Aston University
BIRMINGHAM UK

LEVERAGING MACHINE LEARNING TO PREDICT AIRBNB PRICES IN EUROPEAN CITIES

A comparative Analysis of Predictive Models

Adenike Ayomipo Adeniyi

Candidate Number: 934092

Written under the Supervision of Dr. Seyedmojtaba Sajadi

BDM163 - MSc Business Analytics Business Project

September 2023

Master of Science (MSc) in Business Analytics

This project is submitted in partial fulfillment of the requirements for the master's in business Analytics at Aston University, dated September 29th, 2023

Declaration

I declare that I have personally prepared this report and that it has not in whole or in part been submitted for any other degree or qualification. Nor has it appeared in whole or in part in any textbook, journal or any other document previously published or produced for any purpose. The work described here is my/our own, carried out personally unless otherwise stated. All sources of information, including quotations, are acknowledged by means of reference, both in the final reference section and at the point where they occur in the text.

Acknowledgments

I extend my heartfelt gratitude to God who has enabled me to successfully navigate this project in good health and with a sound mind. My profound gratitude goes to my esteemed supervisor, Dr. Seyedmojtaba Sajadi, whose invaluable insights and unwavering guidance have been instrumental throughout the entirety of this project's journey. My thanks also extend to the dedicated faculty members and tutors within the Business Analytics department. Their dedication in imparting knowledge and nurturing skillsets has been pivotal in equipping me to accomplish this project.

I am forever indebted to my family for their unwavering and unconditional support. Their presence and encouragement have been a pillar of strength not only during the process of crafting this project but also throughout my academic pursuit and life's endeavors. To my circle of friends, your consistent emotional and physical support has been a source of motivation that I deeply appreciate. As I conclude this chapter, I want to acknowledge the collective contributions that have played an integral role in shaping this endeavor. My gratitude remains immeasurable, and I am humbled by the network of individuals who have stood by me every step of the way.

Abstract

In the ever-evolving landscape of European cities, the tourism and hospitality industry has undergone a transformative shift, largely fueled by the advent of home-sharing platforms like Airbnb. This shift has presented both opportunities and challenges for hosts, guests, and the industry at large. At the core of these challenges lies the intricate task of accurately predicting Airbnb prices, a task that holds profound implications for all stakeholders involved. This research embarks on a comprehensive exploration aimed at shedding light on the far-reaching importance of predictive modeling within the tourism and hospitality sector. It seeks to address the intricate problem of price prediction by examining it through a multifaceted lens, encompassing a range of factors that impact pricing dynamics.

The importance of this research lies in its potential to empower both Airbnb hosts and guests. Accurate price predictions can lead to improved decision-making for hosts, enhancing revenue and occupancy rates. Simultaneously, guests benefit from fair and transparent pricing, resulting in enhanced satisfaction. Predicting Airbnb prices in European cities is a complex challenge due to the myriad factors influencing pricing. This study aims to tackle this problem by harnessing the capabilities of three distinct machine learning models: Random Forest, XGBoost, and Neural Networks.

The research methodology involves rigorous data preprocessing to refine the dataset, providing a robust foundation for model development. Subsequently, a meticulous comparative analysis of the three models is conducted. This analysis unveils their individual strengths and limitations in the context of European city-specific price prediction. Among the models, the Random Forest model demonstrates remarkable consistency across diverse cities, showcasing its proficiency in delivering reliable predictions. In parallel, the XGBoost model excels in specific locales, capturing nuanced pricing patterns effectively. However, it's noteworthy that the Neural Networks model falls short in achieving consistently high performance across the diverse European cities analyzed. While it exhibits potential for localized mastery, its limitations in delivering competitive predictive accuracy are evident.

In summary, this research advances our understanding of predictive accuracy in the Airbnb pricing landscape, offering valuable insights for both hosts and guests. By delineating the strengths and limitations of different machine learning models, it provides a foundation for informed decision-making and enhanced user experiences.

Key Words: Machine Learning, Python, Tourism, Airbnb, Cities, Cultural preferences, Hospitality, Dynamic pricing, Data visualization, Economic conditions, Prediction, Comparative analysis, Model development, Random Forest, XGBoost, Neural Networks, Guest satisfaction

Word count: Around 20,000 words.

List of Figures

List of Tables

Table of Contents

CHAPTER ONE: INTRODUCTION

1. Introduction

1.1. Background

Airbnb, founded in 2008 in the USA, has transformed the way people access temporary accommodations by introducing a novel business model in the form of a digital marketplace (Zhu, Li, and Xie, 2020). Unlike traditional hotels, Airbnb does not own any properties but instead connects property owners with travelers seeking short-term stays. In May 2011, the company took a significant step towards international growth by acquiring the German company Accoleo, marking its first venture into the global market merely three years after launching its website. Building on this momentum, Airbnb established a London office in October 2011, further solidifying its presence in Europe. The year 2012 proved to be a turning point for Airbnb's international outreach as the platform experienced a substantial influx of users from around the world. Seizing the opportunity, Airbnb expanded its operations beyond the United States by setting up offices in key cities such as Barcelona, Paris, Copenhagen, Sao Paulo, Moscow, and Milan and These additions complemented the existing four offices in Berlin, San Francisco, Hamburg, and London, establishing a robust global network. Demonstrating its commitment to international growth, Airbnb witnessed a significant shift in its user base. As of June 2016, only 16 percent of the listings originated from the United States, signifying a transformative shift towards becoming a genuinely internationally oriented business. Paris emerged as the leading market for Airbnb, underscoring the platform's popularity and success in attracting global users. While New York and San Francisco remained prominent US markets, they were the only American cities to secure a spot in Airbnb's list of top ten markets.

This further highlights the company's focus on expanding its reach and catering to a diverse range of travelers and hosts worldwide. With a strong foothold in various international cities and a growing global user base, Airbnb has firmly established itself as a truly global platform for accommodation and travel experiences (Bowers, 2017). As of December 2022, Airbnb has an impressive 6.6 million listings available in over 220 countries and regions, spanning across 100,000 cities (Airbnb, 2023). With a steady average of two million guests nightly, it's evident that Airbnb is a top choice for travelers worldwide. The platform offers a vast array of lodging options, with over 700 million choices available across nearly every country worldwide, making it a popular choice for many travelers (Yang, 2021). As of 2017, a report by Hartmans (2017) stated that Airbnb is now bigger than the top five hotel brands put together. The New York Times reports that Airbnb has become a major contender in the short-term rental industry, contributing to the transformation of residential areas into popular destinations for tourists. Although the recent Covid outbreak has impacted travel, the city's growth continues to drive this trend. (Mahyoub et al., 2023).

1.2. Airbnb Business Model

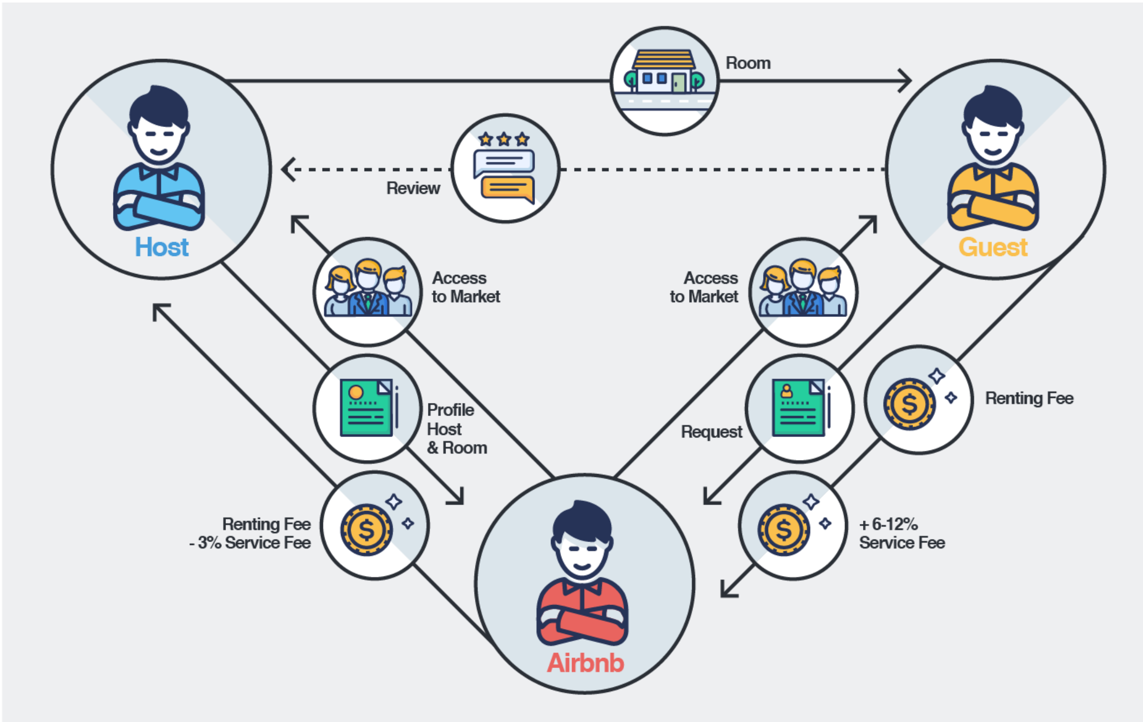

Figure 1: Airbnb Business Model
Source: Airbnb Flowchart - Bing

Figure 1 above illustrates the Airbnb business model showcasing how hosts list their accommodations on the Airbnb platform and how guests place booking requests for these properties. When a guest makes a reservation, they pay a service fee to Airbnb, which facilitates the transaction. From this service fee, Airbnb deducts its commission and then pays the host the remaining amount as their rental fee. The Airbnb business model is user-friendly and a straightforward process, making it convenient for travelers to find suitable accommodations that match their specific needs. By leveraging the platform, guests can access a diverse range of lodging options quickly and easily, gaining access to unique and authentic experiences that go beyond traditional hotel stays. For hosts, the Airbnb model presents an opportunity to monetize their unused or spare spaces, allowing them to generate additional income by listing their properties on the platform. It empowers individuals to become micro-entrepreneurs and share their homes or properties with travelers from around the world.

In summary, Airbnb has disrupted the traditional hotel industry by offering a digital platform that connects property owners with travelers seeking temporary accommodations. With a vast number of listings worldwide, Airbnb provides a wider range of experiences and prices compared to conventional hotels.

1.3. Airbnb in Europe

Airbnb's entry into the European market in 2010 marked the beginning of its remarkable growth in the region. As of 2021, Europe boasted the highest number of listings on the platform, with a staggering 4.84 million properties available for booking. This was followed by the Asia Pacific region with 2.21 million listings and North America with 2.55 million (Statista, 2021). The disruptive nature of Airbnb has had a profound impact on the traditional hotel industry and the private rental market in Europe. It has revolutionized the way travelers search for and book accommodations by offering affordable and unique options, while also providing property owners with an opportunity to monetize their unused spaces.

However, alongside its growth, Airbnb has faced criticism and concerns from various industry stakeholders, particularly hoteliers. Hoteliers argue that Airbnb operates in an unfair competitive environment as it is not subject to the same tax and regulatory constraints as traditional hotels (Dredge & Gyimothy, 2015). The lack of regulatory oversight has also raised concerns regarding safety, security, and the potential disruption caused by transient populations in residential areas (Kushlev & Dunn, 2015).

Although there are concerns, Airbnb's popularity continues to grow in Europe and many cities depend on its revenue. The European Travel Commission reported a 70% increase in Airbnb reservations in European cities over the past five years, prompting new regulatory measures in certain cities such as restrictions on the number of nights a property can be rented out and registration requirements for hosts. Roughly 25% of the EU's tourist accommodation supply is made up of short-term rental hosts, including those who host on Airbnb, proving to be an economic lifeline for many people. Recognizing the importance of hosting in the lives of everyday Europeans is crucial as the EU contemplates its short-term rental proposal.

According to a recent release on Airbnb News, the European Commission released a proposition in November 2022 for EU-wide Short-Term Rental (STR) regulations. These regulations aim to provide a unified framework for data sharing to enhance authorities' accessibility, provided that local regulations are straightforward and easy to comply with for everyday hosts. Hosts from different countries in the European Union have voiced their opinions on how to unleash the advantages of hosting for ordinary Europeans as the proposal advances through the EU institutions. Many individuals have been unable to benefit from hosting due to inconsistent and disproportionately restrictive local regulations (Airbnb, 2023).

1.4. Importance of the Tourism and Hospitality Industry

The tourism and hospitality industry, which includes platforms like Airbnb, stands as one of the globe's most lucrative sectors, consistently generating substantial revenue. This dynamic field continually evolves, witnessing the emergence of fresh niches at regular intervals. In addition to its economic significance, the tourism and hospitality industry, including Airbnb, plays a pivotal role in several as depicted in Figure 2 below:

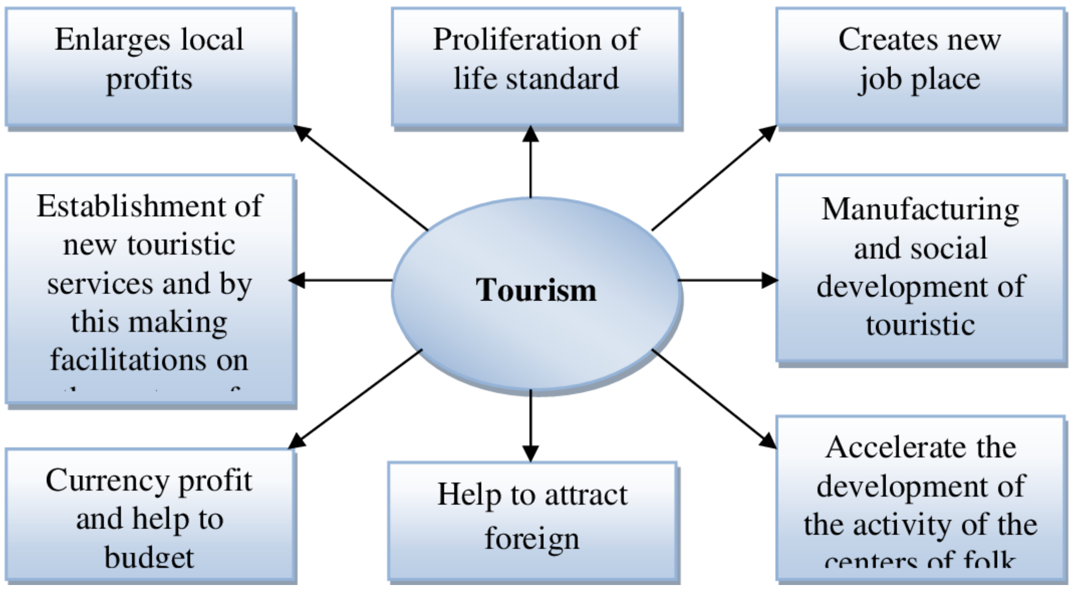

Figure 2: Importance of Tourism and Hospitality

The tourism industry stands as one of the globe's most lucrative sectors, consistently generating substantial revenue. This dynamic field continually evolves, witnessing the emergence of fresh niches at regular intervals. Furthermore, utilizing tourism development strategies as a means for fostering economic growth, particularly in identifying sectors necessitating such strategies, can yield benefits such as product development, marketing enhancements, increased profitability, attraction of investments, defining the roles of government and businesses in resource allocation, and providing support for regional planning. (Nafisa Teshaboyeva, 2023).

