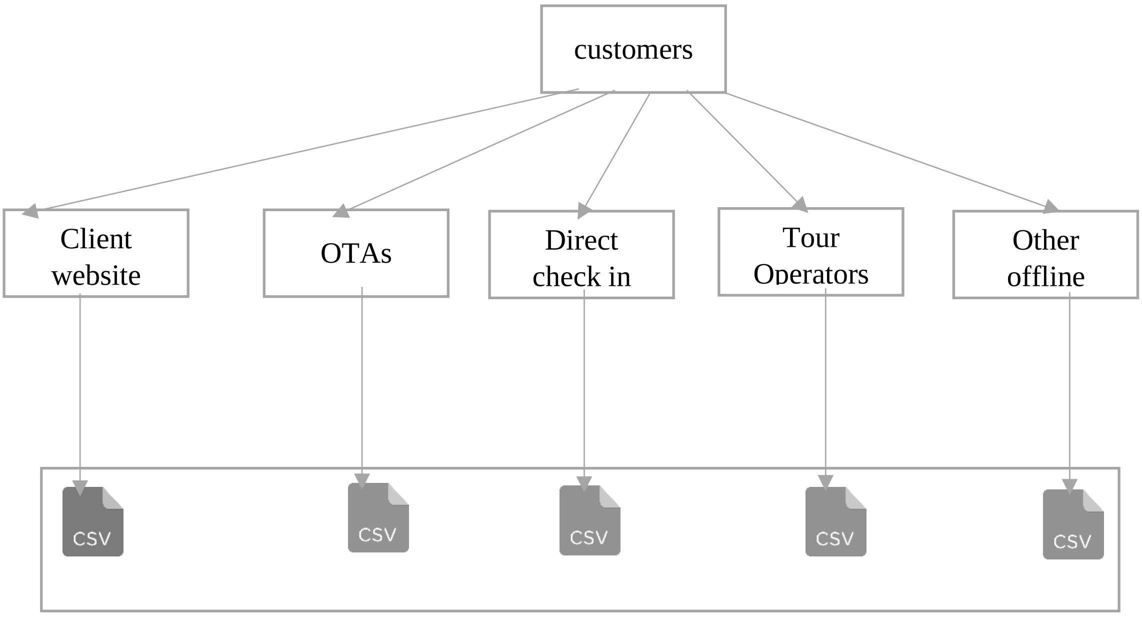


Figure 4.1 Data Collection Flow.

The data originates from various locations, with each CSV file containing bookings received from multiple sources (i.e., OTAs/channels). This data is received multiple times a day and each file comprise distinct columns, exclusive of the primary columns. Hence, unless we are conducting separate analyses for OTAs, the surplus columns are redundant, as our objective is to perform centralized analysis. It is advisable to eliminate these extraneous columns for the sake of efficiency and accuracy in our analytical endeavours.

Additionally, it is imperative to comprehend that the data pertaining to bookings includes a multitude of factors that contribute to the final price. These factors include the base price, additional services such as breakfast and bread, discount coupons, and various other hidden charges. As a result, it is not feasible to extract the base price directly from each Online Travel Agency (OTA). Therefore, it is necessary to devise a methodology for examining the prices and correlating them to the base prices in a meaningful manner. It is essential to note that although the prices appear to be continuous values in the data, they are derived from bar prices, which are predetermined levels assigned to each room. These bar prices are adjusted to modify the final price. Furthermore, it is imperative that price fluctuations remain within a reasonable range, as excessive variations may prompt potential customers to seek alternative vacation destinations. Hence, as a means of adhering to industry norms and maintaining pricing consistency, it is established that adjustments to prices shall be limited to a single bar increase or decrease, rather than several. This approach ensures that prices remain competitive without compromising the resort's market share or profitability. Moreover, it underscores the resort's commitment to providing a superior vacation experience for its patrons.

The data exhibits a multidimensional nature due to the fact that customers not only reserve bookings for the current day, but also make advance bookings for several days into the future, with the belief that such an approach would prove more cost-effective than making a last-minute booking. To regulate this trend, pricing is determined such that the rates for rooms are set lower for bookings that are made further in advance and gradually increase as the demand for the bookings grows and the dates of the bookings draw closer. However, this results in a significant problem between the customers and the owners of the resorts, as the latter are inclined to prioritize profit maximization and maintain market competitiveness by setting prices that are in line with the best available prices, as opposed to offering discounts. This issue shall be comprehensively addressed in subsequent chapters.

While the management of this situation may seem complex, it can be effectively navigated with a thorough understanding and in-depth analysis. However, a unique challenge arises when there is a lack of data for a specific period. As our data collection only dates to 2018, we had assumed that at least three to four years of data would provide a sufficient sample size for analysis. Unfortunately, the sudden outbreak of the COVID-19 pandemic has created a global catastrophe, resulting in the shutdown of all hotels and resorts.

Under normal circumstances, this might not pose a significant problem if there were no bookings scheduled during the period of shutdown, which is nearly two years in duration. However, due to the proactive nature of customers who often make advanced bookings, the hotel owner is required to keep bookings open, albeit not for the current period, but for the upcoming season that is six months away. Despite the lack of clarity regarding when the shutdown will end, the owner remains optimistic and maintains an open booking policy.

Furthermore, many customers are eager to leave their homes as soon as the shutdown is lifted, and hence they have made bookings for future nights with the option of cancelling them later. These factors pose unique challenges that must be addressed and managed effectively to ensure the continued success of the hotel.

The various scenarios that transpired during the pandemic had a significant impact, particularly once it concluded. Not only did we have to re-evaluate our models to account for the pandemic as an anomaly, but we also had to consider the altered mindset of users who had grown accustomed to extended periods of isolation at home. Following the initial phase after the pandemic, there was a considerable surge in bookings, leading to overbooking. Furthermore, several bookings made during the pandemic were subsequently cancelled by customers due to various reasons, adding to the challenge of adapting our model. As we anticipated that this was only a temporary situation, we needed to adjust our model to restore equilibrium as soon as possible.

Considering these developments, we recognized that existing solutions were no longer adequate and needed to be replaced with new solutions that could accommodate seasonality and trends. As a result, a series of experiments and proposed solutions were undertaken, which are discussed in the next chapters.

4.2 Existing Solution

The Online Travel Agencies (OTAs) have already implemented their individual analytics, while for central analytics, the concept of price elasticity of demand is utilized. In order to understand the characteristics of price with respect to changes in occupancy rate, the demand curve is drawn, and the price elasticity coefficient is examined. As per the law of demand, the quantity demanded of a good or service is inversely proportional to its price. However, due to the consideration of customer behavioural psychology and the case of luxury goods, this law appears to be uncertain since rooms with very low prices are considered to be of poor quality and those with very high prices are often overlooked by a certain class of customers. For this reason, it is widely believed that setting the appropriate price for a room at the correct time would result in an equilibrium between the price and demand.

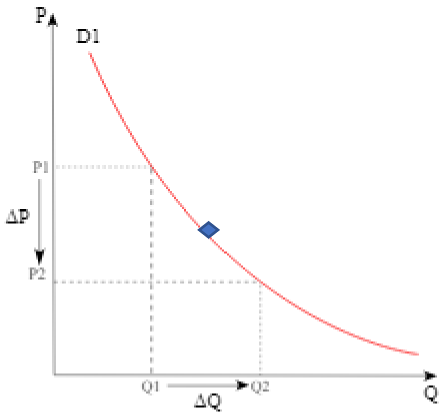


Figure 4.2 price elasticity of demand by Pawel Zdziarski under CC BY SA.

Figure 4.2 depicts the relationship between the quantity and price of a product or service. It shows that as the price of the product decreases, the quantity demanded increases. However, it is noteworthy that the primary objective of any business is not solely to increase the quantity but to generate revenue. This fact is also applicable in the case of resorts, where the goal is to maximize profit while maintaining an acceptable quantity of services provided.

In this context, it is important to find an optimal point, indicated by the blue point in the curve, where the profit can be maximized without compromising the quantity provided. The determination of this optimal point is facilitated by the concept of price elasticity of demand.

As show in Figure 4.2 for a given period of time the change in quantity from Q1 to Q2 is designated with △ Q and its price impact on y-axis and their difference from P1 to P2 is defined by △ P then

$$\textbf{Elasticity ratio} \;=\; \frac{\%\textbf{change in the quantity Demanded}(\triangle \mathrm{Q})}{\%\ \textbf{change in the price}(\triangle \mathrm{P})}\ .$$

The whole pricing strategy revolves around the concept of elasticity ratio which determines the sensitivity of demand towards price changes. This ratio is characterized by two distinct scenarios that have a significant impact on the pricing technique.

Firstly, when the elasticity ratio is greater than one, it implies that the demand is highly responsive to price changes, and even a slight alteration in price leads to a considerable impact on the quantity demanded. Secondly, if the elasticity ratio is less than one, it suggests that the demand is not affected by price changes.

Based on these two cases, it has been determined that on regular days, without any external factors influencing customer behaviour towards the night of stay, it is advisable to set the price at a minimal level to attract more customers. However, in the case of days where external events such as holidays, festivals, and weekends have a positive effect on bookings, the price is observed to be inelastic. Even with an increase in room rates, bookings still occur at an acceptable rate, thereby maximizing profits.

The goal of this thesis is to determine the occupancy rate for a particular check in date given historical data belonging to it. It can be formulated as a regression problem since occupancy rate is a continuous variable and historical data can be considered as input variables and occupancy rate is output variable. In terms of formal definition:

$$\mathrm{O_i} = f(x_{i1}, x_{i2}, x_{i3}, x_{i4} \dots x_{in}) + e \ ,$$

$\mathrm{O_i}$ is the occupancy rate for a given check in time in future be it tomorrow, next week, or next month

$\mathrm{x_{i1},x_{i2,\dots..}x_{in}}$ are features in the data listed in Table 6.1

e is the random noise which is not directly observed in data but observed during statistical analysis

f is the function that describe the relationship between input variables and the occupancy rate output variable.

In this thesis we model various algorithms explained in detail in Chapters 5 namely Ridge Regression, RandomForest, LSTM and best model is found based on features and the minimal errors obtained from the difference of actual of predicted values using several cross validation techniques explained in Chapter 3.8

5. Proposed Methodology

Due to the recent pandemic events and highly unpredictable demand, it was deemed necessary to adopt different methods. In this chapter, we will outline the steps that were taken to arrive at the final solution that is currently being utilized by our clients.

The first step in this process was to comprehend the various csv files that are received from the OTAs and align the data to the central database. While collecting and saving the data to the database, several challenges were encountered, which will be expounded upon in Section 5.1.

Next, we developed charts to identify patterns within the data, in order to determine the most appropriate model to employ. Since the price is not continuous and required an adjustment, we focused on determining the occupancy rate of each room night and adjusting the price accordingly. Through an extensive evaluation process, we identified a set of features that correlated with occupancy and could be utilized during modelling. We experimented with several algorithms that were well-suited to the chosen features, including Ridge Regression, LSTM, RandomForest and others.

We validated the accuracy scores by cross-testing them against individual resorts, as the data covers a chain of resorts. Ultimately, we selected the algorithm that yielded minimal error and a higher accuracy score, and created a hybrid algorithm that corrects the room rate based on the predicted occupancy rate.

We have also performed several experiments on different concepts like periodic cancellation rate, daily average return, revenue pickup compared to last year and monthly forecast and each of these are tested on global level as well as individual resort level too. The following algorithm provides a high-level description of our implementation.

Algorithm 2 Adjusting room rate with occupancy rate

Input: *D*: the dataset, *h*: hyper parameters of the model, C_{bars}: existing room rates(defined in categorical BAR(Best Available Rate) prices) , *P:* future time frame
Output: *DF*: the output dataset with adjusted room rates, *score*: accuracy and MSE

```
function   forecast_occupancy_rate (D, h, P)
        $D_p$ ⟵  preprocess(D)
        $X_{train,}$ $X_{test}$ ,$Y_{train}$  ⟵  ,$Y_{test}$   train_test_split($D_{p)}$
        model  ⟵  ML_algorithm(h)
        model.fit($X_{train}$ , $Y_{train}$ )
        $P_{train}$ ⟵  model.predict($X_{train}$ )
        Acc  ⟵  model.best_score_
        $score_{train}$ ⟵  root_mean_squared_error ($P_{train}$ , $Y_{train}$ )
        $P_{test}$ ⟵  model.predict($X_{test}$ )
        $score_{test}$ ⟵  root_mean_squared_error ($P_{test}$ , $Y_{test}$ )
        DF ⟵  model.predict(P)
        score ⟵  (Acc, $score_{train,}$ $score_{test}$ )
        return DF, score
occupany_rates,   ⟵  score forecast_occupancy_rate(D,h,P)
function  adjust_room_rate(occupancy_rates, $C_{bars}$)
       $P_{bars}$ ⟵  assign_bar_level_using_if_else (occupany_rates)
       for future_date in P do
              if  absolute($P_{bars}$ [future_date] – $C_{bars}$ [ future_date] >= 2)  and
              $P_{bars}$ [future_date] > $C_{bars}$ [ future_date] then
                     $P_{bars}$ [future_date] = $P_{bars}$ [future_date]+1
              else if  absolute($P_{bars}$ [future_date] – $C_{bars}$ [ future_date] >= 2)
              and $P_{bars}$ [future_date] < $C_{bars}$ [ future_date] then
                     $P_{bars}$ [future_date] = $P_{bars}$ [future_date]-1
       return $P_{bars}$
```

As stated in Algorithm 2 we have two separate functions during the whole process of adjusting the prices of the room. The first function forecast_occupancy handle the operations like reading the data from the files, processing and transformation for several feature columns and defining the model, splitting the data into train and test using techniques like train_test_split provided by scikit-learn and fitting the training data and validating it against test data. In order to avoid bias the model is run multiple times with different parameters and best performed model is selected using the GridSearchCV and finally the performance scores are measured using the RMSE(Root Mean Square Error) scores and the model with least score is utilized to forecast the occupancy rates for the prepared forecast data and passed these values to the other function adjust_room_rate. Which contains different if else conditions relating to the occupancy rate and adjusting BAR(Best Available Rates) prices from the provided forecast occupancy rates and current prices.

5.1 Data Pre-processing

During the course of our data analysis project, we encountered various challenges associated with the data files that we were working with. Each file had its own distinct structure and format, which required us to spend considerable time on several issues such as mapping the columns between the files and the database, converting numbers to their appropriate format as specified in the database, and even dealing with prices that were sometimes saved as text and had to be converted to float before saving them as Decimal in the database in order to maintain accuracy and long precision.

As the data in these files constituted a time series and contained numerous date and time columns, formatting the dates posed a major problem. This was because there was no particular standard for the format of dates. To address these issues, we made use of the pandas datetime processor, which enabled us to externally specify the format in which the data had been received from clients. By doing so, we were able to ensure that the data was accurately formatted and ready for analysis.

Furthermore, each booking in the data set had four distinct statuses, namely, ok, pending, suspended, and cancelled, which gave us insights into the behaviour of the customers. These statuses provided us with valuable information that allowed us to draw meaningful conclusions about the data and make informed decisions. We had duplicate rows in the data that arrived regularly. The only discrepancy between the two duplicate rows is the status columns, and if we were to keep the existing row and add the new row, it would produce inaccurate results during total computations. To resolve this issue, we had to pipeline a procedure for removing duplicate records every time a new data file is processed from the warehouse and before data ingestion is started on the database.

As the study was conducted in Italy, a non-English speaking country, some of the columns, such as names and text columns, contained Italian values with special characters that are not available in English. Therefore, we had to convert the data to utf-8 format to save it in the database. To accomplish this, we utilized all the available techniques and tools in Pandas for data processing and saved the final data in the database.

In conclusion, our preliminary analysis was marked by a number of challenges related to the structure and formatting of the data files we were working with. However, through the use of the appropriate tools and techniques, we were able to successfully navigate these challenges and draw meaningful insights from the data.

5.2 Data Insights

Initially, the dataset at our disposal encompassed reservations from 13 resorts belonging to a single chain, each of which received orders through 10 plus channels, both online and offline. In light of this, we resolved to investigate whether any discrepancies existed among the diverse categories. To achieve this, we undertook an analysis of the data from different perspectives, namely by establishing grouped index keys that allowed us to discern historical patterns. Subsequently, we were able to identify a significant number of patterns in customer behaviour towards bookings, both pre- and post-pandemic.