Figure 3: Global Tourism Industry Market Overview
Source: Researchnester

Figure 3 provides a comprehensive overview of the global tourism industry, highlighting Airbnb's prominent position within the broader Tourism and Hospitality sector. Notably, the figure underscores substantial growth prospects in the European tourism market from 2023 to 2035, emphasizing the critical need for precise and reliable predictive modeling for pricing in this region.

1.5. Rationale for Price Prediction

The price of an Airbnb listing is a key determinant of guest booking decisions. By setting the optimal price that aligns with guests' expectations, hosts can maximize their profitability. (Chattopadhyay and Mitra, 2019). Being knowledgeable about the market is essential, especially with the rise in the availability of rental accommodation and the increasing importance of pricing. For those who are budget-conscious and seeking accommodation in major cities like London, Airbnb is a highly recommended alternative to costly hotels, hostels, and motels. (Mahyoub et al., 2023). One major challenge for Airbnb hosts is determining the appropriate pricing for their listings. The unpredictability of Airbnb is evident in the frequent fluctuations in its pricing, making it difficult to predict. Unlike traditional hotel chains that use standardized pricing methods, Airbnb rentals' prices are set directly by the hosts which involves considering several factors before setting the price. One of the key advantages of Airbnb compared to traditional hotels is the wider range of experiences and prices it offers. Unlike hotels that often rely on star ratings and standardized pricing guidelines, pricing on Airbnb can be a complex undertaking within the sharing economy context (Zhu, Li, and Xie, 2020). Hosts on Airbnb have the flexibility to set their own prices based on various factors such as location, amenities, demand, and competition. However, determining a fair and competitive price can be challenging for both hosts and guests.

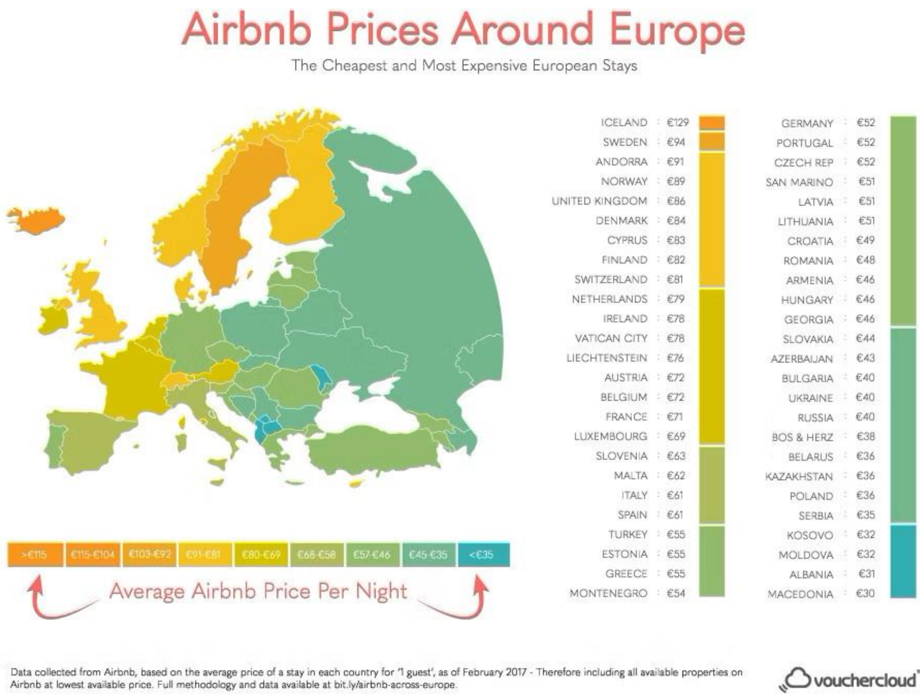

Figure 4: Graph showing Airbnb prices Around Europe as of February 2017.
Source : Average AIrbnb Costs Europe

Figure 4 illustrates the range of prices for Airbnb accommodations in European cities, with an average of £112 per night and a minimum of £26 in February 2017 (Lonely Planet, 2017). Since then, there has been a notable increase in prices over the past six years. Recent data from AllTheRooms (2022) indicates that the current average price for an Airbnb accommodation in Europe is £92. Based on a study conducted by Kwan et al. (2018), Airbnb rentals in Europe have a higher average daily rate compared to hotels. This is uncommon in regions like the USA where hotel prices are typically lower than Airbnb. The study indicates that hosts understand the demand for their properties and are able to adjust their prices accordingly. Moreover, the study highlighted those peripheral areas tend to have lower-priced Airbnb rentals compared to central locations, likely due to lower demand in those areas. The pricing dynamics of Airbnb rentals in Europe are complex and influenced by various factors, necessitating careful consideration by both hosts and guests.

In summary, the growth and impact of Airbnb in Europe have been significant. It has disrupted traditional accommodation models, generated new revenue streams, and posed regulatory and competitive challenges. It is important for both hosts and guests to have a thorough understanding of the intricate pricing system on Airbnb, especially with the constantly evolving European market.

1.6. Why Machine Learning?

Machine learning, a subset of artificial intelligence, has gained significant attention in recent years (Alqahtani et al., 2021). It involves the development of algorithms that learn from data to make predictions or decisions. In the context of Airbnb pricing, machine learning algorithms analyze factors like location, size, amenities, and historical pricing data to predict the optimal price for a listing. This approach enables hosts to maximize profits and ensures fair pricing for guests. However, while some research has explored machine learning for Airbnb pricing in the US, further investigation is needed for European cities due to their unique cultural, economic, and geographic factors. The application of machine learning in Airbnb pricing prediction has gained attention in recent years, with researchers exploring various algorithms and methodologies to improve price estimation accuracy (Zhu, Li, & Xie, 2020). Understanding the pricing structure of Airbnb is crucial, and machine learning algorithms can help hosts navigate the complex pricing landscape, benefiting both hosts and guests. To advance understanding in European cities, comprehensive research considering market characteristics and nuances is essential. Determining fair pricing can be challenging, but machine learning algorithms offer a promising solution by leveraging data and analyzing various factors. Hosts can maximize profits, and guests can make informed choices based on fair and transparent pricing. Overall, machine learning has the potential to revolutionize Airbnb pricing prediction, but further research, especially in European cities, is needed to fully explore its application and understand the influencing factors. By combining different machine learning models and considering local context, more accurate and effective pricing predictions can be achieved, benefiting both hosts and guests.

1.7. Aim, Research Questions, and Impact of Research

1.7.1. Aim of the Research

The aim of this research is to delve into the effectiveness of different machine learning models in predicting Airbnb prices specifically in European cities and build a reliable model that can be applicable for different European cities in predicting Airbnb Prices in their cities. As Airbnb continues to gain popularity as a preferred choice for both travelers and property owners, it becomes crucial to develop reliable predictive models that can accurately forecast the prices of listings. To accomplish this, the study compares and evaluates the performance of various machine learning algorithms. By examining the predictive capabilities of these models, we will aim to identify the most effective approach for predicting Airbnb prices in European cities. This comparative analysis allows for an in-depth understanding of the strengths and weaknesses of each model, providing valuable insights into their applicability in real-world scenarios particularly for different European Cities. The research aims to contribute to the existing knowledge by shedding light on the predictive power of different machine-learning algorithms in the context of Airbnb pricing.

The findings of the study will not only enhance our understanding of the factors influencing Airbnb prices but also provide practical recommendations for property owners, travelers, and other stakeholders in the Airbnb ecosystem. Overall, this research aims to bridge the gap between the increasing demand for accurate Airbnb price predictions and the need for reliable models. By investigating and comparing various machine learning algorithms, the study seeks to provide valuable insights and practical recommendations that can benefit property owners, travelers, and other stakeholders in the Airbnb ecosystem, ultimately contributing to the growth and efficiency of the market.

1.7.2. Research Questions:

The research questions formulated to achieve the above aim are as follows:

- What are the key predictors of Airbnb prices in European cities?
- How effective are predictors in predicting Airbnb prices for European cities?
- How does the performance of the predictive models vary across different European cities?
- What insights can be gained from the analysis of the predictive models in terms of the factors that influence Airbnb prices in European cities?

1.7.3. Impact of research

Answering the research questions outlined above regarding the factors that determine Airbnb prices in European cities can have significant implications for various stakeholders involved in the Airbnb industry.

• **Property owners and hosts** can optimize pricing and attract more customers by understanding key pricing factors, enhancing their listings, and maximizing revenue.

• **Travelers** benefit from accurate price predictions, enabling better trip planning and budgeting, enhancing their travel experience.

• **Policymakers** can use insights to create fair regulations, promote pricing transparency, and support sustainable growth in the Airbnb industry.

- **Researchers and scholars** gain valuable insights into machine learning's application in hospitality, offering comparative analysis for future research and discussions.

In summary, answering the research questions on Airbnb pricing in European cities can provide valuable insights that benefit property owners, hosts, travelers, policymakers, and researchers. It can lead to improved pricing strategies for hosts, better travel planning for guests, fair competition, and regulation for policymakers, and contribute to the broader understanding of machine learning applications in the hospitality industry. Ultimately, this study has the potential to contribute to the sustainable growth and development of the Airbnb industry in European cities.

1.8. Structure of the research

In section two of the research paper, a thorough literature review focuses on predicting Airbnb pricing using Machine Learning methods. The goal is to gather and analyze existing studies and publications related to the topic, understanding the current state of research, identifying gaps, and exploring different approaches. In section three, the methodology is detailed, including data sources, preprocessing techniques, and Machine Learning methods used for prediction. This section outlines the selection of algorithms, feature engineering, and optimization procedures. Section four presents the results, comparing and evaluating model performance using various metrics. The results are presented in an organized manner, potentially using tables, charts, or graphs. Section five discusses insights from the analysis, interpreting results, identifying data patterns, and discussing implications. Strengths, limitations, and influencing factors are critically analyzed. Section six summarizes the paper, highlighting key findings and conclusions. It also offers recommendations for Airbnb stakeholders in European cities, including pricing optimization strategies, suggestions for improving prediction accuracy, and insights into the adoption of Machine Learning models for pricing decisions.

CHAPTER TWO: LITERATURE REVIEW

2. Literature Review

Over the past decade, the emergence and rapid growth of Airbnb has revolutionized the traditional accommodation industry. The platform has gained immense popularity among travelers looking for unique and affordable accommodation options, while also offering property owners an opportunity to earn income by renting out their properties. With Airbnb's growing popularity, there has been a growing interest in predicting Airbnb rental prices, particularly in European cities. As a result, a considerable amount of research has been conducted on this topic, utilizing machine learning and deep learning techniques to develop accurate predictive models. This literature review will provide a comprehensive overview of the existing research on leveraging machine learning to predict Airbnb prices in European cities, highlighting the various models and approaches utilized and their comparative performance while also focusing on the factors that affects Airbnb pricing. This literature review incorporates articles obtained from reputable academic sources such as Google Scholar, IEEE Xplore, and ScienceDirect. The selection of articles primarily focuses on scholarly works published within the last Seven years to ensure the review's relevance.

2.1. Airbnb price determinants and price prediction

The concept of pricing holds immense importance in the realm of any business specifically, within the sharing economy's hospitality industry, platforms like Airbnb heavily rely on the pricing strategies employed. The impact of pricing extends beyond merely influencing guests' choices when it comes to selecting accommodations; it significantly affects the profitability of hosts (Lampinen and Cheshire, 2016). Consequently, some studies have been conducted to investigate the factors that influence pricing.

A study on Airbnb listings price recommendation was conducted by Li et al. (2016), utilizing multi-scale clustering. Their findings revealed that the proximity to the closest landmark, the impact of facilities, and the popularity of the nearest landmark have noteworthy effects on Airbnb listing prices. The study provides valuable insights into the correlation of these factors with pricing, highlighting the significance of location, amenities, and the appeal of the surrounding area. However, the research's significance could be enhanced by elaborating further on methodologies, incorporating additional factors, and discussing generalizability.

Castro and Ferreira's (2018) study unequivocally demonstrates that Airbnb pricing in the hospitality industry is subject to the influence of multiple factors, including star ratings, facilities, consumer ratings, and location. Similarly, Blal, Singal, and Templin's (2018) study emphatically underscores the significant impact of ratings on Airbnb listings and hotel sales growth. It is important to note, however, that these studies may possess limited generalizability due to sample size and representativeness constraints.

A recent study conducted by Barnes and Kirshner (2021) has uncovered a significant link between the facial features of Airbnb hosts and the pricing of their accommodations. The study highlights the importance of using facial image analytics and deep learning to analyze host faces in a large dataset of Airbnb accommodations across 10 cities in the United States. The results indicate that trustworthiness and attractiveness are essential factors in determining the price of accommodations and can increase it by up to 5%. Moreover, the researchers tested the Gray theory of motivation and discovered that trust is more critical when sharing accommodation with strangers. However, the study has limitations as it focuses solely on the perceived trustworthiness and attractiveness of hosts based on their photos and is restricted to the United States.

In conclusion, the topic of pricing is of paramount importance in the Airbnb sector within the sharing economy. Its influence on guests' decisions and hosts' profitability underscores the significance of studying the factors that shape pricing in European cities. The existing research contributes valuable insights into the complex dynamics of pricing and informs strategic decision-making for both hosts and guests. Continued exploration of this topic can further enhance our understanding of Airbnb pricing mechanisms and drive advancements in the industry.

2.2. Adoption of Machine Learning for Increasing Business Performance

Over the past few years, machine learning has shifted from being a subject of academic research curiosity to a discipline capable of addressing practical business challenges. (Paleyes, Urma and Lawrence, 2022) The algorithmic concept of Machine Learning has been employed multiple times. The potential of machine learning (ML) and systems based thereon has grown steadily in recent years with the ability of ML systems to identify relationships rapidly and systematically in large volumes of data, which can be used to analyze new data to make meaningful predictions, enables organizations of all industries to make their processes more effective and efficient (Pumplun, 2022). Both Machine Learning and deep learning have rapidly evolved into powerful resources spanning diverse domains, encompassing endeavors like recognizing images and speech, processing natural language, and even applications in the medical field. These technologies are reshaping our interactions with technology and unleashing novel avenues for innovative advancements. (Sharifani and Amini, 2023). According to (Haug and Drazen, 2023), Advancements in data science go beyond mere enhancements in performance, speed, and storage capacity. Apart from the conventional data available in libraries, data generated within organizations, and structured systems designed for data collection and organization, emerging technologies can also leverage data originating from both human interactions and automated processes. Similarly, Goodell et al. (2021) in their research of adoption of Machine learning in finance discussed how Machine learning provides an effective solution to address the limitations of traditional econometric models when it comes to identifying outliers, extracting features, and conducting classification and regression tasks on intricate datasets.

2.3. Adoption of Machine Learning Algorithm in Predicting Airbnb Prices

The emergence of machine learning techniques has revolutionized the field of price prediction, enabling researchers to develop complex models capable of accurately forecasting prices in various domains. In recent years, there has been a growing interest in applying more advanced machine learning techniques to predict prices in diverse industries, including the real estate market, stock market, and e-commerce platforms.

In their research, Kalehbasti el al (2020) proposed a robust price prediction model that combined machine learning, deep learning, and natural language processing techniques. The objective was to provide a valuable tool for both property owners and customers in making informed decisions. The prediction model incorporated a diverse range of predictors, including rental features, owner characteristics, and customer reviews. These predictors were fed into various machine learning algorithms, such as tree-based models, linear regression, neural networks (NNs), support-vector regression (SVR), and K-means clustering (KMC). The study aimed to develop a model that consistently predicted optimal prices, even when only a limited set of property features were available. By leveraging different machine learning algorithms, the researchers explored the performance of each method and identified the best-performing model. The results of the study revealed that the Support Vector Regression (SVR) algorithm yielded the highest R^2 score and achieved the lowest Mean Squared Error (MSE) among the tested models. This indicates that the SVR model was able to accurately predict prices based on the given set of features, showcasing its effectiveness in the price prediction task. Overall, Kalehbasti et al. (2020) demonstrated the potential of utilizing machine learning, deep learning, and natural language processing techniques in constructing a reliable price prediction model for the rental market. Their findings highlight the superiority of the Support Vector Regression algorithm in achieving accurate price predictions, offering valuable insights for both property owners and customers in the decision-making process. Unfortunately, the paper failed to deliver a comprehensive price prediction model in the end.

In their study, Peng, Li, and Qin (2020) aimed to evaluate the efficacy of advanced algorithms, specifically DNN (Deep Neural Networks) and XGBoost, in comparison to traditional linear-based models like Linear Regression and Support Vector Regression for predicting Airbnb prices. The researchers discovered that the adoption of these exotic algorithms resulted in superior performance when compared to the linear-based models. This suggests that leveraging more sophisticated machine learning techniques can lead to more accurate price predictions in the context of Airbnb listings. However, it is important to note that the study had certain limitations. One of the main constraints was the restricted market range caused by the availability of data solely from Inside Airbnb. This limitation implies that the findings might not fully capture the entire spectrum of Airbnb listings and their corresponding prices. It is possible that the market dynamics and characteristics beyond the scope of Inside Airbnb's dataset were not considered, potentially impacting the generalizability of the research findings. Despite these limitations, the study's results still highlight the potential of utilizing advanced algorithms, such as DNN and XGBoost, for predicting Airbnb prices, indicating their superiority over traditional linear-based models. Future research in this area could overcome the limitations by incorporating a more diverse and comprehensive dataset to obtain a broader understanding of the factors influencing Airbnb prices.

Luo, Zhou, and Zhou (2019) conducted a research study titled "Predicting Airbnb Listing Price Across Different Cities - New York City (NYC), Paris, and Berlin." In their investigation, they employed various machine learning techniques, including Neural Networks, Random Forest, and XGBoost. The results of their study revealed that the Neural Network approach outperformed the other methods in terms of predictive accuracy and performance. Furthermore, the study delved into the impact of label transformations on the models, both with and without thresholding. Notably, the researchers discovered that applying a logarithmic transformation to the labels significantly enhanced the model's performance, even in the absence of thresholding. This finding highlights the importance of label transformations in improving the effectiveness of the predictive models. However, despite the promising outcomes, the researchers acknowledged that there were areas in which the research could be further strengthened. Specifically, they noted that incorporating additional feature extraction techniques could have potentially boosted the performance of the Neural Networks even further. Additionally, conducting hyper-parameter tuning could have fine-tuned the models and optimized their predictive capabilities. These enhancements would have potentially improved the overall effectiveness of the Neural Network approach in predicting Airbnb listing prices.

In a related study, Liu (2021) utilized various machine learning models, including KNN, MLR, LASSO regression, Ridge regression, Random Forest, Gradient Boosting, and XGBoost, to predict pricing for Airbnb listings. The research revealed that XGBoost was the most effective model for predicting pricing, achieving an impressive R^2 score of 0.6321. The performance of different regression models was assessed, leading to varying levels of accuracy. KNN exhibited the lowest performance, with an RMSE of 77.0055 and an R^2 score of 0.3934. Conversely, XGBoost demonstrated the highest performance, with an RMSE of 59.9743 and an R^2 score of 0.6321. Gradient Boosting was closely followed with an RMSE of 60.2121 and an R^2 score of 0.6292. MLR, LASSO regression, and Ridge regression exhibited similar levels of performance, with RMSE values around 64 and R^2 scores around 0.58. In a comparison between the three, MLR slightly outperformed Ridge regression, while Ridge regression marginally outperformed LASSO regression. Notably, despite the increased complexity of regularized regression models compared to MLR, they did not substantially enhance prediction accuracy for this specific dataset. The findings suggested that LASSO regression, despite its feature selection capabilities, did not significantly improve prediction accuracy for the dataset. Additionally, incorporating penalties to address overfitting did not contribute significantly to the predictive power of the models (Liu, 2021).

Furthermore, Kanakaris and Karacapilidis (2023) introduced a novel approach to predicting the prices of Airbnb listings in tourist destinations, specifically focusing on Santorini Island. They incorporated not only the characteristics of individual listings but also the topological and neighborhood properties that influence Airbnb pricing in Santorini. The study employed techniques from graph theory, including Graph Neural Networks (GNNs) and document embeddings. The results revealed that incorporating neighborhood information improved the R^2 coefficient and reduced the mean squared error (MSE). However, it is important to note that despite these positive outcomes, the study did not provide a complete price prediction model. Nonetheless, based on their findings, the researchers concluded that incorporating graph structure information, coupled with a modern GNN-based approach, is more suitable for predicting prices in a network of Airbnb listings compared to traditional neural network models.

2.4. Airbnb Price Prediction Using Machine Learning in Europe

Researchers have successfully utilized machine learning to predict Airbnb prices in the USA. However, the situation in Europe presents a different set of circumstances with not many researchers adopting machine learning for the prediction of Airbnb prices. This section will concentrate on scholars who have implemented machine learning techniques to forecast Airbnb prices across a variety of European cities.

Fitrianingsih, Rahayu, and Zazila (2022) aimed to assist Airbnb hosts in determining the most suitable property price according to the specific season in their study titled "Dynamic Pricing Analytic of Airbnb Amsterdam Using K-Means Clustering," Their research employed the K-Means clustering algorithm to conduct a dynamic pricing analysis using Airbnb data from Amsterdam. The findings revealed that the unsupervised learning model of K-Means Clustering can effectively support dynamic pricing, which is a competitive marketing strategy for companies. It is important to note that the study focused specifically on Airbnb data in Amsterdam and may not be applicable to other locations or platforms.

In a recent empirical study conducted by Mahyoub et al. (2023), a range of machine learning algorithms, namely Linear Regression, XGBoost, Random Forest, Artificial Neural Network (ANN), and K-Nearest Neighbors (KNN), were employed to predict Airbnb prices in the city of London. The comparative analysis aimed to evaluate the performance of these models in terms of price prediction accuracy. The findings revealed that the Random Forest algorithm outperformed the other models, exhibiting the highest accuracy by explaining 86.9% of the price variation. Notably, Linear Regression achieved a respectable R^2 score of 82.44, a root mean squared error (RMSE) of 39.56, and a mean absolute error (MAE) of 25.18. In contrast, the Gradient Boosting Regressor (GBR) demonstrated a significantly lower R^2 score of 68.72, indicating poorer performance compared to Linear Regression. On the other hand, the Random Forest Regressor showcased remarkable predictive capabilities, attaining an impressive R^2 score of 86.94, a low RMSE of 0.38, and a minimal MAE of 0.31. These metrics indicate the robust performance of the Random Forest model. Similarly, the XGB Regressor model yielded comparable results to the Random Forest model, albeit with slightly inferior performance. Specifically, the XGB Regressor achieved an R^2 score of 86.6, an RMSE of 0.39, and an MAE of 0.318. Overall, the study highlights the superiority of the Random Forest algorithm in accurately predicting Airbnb prices in the context of London.

It is however important to note that the study's focus was solely on the dataset from London, which implies a limitation in the generalizability of the findings to other European cities. Each city has its own unique characteristics, including economic factors, demand patterns, and cultural influences, which may significantly impact the factors affecting Airbnb pricing. Therefore, while the Random Forest algorithm exhibited strong predictive power in the context of London, caution should be exercised when applying this model to accurately predict prices in other European cities due to the potential variation in the underlying factors influencing pricing dynamics. Further research would be necessary to develop and validate models that are specifically tailored to individual European cities to ensure accurate price predictions in those locations.

2.5. Advantages of Using Machine Learning Algorithms for Price Predictions

Machine learning proves to be an effective approach for price prediction due to its enhanced precision, efficiency, and autonomous nature, by training the data and developing algorithms that learn from the data, machine learning establishes rules that enable accurate predictions. Through the evaluation of test data, machine learning algorithms can generate results without the need for human intervention, saving time and resources (Dahiya et al., 2022). Compared to traditional statistical methods, machine learning methods offer several advantages in price prediction. Unlike statistical models that often rely on assumptions like linearity, machine learning models can handle complex and non-linear relationships, allowing for more accurate predictions (Thessen, 2016). Additionally, machine learning algorithms have the ability to infer missing data, making them robust in scenarios where data may be incomplete or contain gaps. This capability further enhances the accuracy of price-prediction models.

Furthermore, machine learning reduces the long-term burden of expert annotation. Instead of relying heavily on manual annotation and human expertise, machine learning algorithms can learn directly from data, adapting and improving over time (Thessen, 2016). Once an algorithm acquires the knowledge of how to handle data, it can automatically perform its tasks, streamlining the prediction process and allowing for real-time updates (Mahesh, 2018).

In summary, machine learning's effectiveness in price prediction is attributed to its precision, efficiency, autonomy, ability to handle complex relationships, infer missing data, and reduce the burden of manual annotation. These advantages make machine learning a powerful tool for accurate and automated price forecasting.

2.5. Summary of Literature Review

Table 1, presented in this section, serves as a comprehensive summary of all the relevant literature that has been meticulously reviewed and analyzed in the course of this research. It encapsulates the key findings, methodologies, and insights drawn from a range of scholarly sources, providing a condensed yet informative overview of the extensive body of literature that has contributed to shaping the foundations of this study.

Authors	Year	Origin	Purpose	Type of Source	Methodogy Adopted	Findings	Limitation
Castro and Ferreira	2018	Lisbon and portugal	Study Airbnb pricing factors	Research	Cluster Analysis	Star ratings, facilities, consumer ratings, and location influence pricing.	Limited generalizability due to sample size and representativeness.
Blal, Singal, and Templin	2018	San Fransisco - USA	Explore the impact of ratings on Airbnb listings and hotel sales	Research	Mixed Model - Not specified	Ratings significantly affect Airbnb listings and hotel sales.	Potential limitations due to sample size.
Barnes and Kirshner	2021	United States	Investigate the link between host facial features and pricing	Research	Facial image analytics and deep learning	Trustworthiness and attractiveness of hosts influence pricing. Trust is more critical when sharing with strangers.	Focus on U.S., limited to host photos, and lack of comprehensive coverage of factors.
Kalehbasti et al.	2020	USA	Develop a price prediction model for rentals	Research	SVR, NN, and KMC,	SVR model is the most effective in predicting prices.	Lack of a comprehensive price prediction model.
Peng, Li, and Qin	2020	New York, Paris and Berlin	Evaluate advanced algorithms for Airbnb price prediction	Research	Linear Regression, SVR, XGBoost, deep Neural Network	Advanced algorithms outperform linear-based models.	Limited market range and data source (Inside Airbnb).
Luo, Zhou, and Zhou	2019	USA	Predict Airbnb prices in different cities	Research	Neural Networks and Random Forest,	Neural Network approach outperforms other methods.	Potential improvements in feature extraction and hyper-parameter tuning not explored.
Liu	2021	N/A	Compare various machine learning models for price prediction	Research	KNN, MLR, XGBoost	XGBoost is the most effective model for predicting pricing.	Different models have varying levels of accuracy.
Kanakaris and Karacapilidis	2023	Santorini - Greece	To adopt graph theory in predicting Airbnb prices	Research	Graph Neural Networks and document embeddings	Neighborhood information improves prediction accuracy.	Absence of explainability features, improved document embeddings, enhanced preprocessing and feature selection, and the removal of outlier listings in the proposed approach.
Fitrianingsih, Rahayu, and Zazila	2022	Netherlands	Analyze dynamic pricing with K-Means Clustering	Research	K-Means clustering	K-Means clustering effectively supports dynamic pricing.	Focuses only on Airbnb data in Amsterdam, limited applicability.
Mahyoub et al.	2023	United Kingdom	Compare machine learning algorithms for London Airbnb prices	Research	Linear Regression, Random Forest, ANN, XGBoost	Random Forest outperforms other models for London price prediction.	Limited generalizability to other European cities due to unique characteristics.
Paleyes, Urma and Lawrence	2022	N/A	Discuss the growth of machine learning in addressing business challenges	Discussion	N/A	Machine learning identifies relationships in data for improved predictions.	N/A
Pumplun	2022	N/A	Highlight the use of ML in making processes more efficient	Discussion	N/A	ML systems identify relationships in data for process optimization.	N/A
Sharifani and Amini	2023	N/A	Discuss the evolving applications of ML and deep learning	Discussion	N/A	ML and deep learning reshape interactions with technology and enable innovations.	N/A
Haug and Drazen	2023	N/A	Explore data science advancements	Discussion	N/A	Emerging technologies leverage data from human interactions and automated processes for analysis.	N/A
Goodell et al.	2021	N/A	Discuss the use of ML in finance	Discussion	N/A	ML effectively addresses limitations of traditional econometric models in finance.	N/A

Table 1: Findings of Reviewed Literatures

2.6. Literature Gaps

A noticeable research gap exists concerning the performance evaluation of diverse machine learning models across various European cities, despite the distinctive traits that each city possesses. The crucial task of identifying the most suitable model for each location has yet to be explored comprehensively. This study endeavors to rectify this deficiency by predicting Airbnb prices within eight distinct European cities, thus augmenting the existing knowledge in this domain. While prior research on Airbnb price prediction exists, it is constrained in its ability to be applied to other European cities or regions. The focal points have predominantly been Amsterdam and London, inadvertently neglecting the significance of scrutinizing outcome variations across diverse locales to attain a holistic comprehension of the ever-evolving pricing dynamics within the Airbnb context. Additionally, there has been a dearth of exploration into comparative pricing analysis among assorted European cities, especially within the realm of price prediction. A more profound understanding necessitates an examination and juxtaposition of pricing trends across multiple cities, accounting for their unique idiosyncrasies and attributes. Furthermore, existing research has primarily revolved around factors encompassed by Inside Airbnb's dataset, neglecting potential influences on Airbnb prices that lie beyond its scope. This oversight has the potential to undermine the generalizability of findings. Hence, this study strives to address these gaps by deploying multiple machine learning models for predicting Airbnb prices and conducting a comprehensive comparative analysis spanning eight distinct European cities. This approach enables an enriched comprehension of the fluid pricing dynamics within the Airbnb landscape, with a specific emphasis on the distinctive traits and dynamics characterizing various European cities. Furthermore, this study aims to provide an in-depth model comparison across the different cities.