In order to understand the patterns found in the data, it is necessary to be familiar with certain terminology and pandemic history. The "status" column in the database defines the current situation of each booking with regard to customer behaviour. It is described using the following values:

- "Ok" (indicating that the booking has been done and confirmed)

- "Modified" (indicating that the booking has been made and corrected with some details such as date, age, time of stay, etc.)
- "Suspended" (signifying that the customer did not show up),
- "Pending" (describing a booking that has yet to be confirmed),
- and "Cancelled" (indicating that a booking was made but subsequently cancelled).

These statuses have been grouped into four categories, each with a specific meaning:

- T1 (Ok, Modified, Suspended, Pending, Cancelled) represents the total number of bookings in the database.
- T2 (Ok, Modified) represents the total number of bookings that have been confirmed.
- T3 (Pending, Modified, Cancelled) describes bookings that do not generate revenue.
- T4 (Cancelled) represents all cancellations.

We show these four T1, T2,T3, T4 categories as different histogram bars for various hotels in the following graphs.

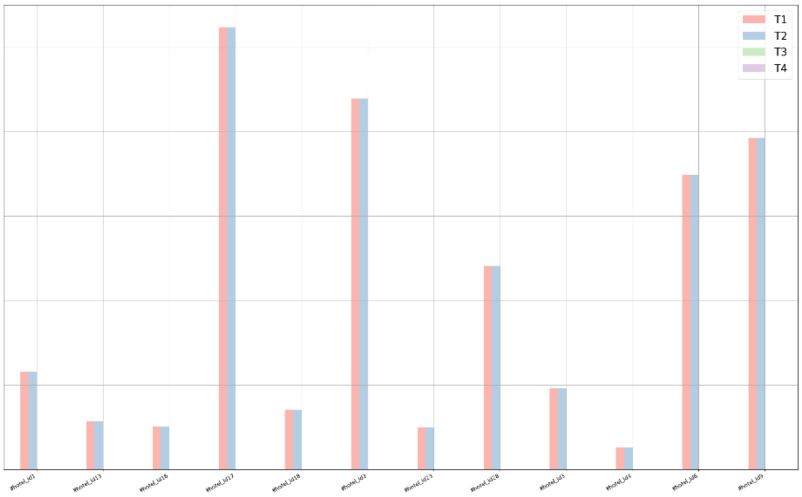


Figure 5.1 Bookings patterns before covid 2019.

The Figure 5.1 illustrates that T1 and T2 exhibit consistent patterns, with a negligible cancellation rate (T3) compared to confirmed bookings (T2). This data suggests that customer behaviour is readily predictable. This is an ideal expected situation where customers make reservation and they showed on the date of check in and so we do not need to worry mostly about the cancellation. However, with the onset of the pandemic, the situation has drastically changed, causing heightened levels of tension and fear among consumers. Consequently, their behaviour has become largely unpredictable. In addition, the government has implemented various sudden notifications and shutdowns, resulting in an entire year spent at home.

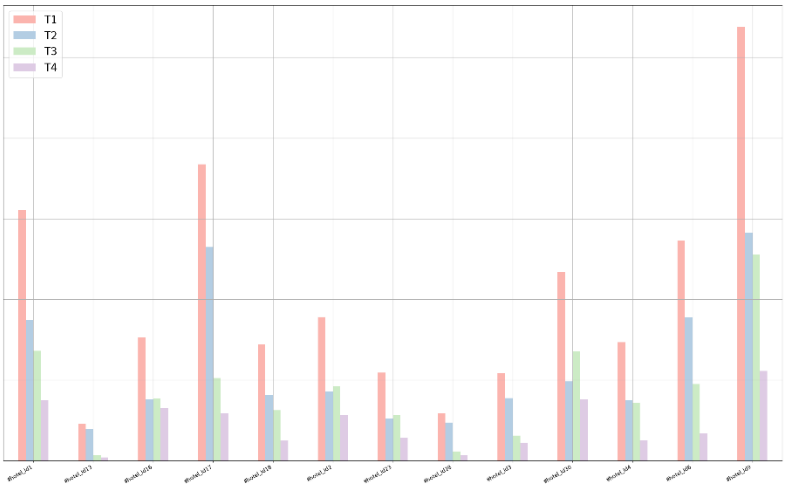


Figure 5.2 Booking patterns during covid 2021.

In the year 2021, the prevailing situation remained unchanged, and customers, despite their desperation, made bookings for the end of the season with the understanding that they could cancel later. Figure 5.2 clearly shows that customers made bookings despite the confusion about the lift of lockdown and status of pandemic hoping it would be cleared out soon. In doing so, customers hoped that normalcy would return and enable them to travel. Unfortunately, this was not the case, and most customers cancelled(T4) their bookings while some remained in a suspended state(T3-T4) difference of T3 and T4 bars. Additionally, the patterns observed were highly unpredictable since the T3 and T4 bars in Figure 5.1 are negligible but suddenly saw a spike in Figure 5.2 showing sudden deviation from historical data, rendering the date before pandemic useless for training and testing.

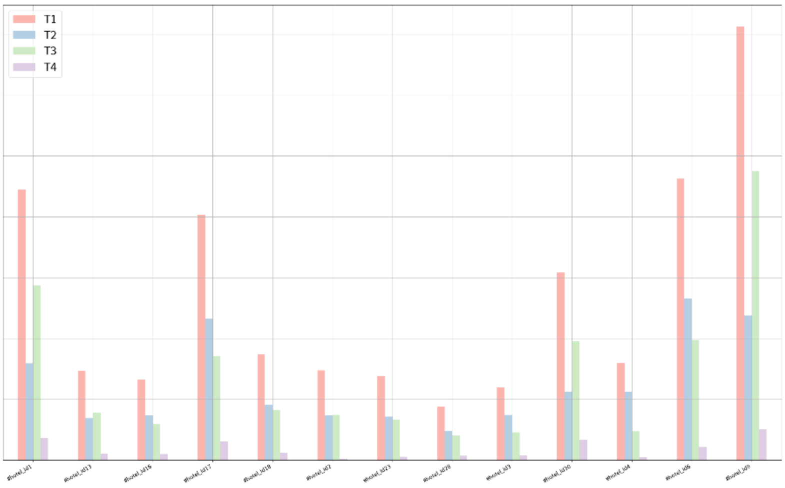


Figure 5.3 Booking patterns post covid 2022.

Upon the lifting of the Covid lockdown in 2022, it was expected that customer behaviour would revert to pre-pandemic patterns and was thought that plotted graph would be similar to Figure 5.1 in best case , but this assumption proved to be incorrect and patterns were not similar to both Figure 5.1 and Figure 5.2 but creating a whole new set of distinct patterns shown in Figure 5.3. In the above Figure 5.3 we can observe that difference between T3 and T4 is higher which means Customers exhibited constantly changing plans, and due to inflation and budgetary constraints, several customers cancelled their bookings. As a result, they can neither confirm bookings nor cancel bookings and left those in suspended or waiting states which is the meaning of T3-T4 from Figure 5.3.

To validate these theories a data set spanning four years was collected, including a year in normal conditions (2019), a year of lockdown (2020), a year of pandemic (2021), and a year of post-pandemic conditions (2022), each with distinct patterns within the data and tested under different periods and observed the patterns .

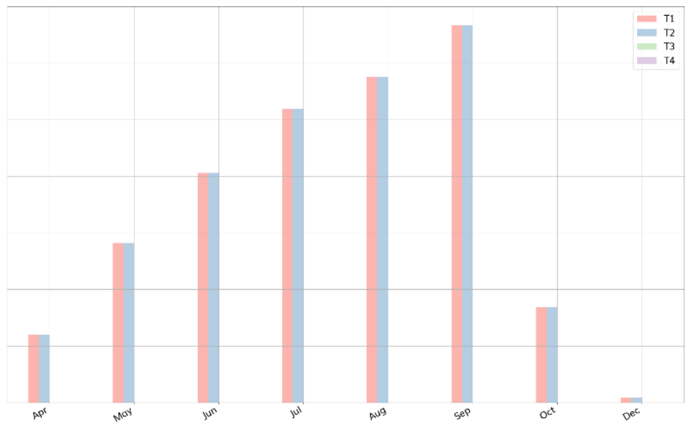


Figure 5.4 monthly booking date before covid 2019.

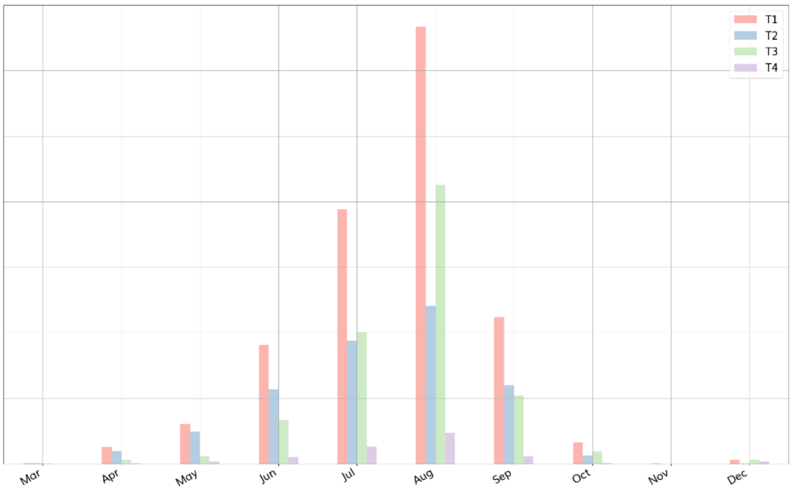


Figure 5.5 monthly booking data after covid 2022.

The Figure 5.4 and Figure 5.5 shows that the distortion in data related to bookings prior to the onset of the COVID-19 pandemic was not limited to the cancellations alone. The data was found to be skewed, thereby making it difficult to discern which period witnessed higher levels of bookings. Moreover, the impact of COVID-19 on the data has been less pronounced in terms of left-skewing. As a result, there has been limited opportunity to obtain a clear understanding of the booking trends during this period.

A correlation analysis was conducted to examine the relationship between variables and the total number of bookings, while taking into consideration other variables such as the type of room, online travel agencies (OTAs), and the hotel, as illustrated in the Figure 5.6. However, it was discovered that the data contained numerous gaps and was sparsely distributed across the timeline, thereby limiting the accuracy and reliability of the analysis.

Some of the main variable used for corelation analysis are:

Room_reference_type: unique room code defining the type of room (numeric code)

Total : total price paid by customer for a booking(decimal values)

Channel: source from where booking comes from (constant values)

Group_channel: Constant value defining whether booking is for a individual or bunch of people

Status : status of the booking record in the database (constant values)

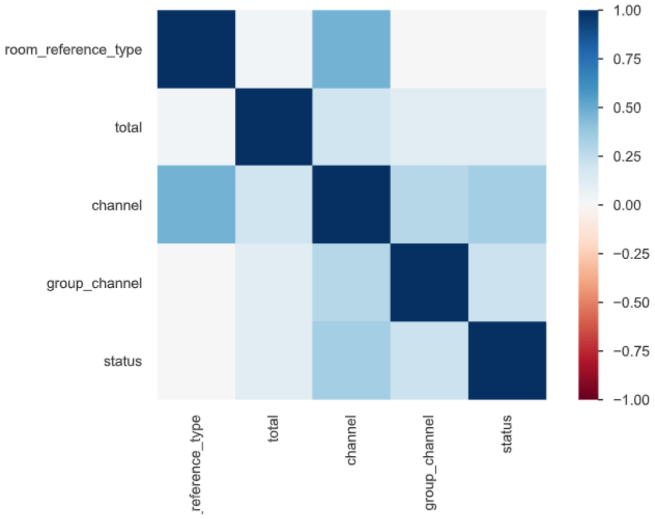


Figure 5.6 corelation heatmap among booking variables.

Consequently, a decision was made to develop a model that leveraged short-term data rather than long-term data to enable more accurate predictions. So, we started looking at short term timestep lags for changes in occupancy rates and surprisingly they seem to fit more than we expected.

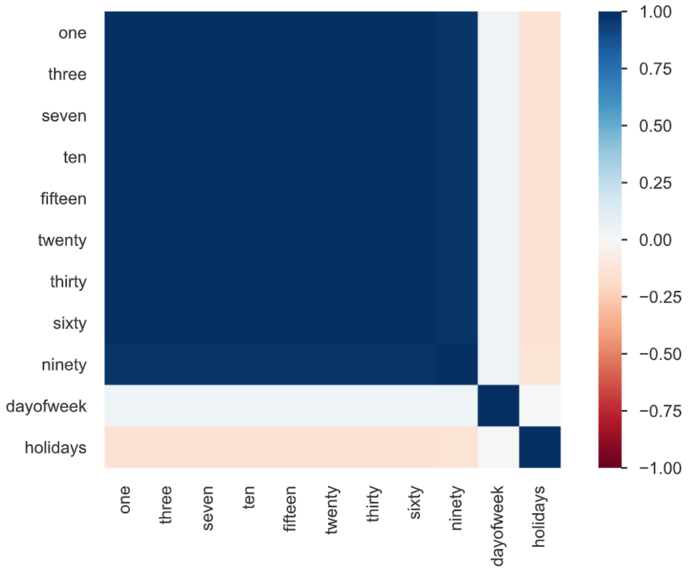


Figure 5.7 correlation heatmap of occupancy rate between various time lags.

We then tried to understand which are best timestep lags could perform well since all of them seem to fit well and as the lag goes far from day zero the correlation seems to become less.

The correlation matrix exhibited evidence that there exists high collinearity between lagged timesteps compared to that other independent variables shown in Figure 5.6. So we believed that these time steps could be a potential solution to our problem of occupancy rate prediction.

5.3 Data preparation

After conducting an extensive analysis, we have arrived at the decision to employ timestep lags as training data. The rationale behind this choice is the sparse nature of the data and the existence of independent patterns among data from historical periods. It is worth noting that the occupancy rate is not a direct column in our data, but rather a derived variable, necessitating the need to prepare the data ourselves.

Algorithm 3 Data preparation with occupancy rate

Input: *h*: list of hotels *c*: hotel room capacities, P: future time frame left in the season, timelags: list of timelags
Output: *DF*: the output dataset with occupancy rate with various timestep lags

```
function get_occupancy_rate (hotel_id, capacity, booking_dt ,stay_night,delta)
        Occ $\leftarrow$ SELECT ( occupancy rate data from database for given time booking_dt)
        return Occ
for dt in P do
    for lag in timelags do
        DF[dt][lag] $\leftarrow$ get_occupancy_rate( hotel_id, C[hotel_id],dt,dt,)
    End for
End for
return DF
```

As described in Algorithm 3 we create a new pipeline procedure which runs at predefined timeslots to prepare the training data for list of provided timelags. We can always adjust the number of timelags at any given point of time. We query the total bookings from the database along with capacity of the room from inventory table and calculate the occupancy rate for a given and timelag.

The data preparation process for all resorts under the chain was executed using the aforementioned algorithm, resulting in the creation of separate CSV files. However, due to the time step lags being farther from zero, the computation time and memory usage were significantly higher for larger lag values (such as lag=60) in comparison to smaller lag values (such as lag=30). As a result, a decision was made to limit the time step lags up to lag=30 in order to reduce processing time and memory usage.

The data in question is presented in decimal format, and it is expected that values should fall within the range of 0.0 to 100%. However, upon examination, it was observed that certain rooms had experienced overbookings resulting in occupancy rates exceeding 120%. Conversely, there were instances where room occupancy rates were extremely low, with percentages close to <5%. These findings suggest that the data is not normalized, which prompted the decision to normalize the data using the sklearn pre-processing package, with values being standardized to fall between 0 and 1.

5.4 Model Implementation

Our model utilizes a list of multivariate algorithms that utilize independent variables to predict target variables. One primary distinction in our algorithm is the number of time step lags used in the data to forecast the target output, which is the occupancy rate on day zero. It is important to note that the occupancy rate may either increase or decrease over time due to cancellations.