In conclusion, the current literature on Airbnb pricing prediction demonstrates a widespread utilization of machine learning techniques to create precise predictive models. Many researchers have successfully employed Random Forest and XGBoost in their analyses, obtaining positive results for various cities. In our study, we aim to build upon this existing research by adopting machine learning methods and applying them to eight different cities in Europe. The primary objective is to identify the most suitable model for accurate price prediction in each city. Comparative studies have been conducted to assess the performance of different algorithms and models, highlighting the strengths and weaknesses of various approaches. Researchers have made significant progress in developing reliable pricing models, ranging from robust models involving machine learning, deep learning, and natural language processing, to investigating the impact of outlier detection and specific features on prediction accuracy. The ultimate goal of this research is to contribute to the existing body of knowledge by providing a more dependable and comprehensive price prediction model for Airbnb listings. We will compare the accuracy of the different models in European cities to offer valuable insights for both hosts and guests, enabling them to make informed decisions. By enhancing the efficiency and transparency of the European Airbnb marketplace, our research aims to benefit all stakeholders involved.

CHAPTER THREE: METHODOLOGY

3. Methodology

The objective of this research is to predict Airbnb prices in European cities through the use of machine learning techniques. To accomplish this goal, we will train various models utilizing different combinations of data, methods, and target variables. The approach will involve utilizing 10 predictors that impact Airbnb pricing to create precise predictions. Table 2 provides a summary of the predictors, machine learning methods, and specific European cities included in the study.

Predictors	Methods	Cities
▪ Day	▪ Random Forest	▪ Amsterdam
▪ Room Type	▪ Neural Network	▪ Athens
▪ Person Capacity	▪ XGBoost	▪ Paris
▪ Super host		▪ Rome
▪ Multiple Rooms		▪ Vienna
▪ Cleanliness Rating		▪ Barcelona
▪ Guest Satisfaction		▪ Lisbon
▪ Number of Bedrooms		▪ Berlin
▪ City Centre (km)		
▪ Metro Distance (km)		

Table 2: Summary of Predictors, Machine Learning Methods, and Cities

Data was sourced from a total of Eight (8) Famous European Cities, and we will construct models for each of the Cities making a total of 24 different predictive models by combining various predictors, methods, and target variables for each of the selected European cities. These models aim to predict Airbnb prices accurately for each city. By creating these 24 models, we will be able to compare their performances and address How the predictive models' performances vary across the different European cities. Through this comprehensive comparison, we seek to gain valuable insights into the effectiveness and suitability of the various models in predicting Airbnb prices in distinct European locations. This analysis will shed light on the factors that influence pricing patterns and uncover any city-specific nuances that could impact the model's accuracy and applicability. Ultimately, the results will contribute to a deeper understanding of the predictive capabilities of machine learning in the context of Airbnb pricing across diverse European cities.

3.1. The Dataset

In this research, **a quantitative approach** was employed to collect and analyse the data with the aim of drawing conclusions and making precise price predictions. The data utilized for this research was obtained from Kaggle, a well-known platform for sharing and exploring datasets, no ethical codes were broken in obtaining these datasets. The source of the dataset was duly referenced. The data was sourced in a CSV format. The original dataset comprised 37,693 rows and 12 columns. Each row in the dataset representing a distinct data point, corresponding to the recorded Airbnb apartment While each column represents the specific attribute or characteristic associated with the apartments in the dataset.

For this analysis, the data set used was the **Cross-Sectional Dataset** as the observations or cases were collected at a single point in time, rather than over a period of time. Table 3 provides a brief description and roles of the target and predictors of the dataset that was used in this analysis.

Variable Name	***Description***	***Role***	***Variable Type***
Price	Price of Airbnb in Euros	Target/Dependent Variable	Continuous
Day	This pertains to whether it is a weekday or weekend.	Predictor/Independent Variable	Categorical
Room Type	Type of Airbnb accommodations available, which include Entire Apartments, Private Rooms, and Shared Rooms.	Predictor/Independent Variable	Categorical
Person Capacity	This refers to the maximum number of individuals that can occupy the Airbnb space.	Predictor/Independent Variable	Ordinal
Super host	A rating that classifies the host based on the customers experience.	Predictor/Independent Variable	Nominal
Multiple Rooms	Airbnb with more than one room. Ideally two to four rooms.	Predictor/Independent Variable	Categorical
Cleanliness Rating	A rating that classifies the cleanliness of the house	Predictor/Independent Variable	Nominal
Guest Satisfaction	Satisfaction rating of the host ranging from 1-10.	Predictor/Independent Variable	Continuous
Number of Bedrooms	Number of bedrooms available in the Airbnb apartment	Predictor/Independent Variable	Categorical
City Center (km)	Distance of the Airbnb apartment from the City Centre	Predictor/Independent Variable	Continuous
Metro Distance (km)	Distance of the Airbnb apartment from the Metro Stations	Predictor/Independent Variable	Continuous

Table 3: Description of the dataset.

3.1.1. The Target Variable.

The target variable in this study as highlighted in table 2 above is the listing price of accommodations offered on the Airbnb platform in different European cities the price is represented **Euros (€).** Accurately predicting these Airbnb prices is essential for optimizing pricing strategies for all the stakeholders of Airbnb. Our analysis focuses on utilizing machine learning models to predict the prices of Airbnb listings in eight European cities.

3.1.1.1. Price Distribution of Airbnb in European Cities

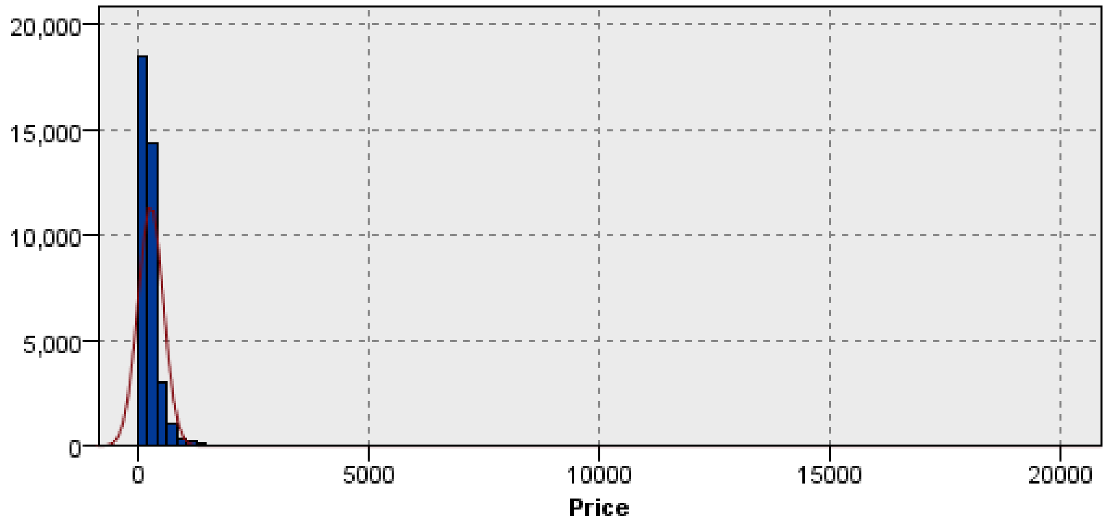

Figure 5: Histogram showing the price distribution in the dataset.

Figure 5 illustrates the distribution of Airbnb prices across various European cities, as per our dataset. It is evident from the right-skewed histogram that a significant portion of the Airbnb prices tends to cluster around the lower end of the scale, while there are only a few instances of very high prices on the upper end. This observation is intriguing, and in our further analysis, we will delve into these standout prices, particularly focusing on Athens, which has the lowest average price among the European cities. Two notable outliers in the data are an Airbnb accommodation in Paris and Athens, listed at an unusually high price of €16445.61 and €18545.45. This prices significantly deviates from the average, and it is possible that such extreme values could be attributed to outliers or factors not accounted for in our current analysis.

3.1.1.2. Average Price of Listings in European Cities

To understand the pricing of Airbnb further comprehensively and how it differs from city to city in the across the European region, Figure 4 below shows the average prices of Airbnb prices in the different European cities:

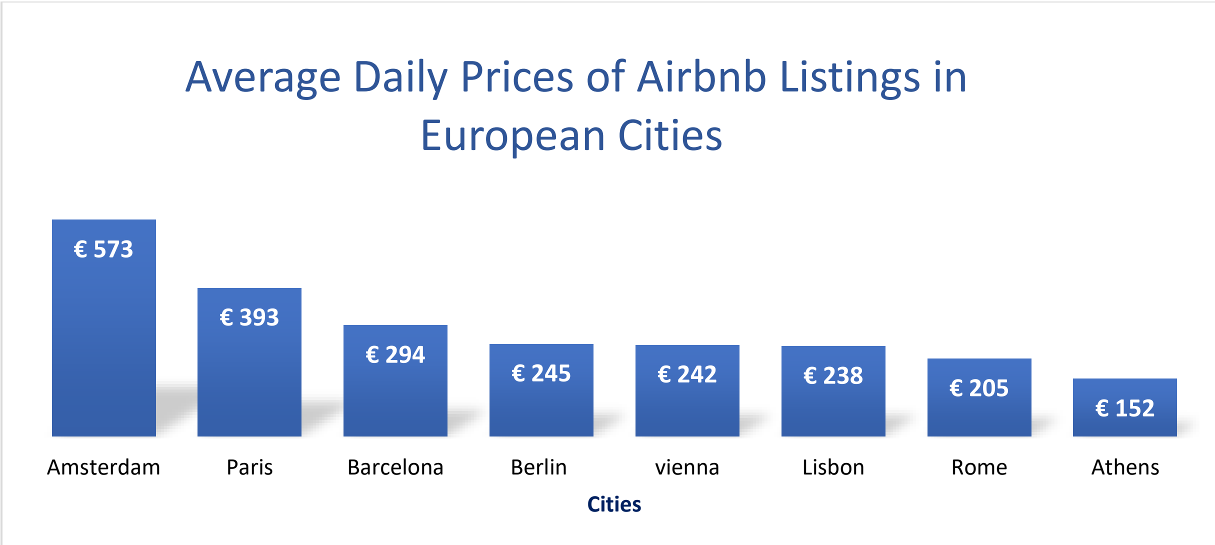

Figure 6: Bar Chart showing Average prices of Airbnb per day in Europe.

Amsterdam stands out with the highest average price of €573, indicating its relatively higher cost compared to other cities. Paris follows with a lower average price of €393 but remains relatively higher than some other cities. Barcelona, although lower than Amsterdam and Paris, still maintains a relatively higher position than the other cities in the Region. Moving down, Berlin offers more affordable options compared to the aforementioned cities. Vienna and Lisbon fall within the mid-range in terms of prices, while Rome is among the more affordable cities. Athens has the lowest average price of €152 per night, making it the most affordable option between the cities in this study. However, it is important to consider factors such as apartment capacity, number of bedrooms, proximity to the city centre, and metro distance when interpreting these prices. Additionally, prices can fluctuate throughout the year, with summer generally being a more expensive period due to increased tourist demand.

3.1.2. The Predictors

In constructing the predictive model for pricing, we incorporated ten predictive variables categorized as both continuous and categorical variables that influence the prices of Airbnb listings in European cities.

3.1.2.1. Exploring The Continuous Variables

We will seek to explore the continuous variables in our dataset, The descriptive statistics accurately showcases the essential metrics such as the average and most frequent values. These statistics offer valuable insights into the central tendencies and the level of variability within the dataset, shedding light on the underlying patterns and characteristics of the continuous variables. By utilizing descriptive statistics, we aim to provide a comprehensive and detailed understanding of the data distribution, enabling informed analysis and decision-making processes.

	Guest Satisfaction	***Metro Distance (km)***	***City Centre (km)***	***Person Capacity***	***Cleanliness Rating***
Mean	93	0.6103	2.7659	3	9
Median	95	0.3983	2.3828	3	10
Mode	100	0.5084	2.8021	2	10
Minimum	20	0.0032	0.0150	2	2
Maximum	100	14.2736	25.2846	6	10

Table 4: A descriptive Statistics of Continuous Variables

As represented in Table 4, we can gather valuable insights about the continuous predictor dataset for Airbnb properties in Europe. The average guest satisfaction score for these properties is notably high at 93 out of 100, showcasing the overall positive experiences of guests. Moreover, the proximity to essential amenities is evident, with the average distance to the nearest metro station being approximately 0.61 kilometers, ensuring convenient access to public transportation. Similarly, the average distance from the city center is around 2.77 kilometers, offering a balance between being close to urban attractions and providing a quieter residential setting. Regarding the capacity of these properties, they are generally well-suited for small to medium-sized groups, as the average person capacity is approximately 3 individuals. However, the dataset reflects variability in this aspect, ranging from accommodating 2 to 6 people. Furthermore, about 30% of the listed properties provide multiple rooms, likely appealing to families or larger travel groups, while the majority offer single-room accommodations, catering to individual travelers or smaller parties. Additionally, the cleanliness ratings are commendable, with the average cleanliness score standing at 9 out of 10. Guests seem to highly value cleanliness when choosing their accommodations, as reflected by the high ratings. However, it is worth noting that ratings range from 2 to a perfect 10, suggesting that while most properties maintain excellent cleanliness standards, some might have room for improvement in this aspect.

3.1.2.2. Exploring The Categorical Variables

Room Type

As shown in Figure 7 below, a significant proportion of room types available for Airbnb rentals in European cities consists of Entire homes/apartments, accounting for approximately 65% of the listings. Private rooms make up approximately 34% of the rentals, while a very small percentage, merely 1% of the 37,692 houses in our dataset, are listed as shared rooms. This observation highlights a clear preference for Entire homes/apartments as the most sought-after accommodation type for Airbnb users in European cities.

Private rooms also attract a substantial number of renters, offering a viable alternative to those seeking a more budget-friendly or intimate stay. On the other hand, shared rooms appear to be relatively uncommon, possibly appealing to a niche audience or limited availability of such offerings.

The preference for Entire homes/apartments on Airbnb in European cities could be due to both pricing and individual privacy concerns. Some guests prioritize personal space and may avoid shared rooms to ensure a comfortable and private stay. Understanding these preferences can help hosts tailor their listings and pricing strategies to cater to a diverse range of travelers.

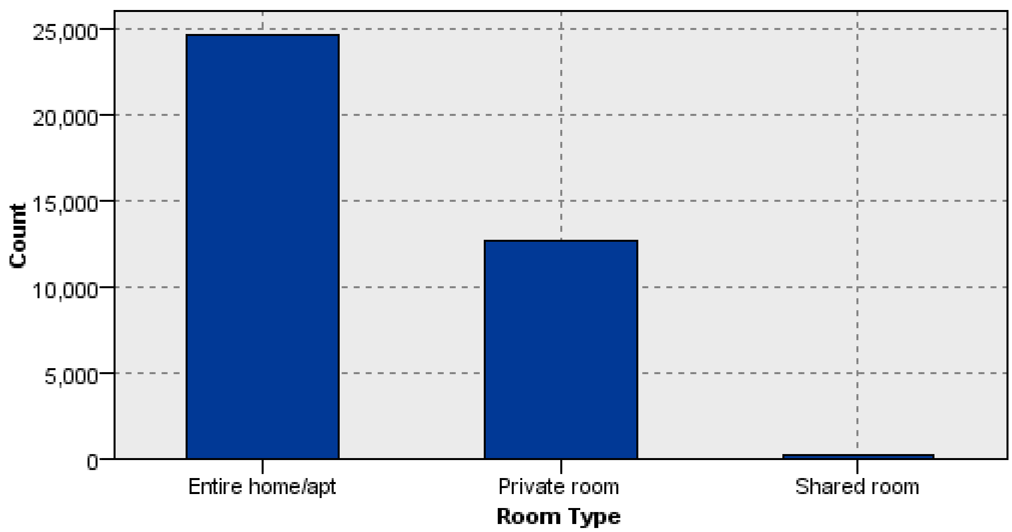

Figure 7: Bar chart showing the distribution of room types in the dataset.

Days

From Figure 8, it is interesting to see that people rent out Airbnb spaces on weekdays just as much as they do on weekends. This observation is unexpected, as one might anticipate higher demand for rentals during weekends when individuals typically have more leisure time for travel and vacations. However, the data suggests that the frequency of bookings or rentals for Airbnb spaces remains consistent throughout the week. Several factors could contribute to this pattern, such as business travel, local events, tourism patterns, or the emerging trend of workcations.

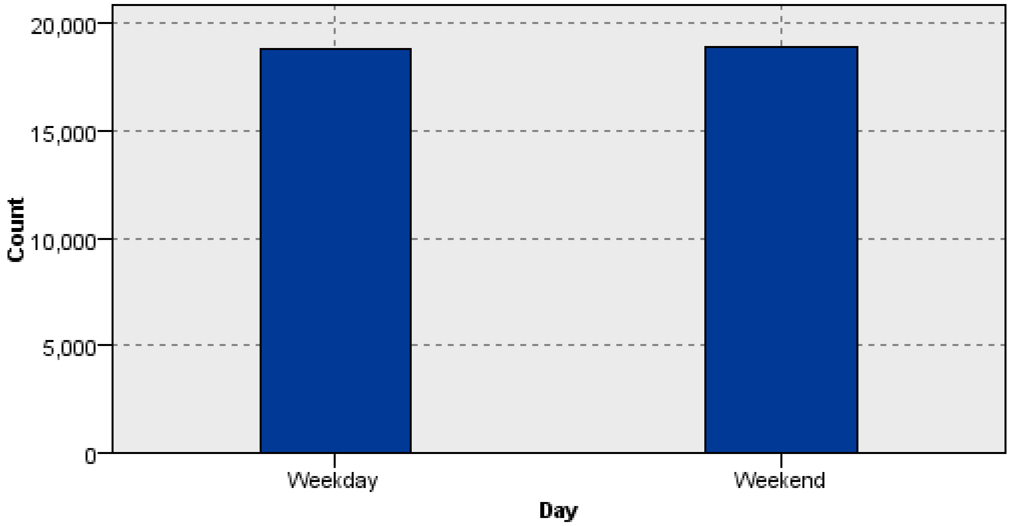

Figure 8: Bar chart showing the Days Airbnb Apartments are rented.

Bedrooms

Based on the data presented in Figure 9, it can be observed that approximately 72% of the Airbnb accommodations available in European cities consist of one bedroom. Following this, around 17% of the listings are for houses with two bedrooms. Interestingly, there is a notably small percentage of properties offering three or more bedrooms available for Airbnb rentals across European cities, as per the information extracted from our dataset. This information highlights the prevalent distribution of Airbnb listings based on the number of bedrooms, and it can be useful for both hosts and guests in understanding the accommodation options available in various European cities. The data may also be valuable for policymakers and stakeholders in the hospitality industry to gain insights into the market trends and preferences regarding the size of properties preferred for short-term rentals.

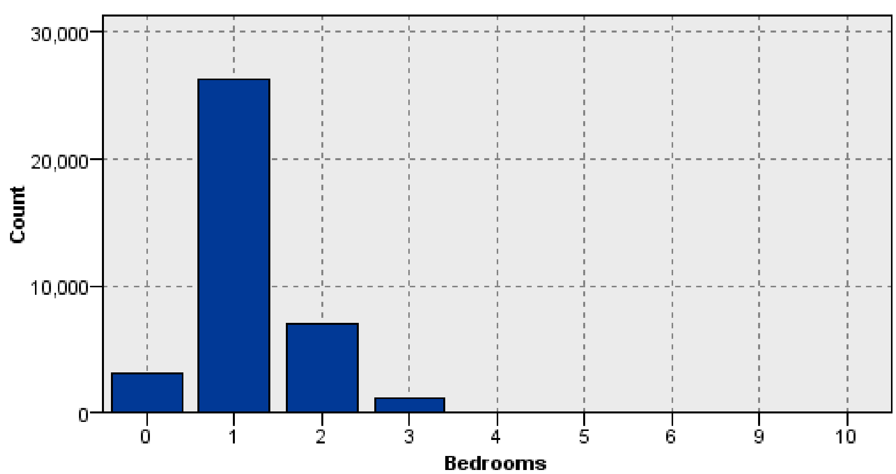

Figure 9: Distribution of bedrooms in Airbnb Rental Homes in Europe.

3.1.3. Correlation of Independent Variables to Price

We performed feature selection by generating a correlation matrix, which is presented below. The correlation matrix helps identify the predictors that exhibit strong correlations with our target variable, "price.". This feature selection step is essential in building an effective and efficient predictive model for the target variable "price."

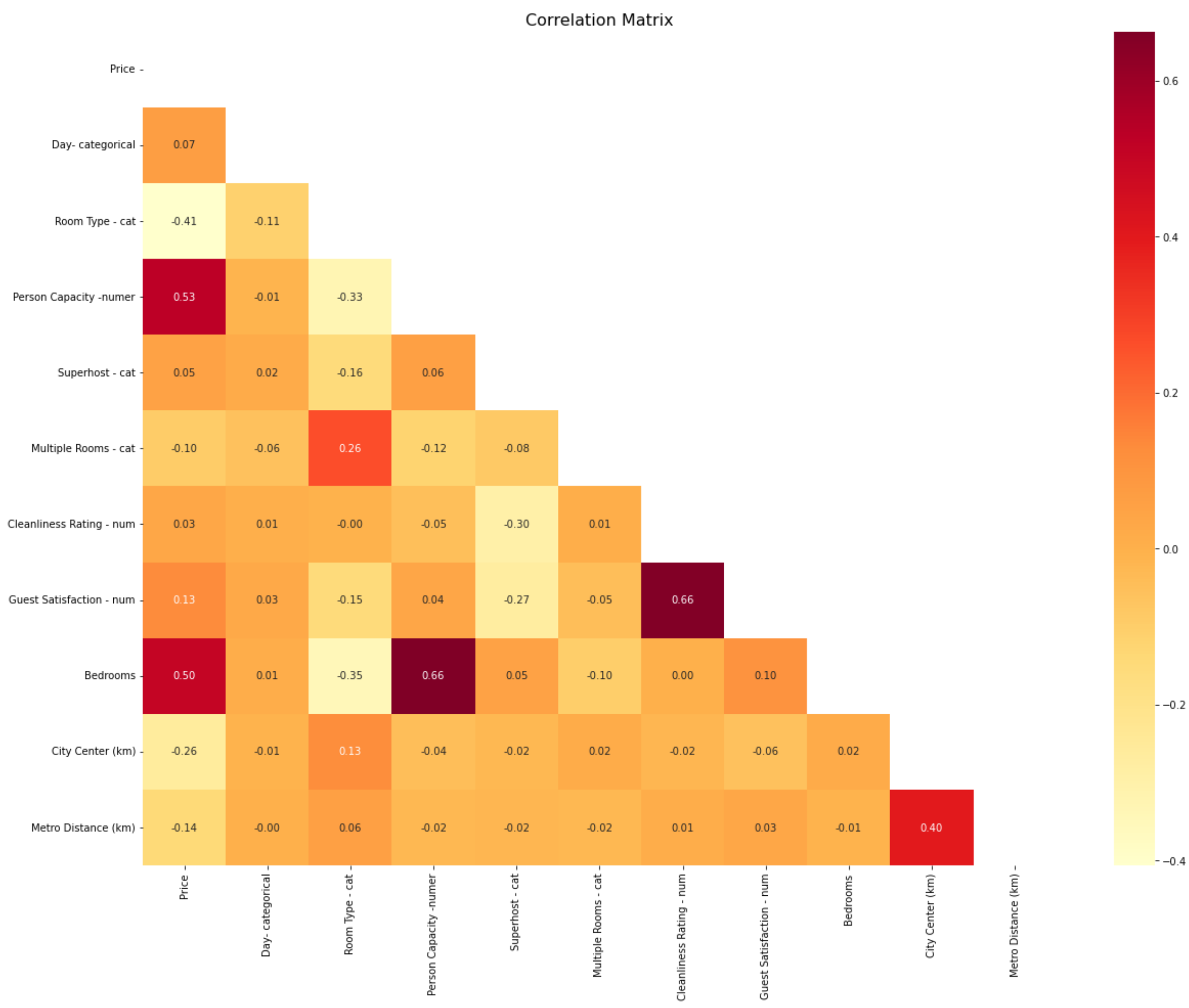

Figure 10: Correlation matrix of independent variables in relation to price

From Figure 10, it is evident that the "person capacity" feature exhibits the highest positive correlation with the price of accommodations in European cities. The correlation value of 0.53, while not particularly strong, indicates that a greater capacity of the house to accommodate guests corresponds to higher prices set by Airbnb hosts. This finding aligns with the observation that the number of bedrooms available in the house also shows a considerable positive correlation (0.50) with accommodation prices. Essentially, the more bedrooms a property has, the higher its price tends to be.

On the other hand, "guest satisfaction" comes next in terms of positive correlation, with a relatively low value of 0.13. This suggests that guest satisfaction has a minor impact on the prices hosts can charge for their Airbnb listings, based on the dataset used for analysis. Although this correlation is modest, it still holds more significance compared to other factors such as "day of week," "rating of the host as a superhost," and "cleanliness rating of the accommodation," all of which show even lower correlation values of 0.06, 0.04, and 0.03, respectively. Consequently, these latter features do not seem to play a critical role in determining pricing for Airbnb rentals in the European region.

Regarding negative correlations, "distance to the city" exhibits a higher negative correlation with price (-0.26) compared to "distance to the metro" (-0.14). However, both correlations indicate a weak relationship. This means that as the distance from the city centre or metro station increases, the price of the house tends to decrease, albeit by a minimal amount.

In conclusion, based on the analysis of the correlation matrix, the person capacity and number of bedrooms appear to be more influential factors in determining the prices of Airbnb accommodations in European cities. Guest satisfaction also plays a minor role, while other features like day of week, host rating, and accommodation cleanliness seem to have little impact on the pricing of Airbnb listings in the region. Additionally, the distance to the city or metro station has a weak negative relationship with price, suggesting that proximity to these locations may slightly influence accommodation prices.

3.2. Machine Learning Algorithm adopted.

3.2.1. Random Forest:

The Random Forest algorithm, proposed by L. Breiman in 2001, combines multiple randomized decision trees and aggregates their predictions through averaging. It has shown excellent performance in settings where the number of variables greatly outweighs the number of observations (Biau and Scornet, 2016). Random Forest is particularly well-suited for predicting Airbnb prices due to its ability to handle large datasets with numerous variables and its robustness against overfitting. It provides insights into feature importance, which will enable us to understand the factors influencing pricing in the Airbnb Market. Moreover, it can handle both categorical and numerical features, making it suitable for the diverse data encountered in Airbnb pricing prediction. In the realm of predicting Airbnb prices using machine learning techniques, the Random Forest model has gained significant prominence among scholars in the past. Its widespread usage and proven effectiveness in generating accurate predictions make it an ideal choice for adoption in this research. The Random Forest machine learning algorithm is an ensemble method.

3.2.2. Neural Network

The Neural Network Model, initially inspired by the learning process of the human brain, has emerged as a potent technique in supervised learning. Renowned for its capacity to optimize prediction outcomes by learning from training data and refining its predictive methods, this model undergoes an evaluation of accuracy using a separate test dataset (Yang, 2021). Its utility as a forecasting model has garnered substantial recognition across various domains, including social sciences, engineering, economics, business, finance, foreign exchange, and stock markets (Adebiyi, Adewumi, and Ayo, 2014). When applied to the prediction of Airbnb prices, neural networks showcase their efficacy. Through its ability to analyze intricate data patterns, these networks can grasp the multitude of factors that impact Airbnb pricing, encompassing aspects like location, property attributes, availability, seasonality, and market demand. By training the network with pertinent features and historical data, it can uncover underlying patterns and relationships, leading to precise predictions of Airbnb prices. What sets the Neural Network machine learning method apart is its proficiency in capturing non-linear relationships between the dependent variable and the independent variables.

3.2.3. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) stands out as a powerful tree-based regression model widely employed in various machine learning applications. In a study conducted by Yang in 2021, XGBoost was among several machine learning techniques utilized to forecast Airbnb prices in Beijing. The research revealed that XGBoost outperformed other models, demonstrating its capability to yield superior results. Specifically, it exhibited the lowest Mean Squared Error (MSE) and highest R-squared (R^2) values, highlighting its exceptional predictive accuracy.

XGBoost's effectiveness in predicting Airbnb prices can be attributed to its ensemble nature, which involves constructing an ensemble of decision trees to collectively generate precise predictions. This methodology proves particularly advantageous for predictions such as Airbnb price, where complex relationships and interactions among various features impact the final outcome. Notably, XGBoost's ability to handle sizable datasets, exemplified by the datasets employed in this study, positions it as a favorable machine learning technique for predicting prices within the European Airbnb markets. Another noteworthy attribute of XGBoost is its adeptness at mitigating overfitting, a common pitfall in model development. XGBoost incorporates mechanisms to counteract this tendency, enhancing the model's ability to generalize effectively to new instances. In the context of predicting Airbnb prices, this attribute is of paramount importance as it ensures the model's reliability and robustness in real-world scenarios.

In essence, XGBoost's amalgamation of ensemble-based decision trees, capacity to handle extensive datasets, and resilience against overfitting renders it a highly suitable and proficient choice for predictive modeling tasks, particularly in domains like Airbnb price prediction where accuracy and generalization are pivotal.