Initially, we utilized the Ridge Regression algorithm, which penalizes the results and checks for the total errors, along with a penalized cost function explained in section 3.5.2. This method enabled us to determine the best time lags and subsequently use them in both training and testing data. The model was then able to return the occupancy rate of day zero. After multiple rounds of cross-validation testing using the Algorithm 1 describe above, we selected the best model and applied the final obtained results in the Algorithm 2 described in Section 5.1.

To further improve the accuracy of the results, we performed several hyper-tuning of parameters for the model and cross-validated each of them. The final results obtained after applying Algorithm 2 were the adjusted room rates, which were saved in the database for future usage.

Once the model was fully prepared, we deployed it on an AWS EC2 instance and set cron jobs to run and update the model every week. We assumed that we would need at least a week to make any changes to the model due to data gathering and the frequency of changes to reflect.

Algorithm 4 Weekly Model Updates

Input: *train*: train dataset , *test*: test dataset, *model_file_name*: previously saved model in pickl format

```
function EvaluateModelWeekly (train,test,P,model_file_name)
        model ⟵ load_model_from_file(file_name)
        model.fit(train , test )
        Pred ⟵ model.predict(P)
        model.save(model.pickl)
        return Pred
```

6. Experiments

This chapter provides a detailed account of the experiments we have conducted and the corresponding results. As stated in the previous chapter, the cost of a room in a resort is influenced by various factors. Our main objective was to forecast changes in occupancy rates and make necessary price adjustments accordingly. In addition, we conducted various other experiments related to booking statuses, considering that there are several types that can impact the price. To achieve our goals, we employed a range of algorithms to predict the percentage of bookings associated with particular booking statuses. This helped us to anticipate situations of overbookings and under bookings, allowing the resorts to make necessary preparations. The first six months of our project were devoted to understanding, analysing, and training the model that was presented in Chapter 5. The remaining months were dedicated to model updating, validation, and experimentation with additional data.

6.1 Dataset Overview

Despite having a limited amount of data at our disposal, we were able to achieve commendable performance owing to the fact that the data emanated from a chain of 13 resorts, centrally located in popular tourism destinations. This positioning accorded us an advantage in terms of the high frequency and volume of bookings. Consequently, we were able to access a substantial quantity of data on a daily basis, ranging from 1000 to 3000 bookings per day. The data set spans a duration of three years, from 2019 to the present. Further information on the input and output variables can be found in Table 6.1 and Table 6.2, respectively.

Table 6.1 Description Of Input Variables(num= numeric , string = descriptive, alphanum= alphanumeric, bool= Boolean, float= decimal value).

List of Input Variables	
Column Name	Description
id_hotel (num)	Unique Identifier for a hotel/resort
id_booking (alphanum)	Unique identifier of a booking
booking_date(Date)	Date of the booking
booking_from(Date)	Customer chose date of check in
booking_to(Date)	Customer chosen date of checkout
Status (constant)	Booking status
Adult(num)	Number of adults
Children(num)	Number of children
Price(double)	Price of the room in decimal format
Promotion(string)	Type of promotion applied if any
Room_reference_type(num)	Unique identifier for Type of room
Room_reference_service(constant)	Additional services like B&B , meals etc
Channel(string)	Source of the bookings(OTAs)
Night(date)	Customer chosen date to spend at the hotel
Room_count(num)	Number of rooms

Continuation of Table 6.1	
Group_channel(constant)	Constant value defining whether booking is for a individual or bunch of people
Capacity(num)	Room capacity for that particular room_reference_type
Current_bar_price(float)	Current base price set for that room decimal format
Derived Variables for night customer choose to check in the resort	
Holiday(bool)	Boolean value defining whether the customer staying night is holiday or not
Week_day(num)	Numerical value representing the day of the week
Weekend(bool)	Boolean value determining sat or sun necessary since they have high bookings theoretically
Month_edge(bool)	Boolean value checking if customer checkin is on first two or last two days of the month
Special_day(bool)	representing if there is any external events
Occupancy_rate(float)	Percentage of occupied capacity for a given room

Table 6.2 Description of Output Variables(num= numeric , string = descriptive, alphanum= alphanumeric, bool= Boolean, float= decimal value).

List of Output Variables	
Predicted_occupancy_rate(float)	Percentage of room capacity that can be occupied in the future
Adjusted_bar_price(float)	Room price after applying algorithm and business rules
Channel_revenue(float)	Forecasted Revenue generated from an OTA
Room_revenue(float)	Forecasted revenue for individual room type
Hotel_revenue(float)	Forecasted revenue for individual hotel

6.2 Procedure

Given that our dataset has multiple dimensions, with each booking date having bookings for several nights and receiving bookings from various sources, searching for transactions in the database requires the use of multiple index keys. Therefore, we conducted experiments on various levels of grouped indexes. Initially, we examined the daily booking patterns and revenue flow for individual hotels, and subsequently developed Key Performance Indicators (KPIs) to comprehend the daily sales changes. One of the KPIs we used was ADR (Average Daily Returns), which is computed as follows:

$$\text{ADR} = \frac{(\text{total revenue generated in a period})}{(\text{total rooms booked in the same period})}\ .$$

Our observations indicate that the Average Daily Returns (ADR) exhibit a volatile daily change. Specifically, we observed ADR fluctuations between 0-1% on some days, while on other days, they can rise up to 10-20% before returning to normal levels the following day. We endeavoured to comprehend the rationale behind this trend by scrutinizing data from comparable days, such as yesterday and the same day last year. However, as elaborated in Chapter 5, the pandemic has resulted in a dearth of data or invalidated established patterns. Due to the requirements of our clients and the shortage of data, we had to develop a model that could function well and adjust to the market, predicting trends and forecasting occupancy rates that would be close to reality (either similar to the existing daily occupancy rate or with slight changes if there were trends with a valid justification for the change). Since we lacked data, we could only do short-range forecasting for up to 3 to 4 weeks in the future, utilizing past data for predictions. We utilized the bookings from the previous month to predict the current month's occupancy rate, and we continued this process for each month as new bookings arrived every day.

Given that we had a group of indexes separating each transaction in the database, we had several options to utilize the occupancy rate by fixing on an index or a subset of indexes. We had mainly three possibilities with the available indexes (booking_date, check_in_night(also referred to as column night in the database), channel) as follows:

1. We could have used booking_date as an index to predict the daily occupancy rate, but we agreed not to due to the volatility of ADR that was explained earlier. Furthermore, this approach reduced the amount of data we had compared to the other two approaches.
2. We could have used channel and booking_date as indexes to forecast the occupancy rate. Although this approach would also reduce the amount of data, it was not much compared to approach 1. However, we decided not to proceed with this approach because existing OTAs are already providing these insights on their platforms.
3. We could have utilized the check-in night as an index to study the historical bookings belonging to that night and forecast the future occupancy rate for the same check-in night.

We decided to proceed with the third approach because it has more benefits compared to the other two approaches. It certainly increases the training and test data for the model. The ADR change with this index is less compared to the first booking_date index. Additionally, it provided us with unintended useful insights for the owners to understand whether the rooms will be fully booked or not for a particular night. Furthermore, marketing teams can take advantage of these insights and focus their efforts on specific nights with promotions.

In order to derive the occupancy rate, we developed an algorithm (Algorithm 3) which is explained in Chapter 5. The input data table was prepared using time lags as columns and night customers chose to check in as the index. Our objective was to establish a relationship between period trends for each check-in night in the season by considering past days. This approach enabled us to obtain data and make it less complex compared to other possibilities.

Figure 5.7 shows that as the distance between the check-in night and booking day becomes large, the correlation seems to reduce. After conducting several trial and error experiments, we selected seven different time lags within the last 30 days, which were one, three, seven, ten, fifteen, twenty, and thirty. The output was day zero, which represents the current day occupancy rate of the check-in night.

To take into account the impact of bookings, we also incorporated two external parameters, namely day of the week and holidays. Finally, we stored the prepared data along with all these variables in csv files named after the hotels.

6.3 Model Training

A pipeline was implemented to process the training data in accordance with the guidelines provided in Chapter 6.2. The data was partitioned into training and test sets using the train/test split method. Subsequently, a ridge model was created, and all its parameters were supplied to the GridsearchCV technique of the sklearn package to determine the optimal parameters for the model. The damping factor and k-fold split values were provided as parameters to GridsearchCV, and the model was executed in a loop to identify the best parameters that produced a good score. The lambda values were examined in a range of 0 to 1 with a precision of digits (i.e., 0.00, 0.01, 0.02, ..., 0.99, 1.00). The scoring mechanism utilized was the "neg_mean_absolute_error" of the sklearn package, as this was a regression problem. The experiment was repeated several times with a constant random state to enable comparison of the results.

After preparing the model, it was fit against the training data and evaluated for errors and performance using the Root Mean Square Error (RMSE). The model was subsequently leveraged to forecast the behaviour of unknown data in the test set. The RepeatedKFold technique with 10 splits was employed to rigorously test the model by looping through subsets of the data split during the process. The best model was returned, which was used for training and testing. Once the RMSE scores were generated for all the models we picked the one with least values and the model was employed to forecast for future periods, such as the next 7 days.

The experiment was repeated using two additional models, namely the RandomForestRegressor and the LSTM. Both models were selected for their ability to handle the type of data under consideration, with the RandomForestRegressor being an ensemble of decision trees on subsamples of data and the LSTM being a neural network model capable of storing and updating information using cell states and ignore gates.

Similar to the ridge model, the data was split using the train_test_split technique and the RandomForestRegressor was defined, with its parameters such as the number of estimators and the maximum features selected using the square root technique from sklearn. The depth of the trees generated was limited to three levels to ensure efficient computation. The RepeatedKFold and GridsearchCV techniques were used for cross-validation, and the RMSE scores for each hotel were stored for future comparison with other models.

For the LSTM model, the two-dimensional data was first converted to three-dimensional data using a defined pipeline before being passed to the model. The model was defined with basic parameters, such as one dense layer and 50 epochs, and a custom RMSE loss function was defined to facilitate result comparison. The ‘rmsprop’ optimizer was applied given the data being numerical and continuous. The data was fitted to the input side and then passed through the dense layers before being inverted to return to the two-dimensional data format. The LSTM model required more computation, time, and complexity compared to the other models.

In summary, the experiment involved the use of three models: the ridge model, the RandomForestRegressor, and the LSTM. Each model was selected based on its ability to handle the data under consideration, and various techniques were employed, such as cross-validation and custom loss functions, to ensure accurate and consistent results. The results from each model were stored for future comparison and analysis.

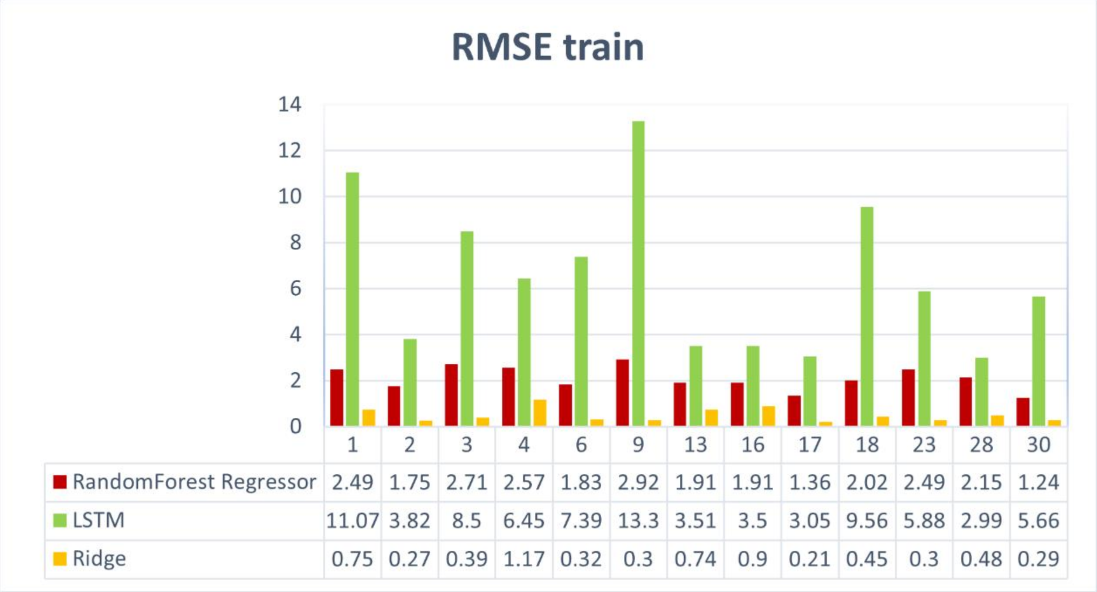
6.4 Results

As explained in Chapter 6.2 we used third approach from the available options(which is taking the time lags i.e. occupancy rate of the past days for a given check in night as the input data and the day zero occupancy rate as the target variable). We focused on performance of the models that had already been created. Each of these models included several hyper parameters, and we tested various combinations of these parameters in order to identify the best settings. We repeated the testing process multiple times and set constraints on resources and time to run the models. For example, we tried to increase the depth of the random forest tree up to 1000 levels, as doing so we have improved accuracy by only a small percentage at a significant cost in terms of time and computing resources.

Despite running models once a week, it still took hours to generate results due to the large number of hotels and nights involved. Furthermore, the random forest model suffered from bias when averaging results from all subtrees due to the limited number of variables included in some trees. Similarly, increasing the number of dense layers in the LSTM model would have required processing and understanding more information, putting us in a similar situation as with the random forest model. In contrast, the ridge model is a type of supervised learning algorithm that can interpolate trends better than other models. As a result, we expected the results from this model to be favourable, as was ultimately the case.

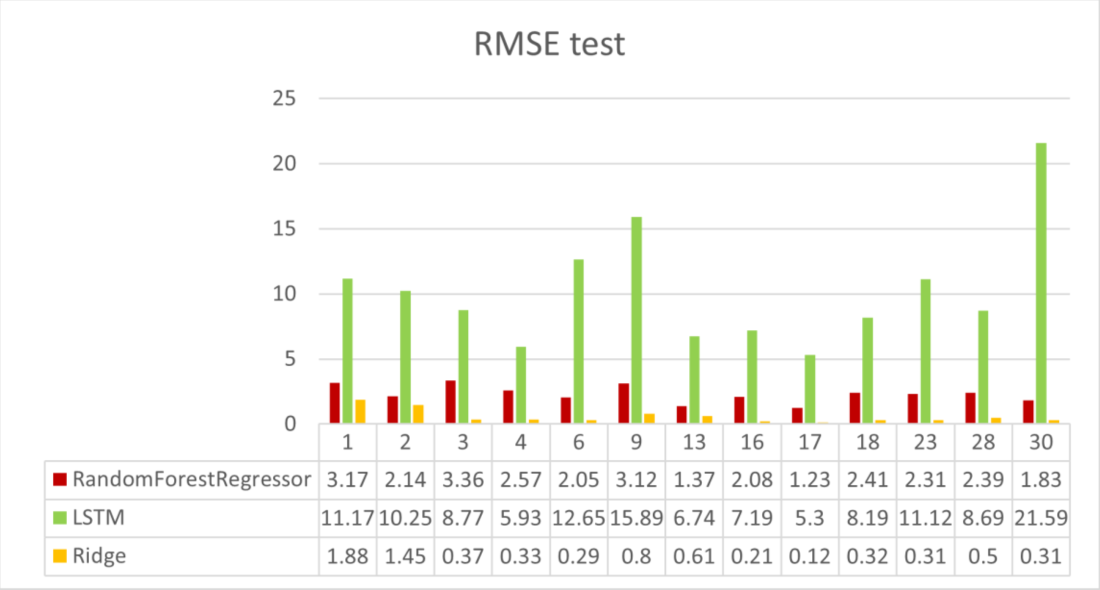
After testing the 3 models and getting the RMSE scores for all the models we compared them in order to understand which works better. The results were shown in figures below related to different hotels and their RMSE score with respect to training and test dataset.

For better understanding of the results we plotted the RMSE score of all the 3 models using the graph representation using hotel number on the x-axis and RMSE score on the y-axis as shown in Figure 6.1 and Figure 6.2 below.