3.3. Justification for Selected Methods over Regression for Airbnb Price Prediction in European Cities

Based on the extensive literature review and the proven success of various scholars in predicting Airbnb prices using machine learning methods, this research has deliberately opted for machine learning techniques over the traditional multivariable regression model for the following reasons:

- **Complexity and Number of Variables:** Airbnb pricing is influenced by a multitude of factors, including location, property attributes, availability, seasonality, market demand, and more. The multivariable regression model may struggle to capture the intricate and often non-linear relationships among these variables effectively. The selected models are adept at handling complex relationships and patterns within the data, making them better suited for this study.
- **Dataset Size and Dimensionality:** Machine learning models are particularly well-equipped to handle large datasets with numerous variables, which is the case in our analysis of Airbnb pricing prediction. Random Forest, for instance, can efficiently process vast amounts of data, ensuring that no valuable information is lost due to data reduction. In contrast, traditional regression models may encounter difficulties with high-dimensional data.
- **Robustness Against Overfitting:** Overfitting is a common concern in predictive modeling, where a model becomes too closely tailored to the training data, leading to poor generalization of new data. Machine learning algorithms, such as Random Forest and XGBoost, incorporate mechanisms to mitigate overfitting, ensuring that the models perform well on unseen data. This is crucial for building reliable predictive models for Airbnb prices, where generalizability is key.
- **Feature Importance Insights:** Machine learning models provide valuable insights into feature importance, enabling us to understand which factors most significantly influence Airbnb pricing. This knowledge can be invaluable for hosts and policymakers, as it helps identify the drivers behind price fluctuations. Traditional multivariable regression models do not offer this level of feature

interpretability. Non-Linear Relationships: Airbnb pricing often involves complex, non-linear relationships between independent variables and the dependent variable (price). Machine learning models, particularly Neural Networks, excel at capturing these non-linear patterns, which may go unnoticed by linear regression models.

- **Proven Effectiveness for Airbnb Dataset:** The adoption of machine learning methods, such as Random Forest and XGBoost, has been prevalent among scholars reviewed in this research, and their effectiveness in generating accurate predictions for Airbnb prices has been well-documented. These models have consistently outperformed traditional regression models in terms of predictive accuracy.

3.4. Prediction Accuracy Measures

Prediction accuracy measures are essential for evaluating the performance of our predictive models. These measures will provide quantitative metrics that will allow us to compare and assess the accuracy and reliability of different models that will be built in this research. By understanding and utilizing these measures, we can make informed decisions, select the most suitable models, and gain confidence in their predictive power. In this section, we will explore various prediction accuracy measures, their calculations, interpretations, and their practical applications in assessing the accuracy of predictive models. As this is a regression problem because our target variable is a numerical, we will adopt the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

3.3.1. Mean Absolute Error (MAE)

In Hodson's 2022 study, the significance of Mean Absolute Error (MAE) as a conventional metric for model evaluation is reaffirmed. Additionally, the research conducted by Willmott and Matsuura in 2005 adds weight to this notion by highlighting MAE's distinctive advantage of offering a clear and unambiguous representation of average error. Their findings stress the appropriateness of MAE, especially in scenarios necessitating dimensional evaluations and comparisons of average model-performance errors. As such, both studies converge on the importance of MAE as a robust tool for accurate and insightful model assessment. The formular for the MAE is represented in (1).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}. \tag{1}$$

3.3.2. Root Mean Squared Error (RMSE)

The RMSE (Root Mean Squared Error) is a widely used metric for evaluating the accuracy of predictive models. It quantifies the average magnitude of differences between predicted and true values, making it effective in accurately measuring model performance. The RMSE's ability to penalize larger errors and maintain compatibility with the original scale of the data enhances its usefulness in assessing pricing accuracy in the dynamic Airbnb price market. However, it is important to consider the potential impact of outliers on the RMSE when evaluating models. It is calculated (2) below:

$$RSME = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N - P}} \tag{2}$$

3.5. Measure of Goodness of Fit

The Coefficient of Determination (R^2) is a pivotal metric utilized to ascertain the quality of fit achieved by a model. In essence, R-squared quantifies the proportion of variance in the dependent variable that is accounted for by the model. Its fundamental role lies in evaluating the extent to which the model can explain the variability within the data. In the context of predicting Airbnb prices across distinct European cities, the application of R^2 serves as a crucial analytical tool. This metric allows for a comparative assessment of various models, facilitating the identification of the model that best captures the dissimilarities between the actual observed data and the model's fitted values. A higher R^2 value signifies a stronger alignment between the model's predictions and the true outcomes, indicative of a superior explanatory capacity of the model. Moreover, R^2 holds significance beyond its role in model comparison. It provides valuable insights into the appropriateness of the selected predictors and the overall reliability of the model's predictions. A high R^2 indicates that a substantial portion of the variation in the dependent variable can be attributed to the predictors, validating the model's explanatory prowess. Conversely, a low R^2 may suggest that the chosen predictors inadequately account for the observed variability, prompting a reevaluation of the model's feature selection and underlying assumptions.

In conclusion, R^2 emerges as a pivotal measure in assessing the goodness of fit of predictive models. Its utilization within the analysis of Airbnb price prediction across diverse European cities offers a comprehensive means to gauge model performance, ultimately guiding the selection of the most suitable and informative model for the task at hand.

3.6. Data Pre-Processing

Within this section, our attention turns to the crucial task of meticulously preparing the dataset to facilitate the ensuing modeling analysis. This pivotal stage involves a series of concerted efforts aimed at enhancing the dataset's quality and relevance, ultimately laying the foundation for robust predictive modeling. The undertaken endeavors are geared towards not only refining the raw data but also imbuing it with attributes that foster meaningful insights. Through meticulous data cleansing, feature engineering, and other preparatory measures, we endeavor to bolster the dataset's efficacy in unraveling the intricate nuances of Airbnb pricing dynamics in European cities.

3.6.1. Feature Engineering Effort

This section encompasses all the feature engineering endeavors undertaken in this study to enhance dataset quality and prepare for model construction.

3.6.1.1. Encoding of Categorical Variables

In this phase, we transformed specific categorical variables as shown in the table below into dummy binary variables. This conversion allowed us to represent categorical information in a format suitable for our predictive model. By creating dummy variables, we converted each category within those variables into binary representations 0, 1 and 2 which represented the different categories of the variables. Treating these dummy binary variables as explanatory variables in our model specification enables us to incorporate the categorical information into our predictive analysis. This process is essential as machine learning algorithms require numerical inputs rather than categorical ones. By doing so, we can effectively capture the effects of these categorical dependent variables have on the price and assess their significance in predicting the outcomes. By including these transformed explanatory variables in our model, we aim to improve the accuracy and comprehensiveness of our price prediction model. The predictive power of our model is enhanced by considering the various categorical features that influence Airbnb prices in different European cities. Table 6 shows the encoding of the variables that were created to replace our categorical variables in the analysis.

	Variables	Dummy Variables
Day	Weekday	0
	Weekend	1
Room Type	Entire home/apt	0
	Private room	1
	Shared Room	2
Super host	TRUE	0
	FALSE	1

Table 5: Dummy variables for Categorical Variables.

3.6.1.2. Data Quality Assessment and Treatment – Outliers, extreme quality

Measures were implemented to enhance data quality and ensure the integrity of the data, with the ultimate goal of building an accurate predictive model. As observed in the price distribution analysis mentioned earlier, there were extreme values in the dataset that deviated significantly from the typical Airbnb listing prices. These extreme values, known as outliers, could potentially hinder the accuracy of our predictive model.

To address this issue and avoid skewness in our analysis, we employed the Interquartile Range (IQR) method to identify the upper and lower bounds and subsequently removed the outlier values. This process resulted in the exclusion of 3644 rows, accounting for approximately 10% of the original dataset, leaving us with a refined dataset consisting of 34049 rows and 12 columns. The histogram depicting the price distribution after removing the outliers can be found in Figure 6.

Additionally, certain columns were eliminated from the dataset to prevent potential issues of multicollinearity. Specifically, a column related to 'private room' or 'shared room' information was removed since this data could be inferred from the existing 'Room Type' column. Moreover, two columns, including the Attraction Index and Normalised Attraction Index, were also removed, as they were redundant and not contributing significantly to the analysis. Furthermore, five columns with null values were excluded to ensure data completeness and integrity, considering that these columns contained missing data for approximately 90% of the entries. By taking these steps, we aimed to ensure the reliability and accuracy of the dataset, thereby enabling us to build a robust predictive model that can provide valuable insights into Airbnb pricing patterns. Figure 11 reflects the distribution of prices in our dataset following the removal of outliers in all our continuous variables.

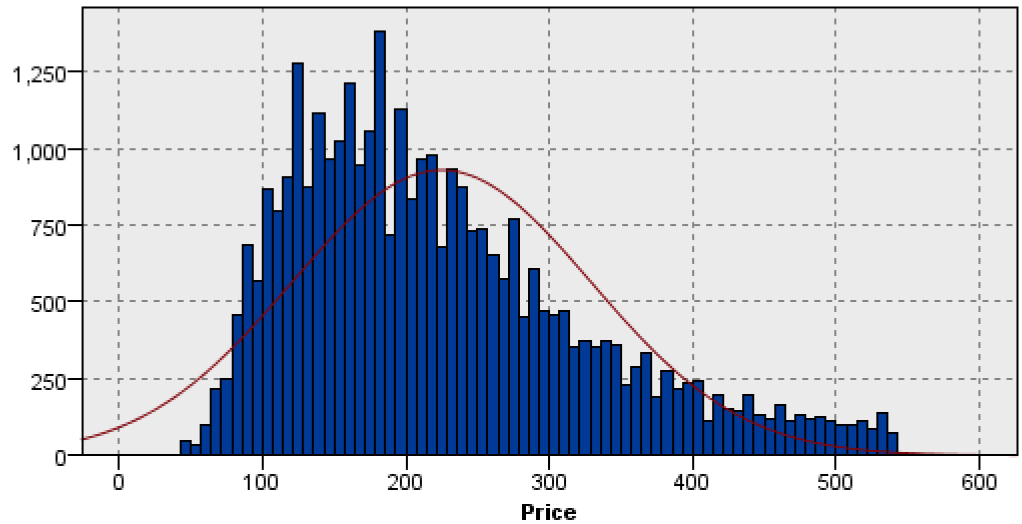

Figure 11: Revised histogram showing distribution of Price in the dataset without Outliers.

3.6.2. Data partitioning

When constructing a predictive model, it is crucial to divide the dataset into different subsets. For this analysis, the dataset will be split into two main sets: the training set and the testing set. The training which will set was used to establish the mathematical relationship between the target variable (the Airbnb price) and the input variables (the independent variable as stated in Table 2). This process involved training the model on the training set, allowing it to learn patterns and correlations within the data. Once the model is trained, it will then be applied to the test sample. The test sample for this analysis serves as an unseen dataset that the algorithm has not encountered during training. By evaluating the model's performance on this independent set, we can assess how well the model generalizes to new, unseen data. This approach ensures that the model's effectiveness can be assessed on new data that it has not been previously exposed to. In this research, 80% of the data will be assigned to the training set, while the remaining 20% will be allocated to the testing set. The random state in the train-test split was set to 8 for reach of the algorithms, which was done to ensure consistent splitting of the data, regardless of how many times the split is performed the consistency of which helped maintain the stability of the model's performance, as the same data points are consistently used for training and testing, leading to consistent model evaluation.

3.7. Tools

For this study, Outliers within the dataset underwent meticulous treatment using the Excel platform, ensuring the robustness of subsequent analyses. For comprehensive data exploration, the SPSS Modeler software played a pivotal role in unraveling intricate patterns. When it came to the construction of predictive models, Python emerged as the tool of choice. The decision to employ Python and Excel was driven by their exceptional efficiency and versatility in addressing different facets of the analysis. Python's extensive array of libraries and packages made it the go-to option for data cleaning, preparation, and visualization. The language's prowess in generating informative and insightful visualizations allowed us to gain a deeper understanding of the dataset's dynamics. Excel, on the other hand, proved indispensable in its ability to effectively handle outliers, reinforcing the integrity of our subsequent analytical endeavors. This dual-tool approach underscored our commitment to meticulous data treatment, exploration, and model development, paving the way for a comprehensive and robust analysis of Airbnb pricing trends in European cities.

CHAPTER FOUR: MODEL ANALYSIS AND RESULTS

4. Result of the Analysis and Findings

The section provides a comprehensive analysis of the findings obtained in the context of leveraging machine learning techniques for price prediction in the selected European cities. We will delve into the results derived from the diverse range of models employed in this study, conducting a thorough investigation to extract valuable insights. A concise summary will be presented, focusing on the key features and essential findings that have emerged from these results. It is of utmost importance to ensure the coherence and dependability of the constructed models, as their primary objective is to deliver robust and accurate price predictions for medium to long-term planning. The models adopted must exhibit consistency and provide reliable forecasts that are reasonably approximate, validating their efficacy in practical applications. This first section will focus solely on using the performance metrics to show how the predictive models performed in the various European cities. The next section will focus on the best performing models for each of the European cities and showing the using the R^2 to show the goodness of fit of the model.

4.1. Comparison of Model Results Across European Cities

In this comprehensive analysis of the various models used in this analysis, we evaluated and compared the performance of three predictive models - Random Forest, Neural Networks, and XGBoost - in predicting Airbnb prices across eight different European cities. The assessment was based on two performance metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), for both the training and testing datasets. Let's delve deeper into the results for each model and compare them to see how these models performed across the various cities.

Random Forest

The first analysis that was conducted was done by building the Random Forest model in predicting Airbnb prices across the eight different European cities. The model's accuracy was assessed using the various aforementioned metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2) scores for both the training and testing datasets.

Model	Cities	Training Performance		Testing Performance	
		MAE	*RMSE*	*MAE*	*RMSE*
Random Forest	Amsterdam	17.53	22.86	41.01	52.69
	Athens	9.78	15.53	25.87	41.44
	Barcelona	14.93	21.63	37.77	54.67
	Berlin	13.70	20.51	34.25	53.00
	Lisbon	11.73	17.46	31.75	45.64
	Paris	17.71	24.21	44.62	60.87
	Rome	11.69	17.10	30.03	43.83
	Vienna	12.41	17.72	33.55	51.04

Table 6: Random Forest Model Performance Accuracy Output

The model demonstrated strong performance during training in Amsterdam, achieving a low MAE of 17.53 and an RMSE of 22.86. However, during testing, a slight decrease in performance was observed, with an MAE of 41.01 and an RMSE of 52.69. This suggests that the model's predictions had larger errors when applied to unseen data, indicating challenges in generalizing to new data points in Amsterdam.

In Athens, the model's training performance resulted in a relatively low MAE of 9.78 and an RMSE of 15.53. These lower values suggest that, on average, the predicted prices had an absolute difference of 9.78 currency units from the actual prices during training. However, during testing, the model's performance was slightly less accurate with an MAE of 25.87 and an RMSE of 41.44. While there was an increase in errors, the overall performance remains reasonable, indicating that the Random Forest model performs well as a suitable model for Airbnb price prediction in Athens.

The Random Forest model's output for Barcelona showed an MAE and RMSE of 14.93 and 21.63, respectively, in the training dataset. Similarly, in the testing set, the performance slightly deteriorated, resulting in higher values of 37.77 for MAE and 54.67 for RMSE. This consistent pattern aligns with the findings for Amsterdam and Athens, highlighting the model's performance trend across these cities.