	1	2	3	4	6	9	13	16	17	18	23	28	30
🟥 RandomForest Regressor	2.49	1.75	2.71	2.57	1.83	2.92	1.91	1.91	1.36	2.02	2.49	2.15	1.24
🟩 LSTM	11.07	3.82	8.5	6.45	7.39	13.3	3.51	3.5	3.05	9.56	5.88	2.99	5.66
🟨 Ridge	0.75	0.27	0.39	1.17	0.32	0.3	0.74	0.9	0.21	0.45	0.3	0.48	0.29

Figure 6.1 RMSE training dataset scores.



	1	2	3	4	6	9	13	16	17	18	23	28	30
🟥 RandomForestRegressor	3.17	2.14	3.36	2.57	2.05	3.12	1.37	2.08	1.23	2.41	2.31	2.39	1.83
🟩 LSTM	11.17	10.25	8.77	5.93	12.65	15.89	6.74	7.19	5.3	8.19	11.12	8.69	21.59
🟨 Ridge	1.88	1.45	0.37	0.33	0.29	0.8	0.61	0.21	0.12	0.32	0.31	0.5	0.31

Figure 6.2 RMSE scores on test dataset.

To verify the model performance, we used K fold cross-validation as we thought our model might be biased if we did the manual test once or twice and compared results. Setting a random seed might result in the same outcomes, so we decided to use K fold cross-validation.

After conducting an analysis of the results obtained from Figure 6.1 and Figure 6.2, we have analysed the results in two ways:

Firstly, we compared the root mean square error (RMSE) scores among all the models for both training and test data. We found that the RMSE scores were almost similar or slightly higher in all the models, except for a few hotels such as hotel 2, hotel 13, and hotel 28. These hotels showed large differences between their training and test data scores. However, we know that these hotels did not receive many bookings from the available data, and they were also closed both online and offline for most of the time during the pandemic. Despite other hotels showing close differences in RMSE scores, we could only verify the results after obtaining more data. As the model is updated continuously, we decided to accept the results and move forward.

Secondly, we compared the RMSE scores among different models and found that each model had its unique ability to process information involving correlations, branching, state changes, and partial information storing and updating the rest like a memory cell. However, we observed that simple models like Ridge performed better than complex models like LSTM and RandomForest Regressor. We assume that this is due to the lack of data to process in LSTM memory states or in splitting enough trees in RandomForestRegressor. Therefore, we decided to keep the Ridge model as primary for the time being and take the next steps while validating the other models periodically and utilizing them as soon as they improve accuracy.

Furthermore, it was observed that the Ridge model outperformed the RandomForestRegressor and LSTM models. In addition, the Ridge model exhibited a lower time complexity and was easier to build and update. Thus, the decision was made to implement the entire system pipeline using the Ridge model. The Ridge model was updated on a weekly basis by fitting new data and storing the resulting models in pickle files for each individual hotel. The output obtained through the forecast was subsequently passed through Algorithm 2, resulting in the generation of final adjusted room rates. These adjusted rates were then stored in a MYSQL database for further use. The models were stored on AWS and scheduled to run once a week, taking into account the frequency of bookings.

7. Conclusion & Future Works

This thesis explores a specific case of dynamic pricing within the hotel/resorts industry, focusing on data scarcity during times of crisis. Despite having three years of data, the pandemic and inconsistencies in the data during the last two years before the study prevented its use. The impact of COVID-19 and new inflation would also continue to affect the industry for at least the next year or two. Initially, a model with synthetic data similar to that generated before the pandemic was proposed, but after discussion and consideration of the severity of the pandemic and news information, a model was developed to adapt to real-time data during the pandemic period where traditional seasonality and trend analysis failed. Several analysis techniques were explored before selecting the occupancy rate of the rooms as the input variable for a hybrid model that adapts to market demand and business rules set by clients. For each night in the season, the model considers the occupancy rate during previous time steps and forecasts the current day occupancy rate, with a continuous update of the model in real-time.

The study involved the evaluation of several regression models, namely Ridge regression, RandomForestRegressor and LSTM models, in order to predict the occupancy rate. Based on the Root Mean Squared Error (RMSE) scores and the computational efficiency, the Ridge model was chosen as the best option for the prediction task. The predicted occupancy rates were then subjected to certain business rules defined by the clients, including a function that contains various if-else conditions to set the initial occupancy rate. The rate was then adjusted based on bar/level up or down to the current price, and the process was repeated on a weekly basis.

Additionally, external factors such as the day of the week and holidays were taken into consideration as they were believed to have a significant impact on booking rates. Special pricing was therefore applied during such periods, with rates set slightly higher than the regular price. It is noteworthy that these factors were determined based on experience and intuition of the clients.

Due to the limitations imposed by both data and business considerations, and the timing of the current season, we have been unable to test all potential theories for improving revenue. However, we believe that by combining certain behavioural tests with pricing techniques, we can increase revenue. Specifically, we propose increasing room prices during special days and periods of high demand at the start of the season when check-in dates are distant, as this will allow the model to adjust prices based on occupancy rates over time. By implementing this pricing strategy at the start of the season, we could have improved profits.

Furthermore, while our analysis has focused on short-term occupancy rates, we anticipate that next season we will have access to an abundance of data that will allow us to incorporate seasonality and long-term trends into our revenue optimization efforts. Finally, although we encountered some challenges in accessing information related to promotions and other variables while processing raw data, we believe that by combining pricing and promotional strategies, we can gain further insights into customer behaviour and improve revenue.

Bibliography

1. **Sangwon Park, Yizhen Yin, Byung-Gak Son.** Understanding of online hotel booking process: A multiple method approach. s.l. : Journal of Vacation Marketing, June 2018.

2. **Christiane Barz, Simon Laumer, Marcel Freyschmidt, Jesús Martínez Blanco.** Discrete dynamic pricing and application of network revenue management for FlixBus. s.l. : Journal of Revenue and Pricing Management, November 2021.

3. **Costa, Joan Carles Cirer.** Price formation and market segmentation in seaside accommodations. s.l. : International Journal of Hospitality Management, 2013.

4. **Anderson, C.K. and Xie, X.** Dynamic pricing in hospitality: overview and opportunities. s.l. : Journal of Revenue Management, 2016.

5. **Petricek, M., Chalupa, S. and Melas, D.** Model of Price Optimization as a Part of Hotel Revenue Management—Stochastic Approach. s.l. : Mathematics journals, July 2021.

6. **Gu, Zheng.** Proposing a room pricing model for optimizing profitability. *Journal of Hospitality Management.* 1997. Vol. 16, 3.

7. **Armstrong, M.** Competition in two-sided markets. s.l. : The RAND Journal of Economics, 2006. Vol. 37, 3, pp. 668–691.

8. **Rama Yelkur, Maria Manuela NeÃveda DaCosta.** Differential pricing and segmentation on the Internet: the case of hotels. 2001.

9. **Croes, R., & Semrad, K. J.** Does Discounting Work in the Lodging Industry? s.l. : Journal of Travel Research, 2012.

10. **Kefela, Mehari Semere.** Determinants of Hotel Room Rates in Stockholm: A Hedonic Pricing Approach. s.l. : Thesis, 2014.

11. **Graziano Abratea, Juan Luis Nicolaub , Giampaolo Vigliac.** The impact of dynamic price variability on revenue maximization. s.l. : Journal of Tourism Management, 2019.

12. **Graziano Abrate, Giovanni Fraquelli , Giampaolo Viglia .** Dynamic pricing strategies and customer heterogeneity: the case of European hotels. s.l. : Working paper, July 2010.

13. **Andrei M. Bandalouski, Natalja G. Egorova, Mikhail Y. Kovalyov, Erwin Pesch · S, Armagan Tarim.** Dynamic pricing with demand disaggregation for hotel revenue management. s.l. : Journal of Heuristics, 2021.

14. **Abd El-Moniem Bayoumia, Mohamed Salehb , Amir F. Atiyaa and Heba Abdel Aziz.** Dynamic pricing for hotel revenue management using price multipliers. s.l. : Journal of Revenue and Pricing Management, 2013.

15. **Pilar Talon-Ballestero, Marta Nieto-García , Lydia Gonzalez-Serrano.** The wheel of dynamic pricing: Towards open pricing and one to one pricing in hotel revenue management. s.l. : International Journal of Hospitality Management, Feb 2022.

16. **Andrea Guizzardi, Luca Vincenzo Ballestra , Enzo D'Innocenzo .** Hotel dynamic pricing, stochastic demand and covid-19. s.l. : Annals of Tourism Research, 2022.

17. **Leroy, Jean Pierre SignoretAlain.** Monte Carlo Simulation. 2021. pp. 547-586.

18. **Qi Tang, Tongmei Fan,Ruchen Shi,Jingyan Huang,Yidan Ma.** *Prediction of financial time series using LSTM and data denoising methods.* Nanjing,China : School of Economics and Management, Southeast University , 2021.

19. **Hristos Tyralis, Georgia Papacharalampous.** Variable Selection in Time Series Forecasting Using Random Forests. 2017. Vol. 114.

20. **Rob Hallak, Ilke Onur, Craig Lee.** Consumer demand for healthy beverages in the hospitality industry: Examining willingness to pay a premium, and barriers to purchase. s.l. : PLOS ONE Research Article, May 2022.

21. **Unji BAEK, Youngseok SIM , Seul-Ki LEE.** Analysis of Hierarchical Competition Structure and Pricing Strategy in the Hotel Industry. s.l. : Journal of Asian Finance, Economics and Business, 2019. Vol. 6, 4.

22. **HAYNES, Natalie.** The evolution of competitor data collection in the hotel industry and its application to revenue management and pricing. s.l. : Journal of Revenue and Pricing Management, 2016. Vol. 15.

23. **Seung Hyun Lee, Robertico Croes, Manuel Rivera.** Exploring the role of human judgment in making discount decisions in the lodging industry. s.l. : Journal of Hospitality Financial Management, 2015. Vol. 23, 1.

24. **Andres-Martinez, et al.** Analysis Of Hotel Internet Booking Users. 2014. Vol. 13, 7.

25. **Harnjo, Edward, Simamora, Javerson and Rotua Hutabarat, Linda.** IDENTIFYING CUSTOMER BEHAVIOR IN HOSPITALITY TO DELIVER QUALITY SERVICE AND CUSTOMER SATISFACTION. 2009. Vol. 2.

26. **Moro, Sérgio Rita, Paulo Oliveira, Cristina.** Factors Influencing Hotels’ Online Prices. 2018. Vol. 27, 4, pp. 443-464.

27. **M, Taufik, F, Renaldi and R, Umbara F.** Implementing Online Analytical Processing in Hotel Customer Relationship Management. 2021. Vol. 1115, 1.

28. **Singh, P Amanpartap KHAIRA, JS.** A comparative review of Extraction, Transformation and Loading tools. 2013. Vol. 4, 2.

29. **Van Rossum, Guido and Drake Jr, Fred L.** Python reference manual. s.l. : Centrum voor Wiskunde en Informatica Amsterdam, 1995.

30. **Naoki, Inada.** pymysql. s.l. : Python Software Foundation, 2021.

31. **McKinney, Wes and others.** Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference.* Austin, TX : s.n., 2010, pp. 51-56.

32. **M, Shereni NChambwe.** Hospitality Big Data Analytics in Developing Countries. 2020. Vol. 21, pp. 361-369.

33. **ALOTAIBI and EID.** Application of Machine Learning in the Hotel Industry: A Critical Review. 2010.

34. **Kim, T. H., and A. White.** On more robust estimation of skewness and kurtosis: simulation and application to the S&P500 index. s.l. : Finance Research Letters, 2014. Vol. 1, pp. 56-70.

35. **Sop, Serhat Adem.** The Effect of Market-Oriented and Brand-Oriented Service Improvement on Hotel Performance. 2021. Vol. 9, 1, pp. 2147-9100.

36. **Meghan K. Cain, Zhiyong Zhang , Ke-Hai Yuan.** Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. 2017. Vol. 49, pp. 1716–1735.

37. **Lendasse, Francesco Corona and Amaury.** *Variable Scaling for Time Series Prediction.* s.l. : Helsinki University of Technology - Laboratory of Computer and Information Science.

38. **K.A. Venkatesh, Dhanajay Mishra and T. Manimozhi.** Model selection and regularization. *Statistical Modeling in Machine Learning.* 2023, pp. 159-178.

39. **Olivier Sigaud, Camille Salaun, vincent padois.** On-line regression algorithms for learning mechanical models of robots: A survey. 2011. Vol. 59, pp. 1115-1129.

40. **Lijuan Huang, Guojie Xie, Dahao Li, and Chunfang Zou.** Predicting and Analyzing E-Logistics Demand in Urban and Rural Areas: An Empirical Approach on Historical Data of China. 2018. Vol. 14, 7.

41. **Snee, Donald W. Marquardt and Ronald D.** Ridge Regression in Practice. 1975. Vol. 29, 1.

42. **Antoni Wibowo, Inten Yasmina, Antoni Wibowo.** Food Price Prediction Using Time Series Linear Ridge Regression with The Best Damping Factor. 2021. Vol. 6, 2, pp. 694-698.

43. **Ertuğrul Egemen, Baytar Zakir,Çatal Çağatay,Muratli, Can.** Performance tuning for machine learning-based software development effort prediction models. 2019. Vol. 27, 2.

44. **Breiman, L.** *Random forests* . s.l. : Machine learning, 2001. pp. 5-32.

45. **Nielsen, Richard A. Davis and Mikkel S.** Modeling of time series using random forests: Theoretical developments. 2020. Vol. 14, pp. 3644-3671.

46. **Ricardo P. Masini, Marcelo C. Medeiros,Eduardo F. Mendes.** *Machine Learning Advances for Time Series Forecasting.* 2021.

47. **W. M. P. van der Aalst, V. Rubin, H. M. W. Verbeek, B. F. van Dongen, E. Kindler & C. W. Günther.** Process mining: A two-step approach to balance between underfitting and overfitting. 2010. Vol. 9, 1.

48. **Palić, Irena Palić, Petra, Banić, Frane.** The pre-pandemic role of customer online satisfaction in price determination: evidence from hotel industry. 2021. Vol. 7, 2.

49. **Mohammed, Ibrahim Guillet, Basak Denizci Law, Rob Rahaman, Wassiuw Abdul.** Predicting the direction of dynamic price adjustment in the Hong Kong hotel industry. 2021. Vol. 27, 2, pp. 346-364.

50. **Brownlee, Jason.** *Master Machine Learning Algorithms Discover How They Work and Implement Them From Scratch.* s.l. : http://machinelearningmastery.com/, 2016.

51. **Andrew, et al.** *HOTEL OCCUPANCY RATES WITH TIME SERIES MODELS: AN EMPIRICAL ANALYSIS.* s.l. : THE COUNCIL ON HOTEL, RESTAURAM AND INSTITUTIONAL EDUCATION , 1991.

52. **Nair, Girish K.** Data driven pricing strategies for hotels during the COVID-19 pandemic. 2021. Vol. 12.

53. **Pereira, Luis Nobre.** An introduction to helpful forecasting methods for hotel revenue management. 2016. Vol. 58, pp. 13-23.

54. **Mitra, Subrata Kumar.** An analysis of asymmetry in dynamic pricing of hospitality industry. 2020. Vol. 89.

55. **Edwin Baidoo, Jennifer L. Priestley.** An Analysis of Accuracy using Logistic Regression and Time Series. s.l. : Grey Literature from PhD Candidates, 2016.

56. **Blengini, Isabella and Heo, Cindy Yoonjoung.** How do hotels adapt their pricing strategies to macroeconomic factors? 2020. Vol. 88.

57. **Irina Karachun, Lyubov Vinnichek,Andrey Tuskov.** Machine learning methods in finance. 2021. Vol. 110.

58. **Denizci Guillet, Basak and Chu, Angela Mai Chi.** Managing hotel revenue amid the COVID-19 crisis. 2021. Vol. 33, 2.

59. **Mohammed, Ibrahim Guillet, Basak Denizci Law, Rob.** Modeling dynamic price dispersion of hotel rooms in a spatially agglomerated tourism city for weekend and midweek stays. 2019. Vol. 25, 8.