For Berlin, the model's training set produced an MAE of 13.70 and an RMSE of 20.51. While the error margin increased slightly in the testing dataset, the Random Forest model still performed well with MAE and RMSE values of 34.25 and 53.00, respectively. This suggests that the model's predictions are capable of providing reasonable price estimates for Airbnb listings in Berlin.

The Random Forest model's training performance in Rome resulted in an MAE of 11.69 and an RMSE of 17.10. These values indicate that, on average, the predicted prices during training had an absolute difference of 11.69 currency units from the actual prices. Similarly, in testing, the model's performance remained reasonably consistent, yielding an MAE of 30.03 and an RMSE of 43.83. This suggests that the model's predictive capabilities extended to Airbnb price prediction in Rome, even when applied to unseen data.

In Lisbon, the Random Forest model exhibited good performance during training, with an MAE of 11.73 and an RMSE of 17.46. These values indicate that the average absolute difference and squared difference between predicted and actual prices were relatively low during training. During testing, the model's performance experienced a slight decline, with an MAE of 31.75 and an RMSE of 45.64. Despite the increase in errors, the model's predictions for Airbnb prices in Lisbon remained reasonable, demonstrating its ability to generalize across new data points.

Similar to the previous cities, the pattern of increased MAE and RMSE errors when transitioning from the training set to the testing set was observed in Paris and Vienna. Despite these increased errors, the Random Forest model maintained an acceptable level of performance in these cities as well.

Overall, the Random Forest model demonstrated promising performance in predicting Airbnb prices across European cities. Notably, the model performed exceptionally well in Rome and Lisbon, with Athens also showcasing strong results. These cities displayed low MAE and RMSE values in the testing dataset, indicating that the model's predictions were relatively accurate and precise. It is worth noting that the model's generalization ability varied across cities, with some cities exhibiting better performance in the training dataset than in testing. This variability suggests that the model might face challenges in accurately predicting Airbnb prices in certain unseen scenarios.

Neural Networks

Model	Cities	Training Performance		Testing Performance	
		MAE	*RMSE*	*MAE*	*RMSE*
Neural Networks	Amsterdam	59.52	73.16	57.14	68.53
	Athens	37.72	54.63	38.20	55.97
	Barcelona	53.20	71.45	50.36	65.25
	Berlin	53.12	69.63	50.18	68.67
	Lisbon	45.41	60.23	47.76	63.17
	Paris	69.87	86.64	71.17	89.16
	Rome	47.35	65.63	45.97	65.54
	Vienna	47.73	63.13	48.88	66.18

Table 7: Neural Network Model Performance Accuracy Output

While evaluating the performance of the Neural Networks model in our prediction of Airbnb prices across the eight different European cities. We gained the following insights:

The Neural Networks model demonstrated strong predictive performance in Athens. During the training phase, it achieved an MAE of 37.72, indicating that, on average, its predictions were only about 37.72 off from the actual prices. The RMSE value of 54.63 signified that the squared differences between predictions and actual prices were relatively low during training. As the model moved to testing, the MAE increased slightly to 38.20, and the RMSE rose to 55.97. Despite this increase, the model's ability to generalize effectively to new data is apparent, as it continued to provide accurate price predictions for Athens.

Similar to Athens, the Neural Networks model performed well in Rome. It achieved an MAE of 47.35 during training, indicating a reasonably small average absolute difference between its predicted and actual prices. The corresponding RMSE value of 65.63 reinforced this accuracy. In testing, the model maintained its proficiency, with a reduced MAE of 45.97 and an RMSE of 65.54. These values underscore the model's ability to consistently and accurately forecast Airbnb prices in Rome, even when presented with previously unseen data, as demonstrated by the reduction in both MAE and RMSE values.

The model's performance in Lisbon was also noteworthy. During training, it achieved an MAE of 45.41, suggesting its predictions were, on average, about 45.41 currency units away from actual prices. The RMSE value of 60.23 reinforced its accuracy during training. In the testing phase, the model's predictions remained stable, with an MAE of 47.76 and an RMSE of 63.17. These values indicate that the model successfully retained its ability to generalize well to new data in Lisbon.

In Vienna, the Neural Networks model demonstrated consistent performance. During training, it achieved an MAE of 47.73 and an RMSE of 63.13, indicating its proficiency in capturing pricing patterns. These capabilities were maintained during testing, with an MAE of 48.88 and an RMSE of 66.18. These values suggest that the model's predictions in Vienna were reliably accurate both during training and when applied to new data.

Contrastingly, the Neural Networks model encountered some challenges in accurately predicting Airbnb prices in the cities of Amsterdam, Barcelona, Berlin, and Paris. During the training phase, the model's MAE values ranged from 53.12 to 69.87, reflecting notable errors in its predictions. Similarly, the corresponding RMSE values ranging from 69.63 to 86.64 underscored the model's little difficulty in capturing the nuances of these cities' pricing dynamics. However, the model demonstrated some improvements during testing with some cities, with Amsterdam, Berlin and Barcelona having a reduced MAE and RMSE in the testing set which implies that the model performed well to unseen data in those regions. However, For Paris the MAE and RMSE values increased to 71.17 and 89.16 respectively.

Overall, The Neural Networks model showcased varying performance across different cities. While it exhibited impressive predictive capabilities in Athens, Rome, Lisbon, Vienna, and Barcelona, the model's performance was less consistent in Amsterdam, Berlin, and Paris. The cities with lower MAE and RMSE values demonstrated more consistent patterns of improvement from training to testing, which indicates that the model was able to adapt well to new data in these cases.

XGBoost

Model	Cities	Training Performance		Testing Performance	
		MAE	***RMSE***	***MAE***	***RMSE***
XgBoost	Amsterdam	7.18	10.16	38.04	49.70
	Athens	14.78	20.75	27.97	43.12
	Barcelona	13.14	18.86	37.82	55.71
	Berlin	8.64	12.33	32.93	51.68
	Lisbon	18.30	25.37	34.67	48.97
	Paris	29.73	40.02	50.93	66.66
	Rome	22.31	30.88	33.60	46.76
	Vienna	12.84	17.87	33.81	51.04

Table 8: XGBoost Model Performance Accuracy Output

During training, the XGBoost model achieved an MAE of 14.78 and an RMSE of 20.75. These values indicated that, on average, the predicted prices had a moderate absolute difference and squared difference from the actual prices during training. However, during testing, the model's performance worsened, resulting in an increased MAE of 27.97 and an RMSE of 43.12. This suggests that the model's predictions had larger errors when applied to unseen data in Athens, indicating challenges in generalizing effectively.

Similarly, In Berlin, the XGBoost model's training performance yielded an MAE of 8.64 and an RMSE of 12.33, indicating accurate predictions during training. The model's testing performance remained reasonably strong, with an MAE of 32.93 and an RMSE of 51.68. This indicates that the model's generalization ability was effective in maintaining its accuracy across new data in Berlin, even though errors increased during testing.

The XGBoost model's performance in Barcelona during training resulted in an MAE of 13.14 and an RMSE of 18.86, indicating relatively accurate predictions. In testing, the model's performance remained consistent, with an MAE of 37.82 and an RMSE of 55.71. This suggests that the model was able to generalize well to unseen data in Barcelona while maintaining its accuracy, despite the increased errors.

For Vienna, the XGBoost model achieved an MAE of 12.84 and an RMSE of 17.87 during training, indicating accurate predictions on average. As the model transitioned to testing, it maintained a similar level of performance, with an MAE of 33.81 and an RMSE of 51.04. This suggests that the model's ability to generalize to new data remained consistent in Vienna, but errors increased when applied to unseen data.

Interestingly, Amsterdam achieved the lowest MAE and RMSE in the Training performance, however, the performance significantly increased in the testing performance. Paris on the other hand, produced the highest MAE and RMSE values both the training and testing performances.

Overall, the XGBoost model exhibited varying levels of performance across different cities. While it demonstrated stronger predictive capabilities in Vienna, Berlin, Barcelona, and Rome, the model's performance was less consistent in Amsterdam, Lisbon, and Paris. The cities with lower MAE and RMSE values showed the model's effectiveness in maintaining accuracy during training, but the increased errors during testing indicated that there's room for improvement in generalization to new data.

4.2. Forecast Patterns for Best Performing Models for Each European Cities

In this section, we will employ a line graph as a powerful visual tool to gain deeper insights and comprehensively evaluate the accuracy between the predicted prices and the actual values for each of the European cities. This assessment will be based on the utilization of the most effective predictive models available. The primary purpose of the line graph is to showcase the patterns and variations that exist between the predicted price values and the corresponding ground truth values. By plotting these two sets of data over time or any relevant dimension, we can easily observe how closely the predictions align with the actual outcomes for each city. Through this graphical representation, we will identify any potential discrepancies or deviations between the predicted prices and the actual values based on the best-performing models for each of the cities. This allows us to gauge the effectiveness of the predictive models and assess the degree to which they accurately capture real-world trends and fluctuations in the Airbnb industry for each European city under consideration. Furthermore, the line graph serves as an essential tool for assessing the overall goodness of fit of the predictive models for the given dataset. It helps us determine whether the models adequately capture the underlying relationships and dynamics between the variables, thereby indicating their overall efficacy and reliability. Alongside the line graph, we will present a comprehensive discussion of the R^2 values (coefficient of determination) for each city.

The R^2 values provides a measure of how well the predictive models explain the variance in the actual price data for the European City. The Higher R^2 values indicates a stronger correlation between the predicted and actual values, implying a more accurate model. Comparing and analyzing these R^2 values for each city will allow us to identify which predictive models have performed the best in terms of accuracy and predictive power.

In conclusion, the line graph and the accompanying analysis of R^2 values will provide valuable insights into the performance and accuracy of the predictive models in the context of European city Airbnb prices.

Certain hyperparameters were adjusted to enhance the performance of the most optimal model in each city. An exhaustive list of all tunings to all for each of cities can be found in Appendix 2.

Amsterdam

City	Model	R^2	
		Train	***Test***
Amsterdam	Random Forest	0.94	0.65
	Neural Network	0.39	0.41
	XGBoost	0.99	0.67

Table 9: Coefficient of Determination(R^2) - Amsterdam

The **XGBoost model** demonstrated superior performance in predicting Airbnb prices for Amsterdam with an R^2 value of 67% which indicates that the independent variables of factors affecting Airbnb prices in Amsterdam explains 67% of the price variatrion. The XGBoost amongst the other models also achieved the lowest MAE and RMSE for the City of Amsterdam. The scatter plot below depicts the predicted and actual price values reveals a tight cluster of points, indicating a strong alignment between the predicted and actual values. This suggests that the model's predictions closely match the actual prices, signifying its accuracy in capturing the underlying patterns in the data. The small variation between the predicted and actual values signifies that the model provides a reliable estimation of Airbnb prices for Amsterdam.

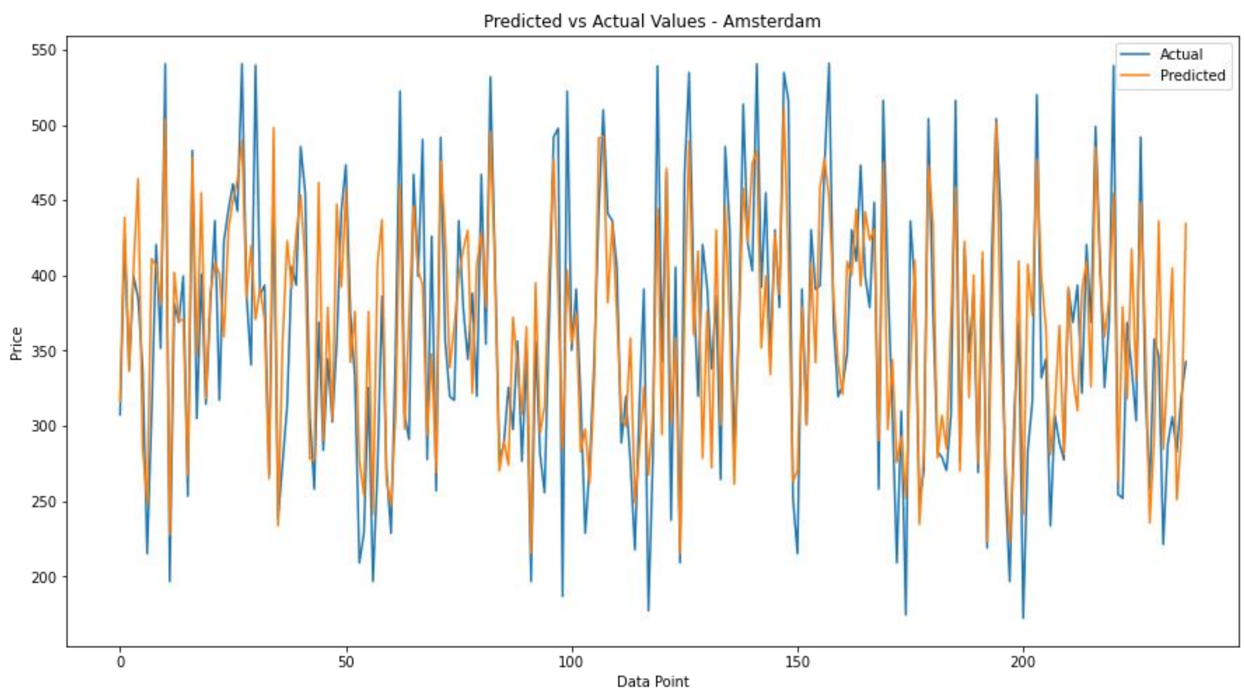

Figure 12: Line chart showing Prediction Vs Actual Values – Amsterdam

Athens

City	Model	R^2	
		Train	***Test***
Athens	Random Forest	0.95	0.65
	Neural Network	0.36	0.37
	XGBoost	0.91	0.63

Table 10: Coefficient of Determination(R^2) - Athens

Among the predictive models utilized, **Random Forest** demonstrated the strongest performance for Athens. Its high R^2 value of 0.65 indicates that the model effectively explains approximately 65% of the price variations in the Athens dataset, showcasing its predictive capability. Similarly, in the performance metric as explained in 4.1 above, the Random Forest model achieved the lowest MAE and RMSE in both the training and testing dataset indicating it is capability to predict Airbnb prices more accurately in Athens when compared to the other Models in this analysis.

The line chart below visually illustrates the predicted prices against the actual values, revealing minimal deviations between the two. This indicates that the Random Forest model is capable of making accurate predictions for Airbnb prices in the Athens region of Europe. Its ability to closely approximate the actual prices suggests that it is a reliable tool for price prediction in this particular city.

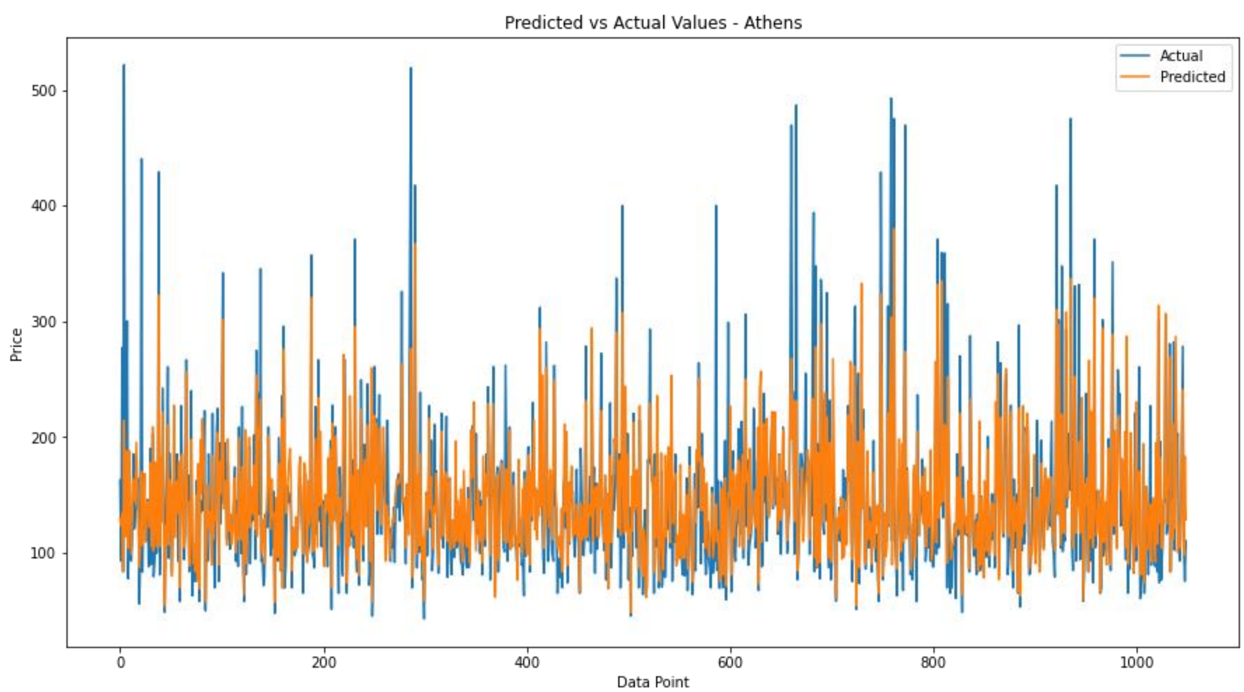

Figure 13: Line chart showing Prediction Vs Actual Values – Athens

Barcelona

City	Model	R^2	
		Training Performance	***Testing Performance***
Barcelona	Random Forest	0.96	0.71
	Neural Network	0.57	0.56
	XGBoost	0.97	0.69

Table 11: Coefficient of Determination(R^2) - Barcelona

Like Athens, the **Random Forest** model also outperformed the other predictive models in the Barcelona region. Although the XGBoost model achieved a close R^2 value of 0.69, the Random Forest model obtained an even higher R^2 value of 0.71, indicating its superior ability to make accurate predictions for Airbnb prices in the region while the Neural Networks performed relatively low with R^2 value of 0.47. In addition, the MAE and RMSE achieved for Random Forest was the lowest compared to the results for Neural Network and XGBoost.

The line chart below provides a visual representation of the predicted values using the Random Forest while comparing them to the actual values datapoint, illustrating the model's accuracy. The closely clustered data points demonstrate the model's capability to closely approximate the actual prices, solidifying its effectiveness in predicting Airbnb prices in the Barcelona region.

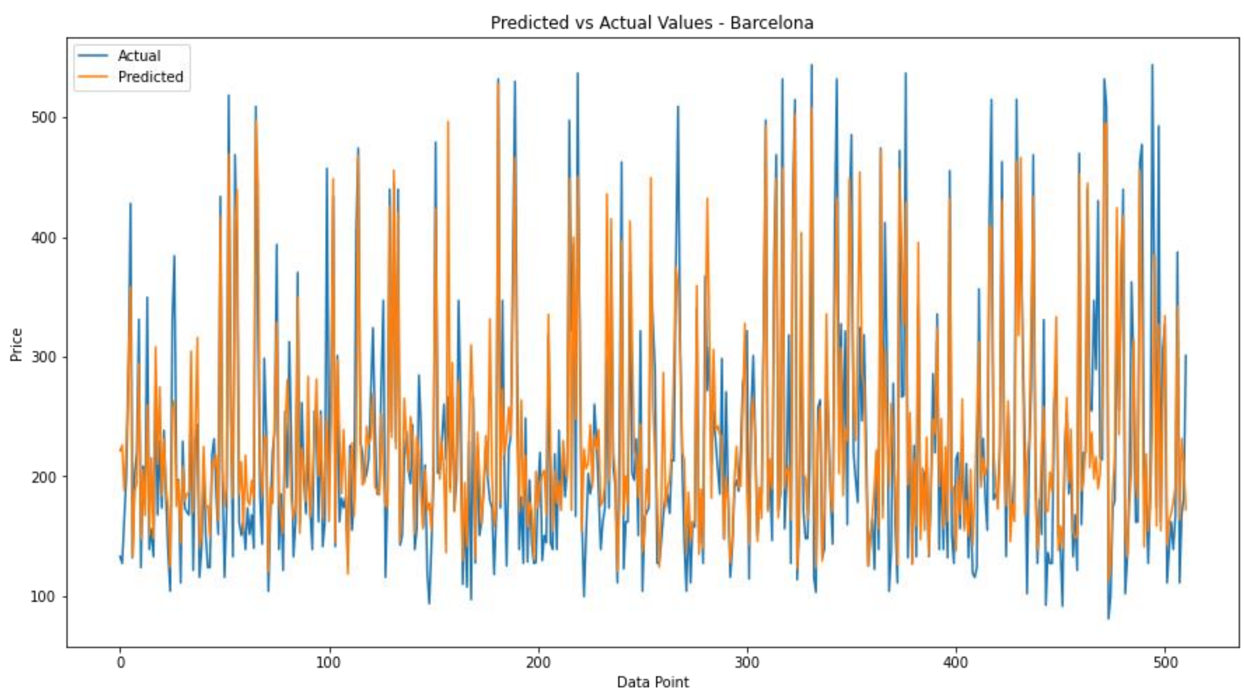

Figure 14: Line chart showing Prediction Vs Actual Values – Barcelona

Berlin

City	Model	R^2	
		Training Performance	***Testing Performance***
Berlin	Random Forest	0.96	0.71
	Neural Network	0.49	0.51
	XGBoost	0.98	0.73

Table 12: Coefficient of Determination(R^2) - Berlin

Among the range of predictive models investigated in this study, **XGBoost** emerged with the highest R^2 value of 0.73. This value suggests that the model can effectively account for around 73% of the fluctuations in prices within the dataset, signifying its aptness for price prediction in the Berlin area of Europe. Notably, all models achieved favorable outcomes, with none registering an R^2 value below 50% in testing performance. Nonetheless, in alignment with the accuracy assessment presented in Section 4.1, XGBoost demonstrated the lowest MAE and RMSE values. This outcome underscores its suitability for precise Airbnb price prediction in the Berlin region. As depicted in the Line chart below, the XGBoost model showcases minimal disparities between the predicted and actual values, reinforcing its effectiveness in price prediction.

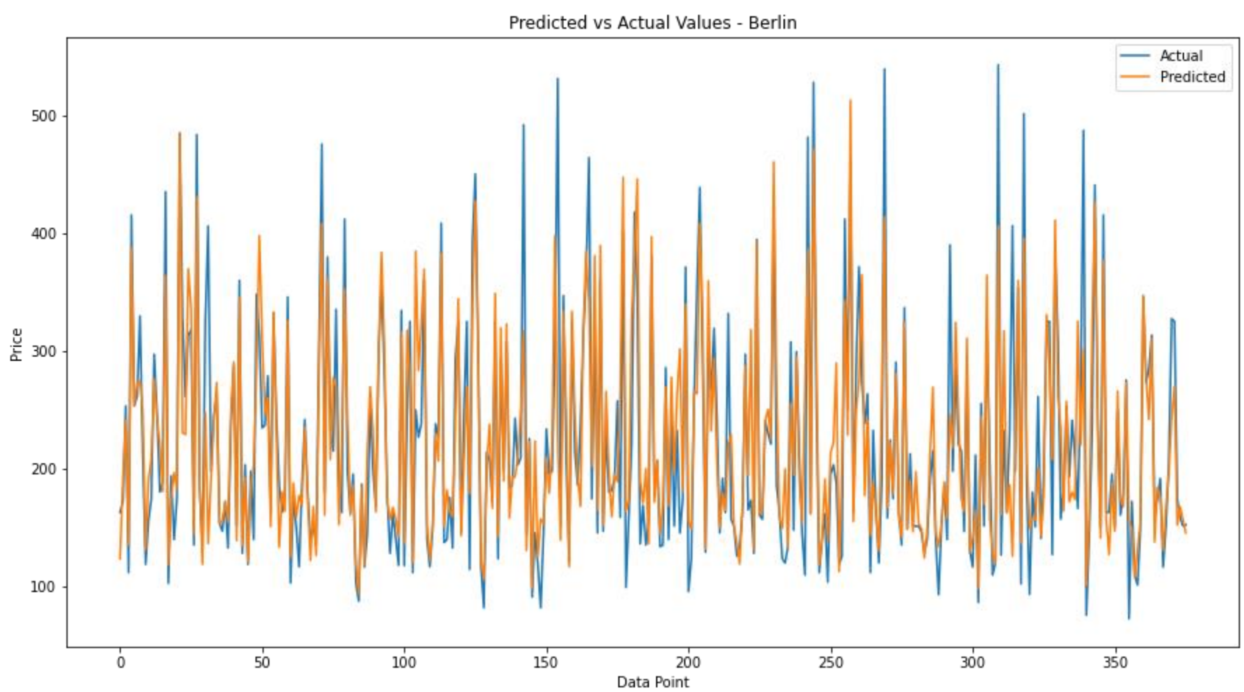

Figure 15: Line chart showing Prediction Vs Actual Values – Berlin

Lisbon

City	Model	R^2	
		Training Performance	***Testing Performance***
Lisbon	Random Forest	0.96	0.76
	Neural Network	0.56	0.53
	XGBoost	0.92	0.73

Table 13: Coefficient of Determination(R^2) - Lisbon

According to our dataset, both the Random Forest and **XGBoost** models exhibited closely aligned R^2 values when forecasting Airbnb prices within the Lisbon area. Specifically, the R^2 values for Random Forest and XGBoost were 0.76 and 0.73, respectively. These values indicate that both models are capable of elucidating roughly 76% and 73% of the variations in prices. This demonstration of predictive power underscores their effectiveness in furnishing accurate predictions for Airbnb prices in Lisbon. However, it's noteworthy that the Random Forest model yielded lower MAE and RMSE values. This outcome highlights Random Forest as the more fitting choice, as it showcases superior predictive performance with diminished errors compared to the alternative models.

The line chart below illustrates the variation between the predicted prices and the actual prices for Lisbon, based on the best performing model, which is the Random Forest model. As seen in the chart, the predicted prices align closely with the actual prices, confirming the model's accuracy in capturing the price variations for Airbnb listings in Lisbon. This graph further reinforces the superiority of the Random Forest model in making precise predictions for Airbnb prices in the Lisbon region, outperforming both the XGBoost and Neural Networks models.

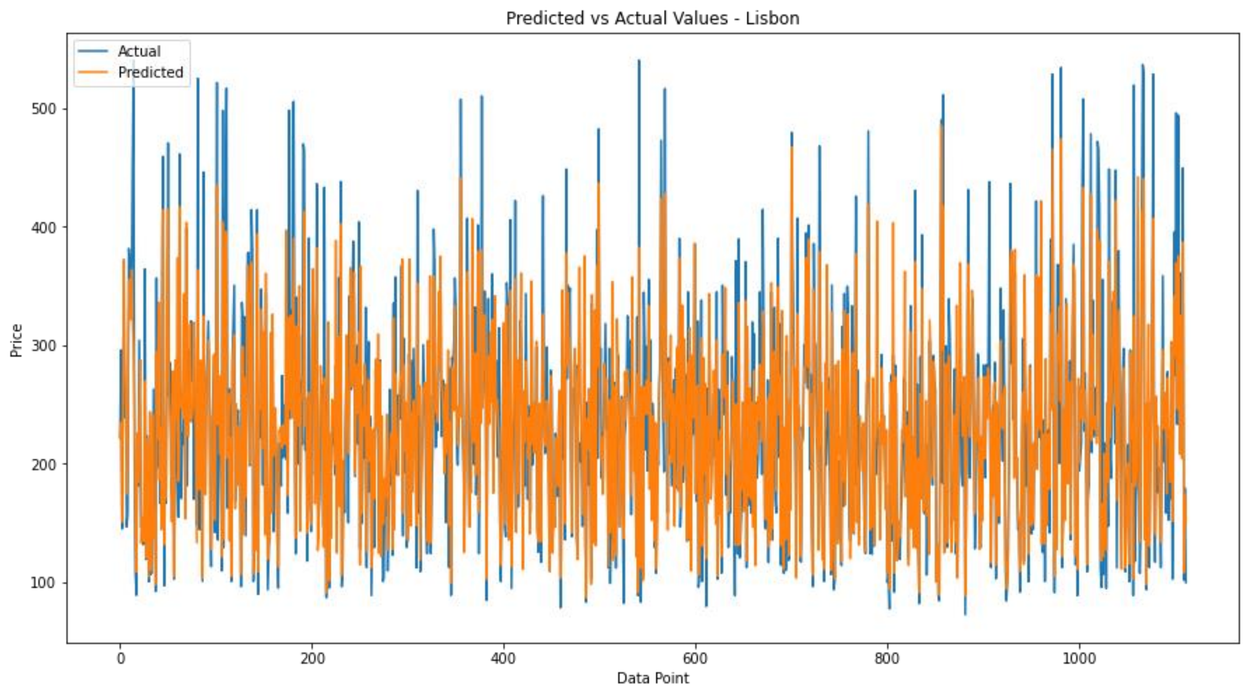

Figure 16: Line chart showing Prediction Vs Actual Values - Lisbon

Paris

City	Model	R^2	
		Training Performance	*Testing Performance*
Paris	Random Forest	0.95	0.60
	Neural Network	0.26	0.26
	XGBoost	0.86	0.55

Table 14: Coefficient of Determination(R^2) - Paris

The **Random Forest** machine learning model displayed superior performance in the context of Paris when compared to the other utilized models. It achieved a relatively higher R^2 value of 0.60, implying its capability to account for approximately 60% of the price variations within the dataset. While this R^2 value stands as reasonable, there exists an opportunity for enhancing the model's predictive precision even further. In contrast, the XGBoost model returned a lower R^2 value of 0.55, suggesting a slightly diminished performance compared to the Random Forest approach. Notably, the Neural Network model exhibited subpar performance with a low R^2 value of 0.26, indicating its challenges in effectively capturing and elucidating the fluctuations in prices for the Paris region. This observation aligns with the model's elevated errors, as evidenced by the MAE and RMSE values.

The line chart below provides a visual representation of the variations between the predicted and actual values from the Random Forest model in the city of Paris. The close alignment between the predicted and actual values on the chart suggests that the XGBoost model is proficient in making accurate price predictions in this specific city, further reinforcing its superior performance over the other models.

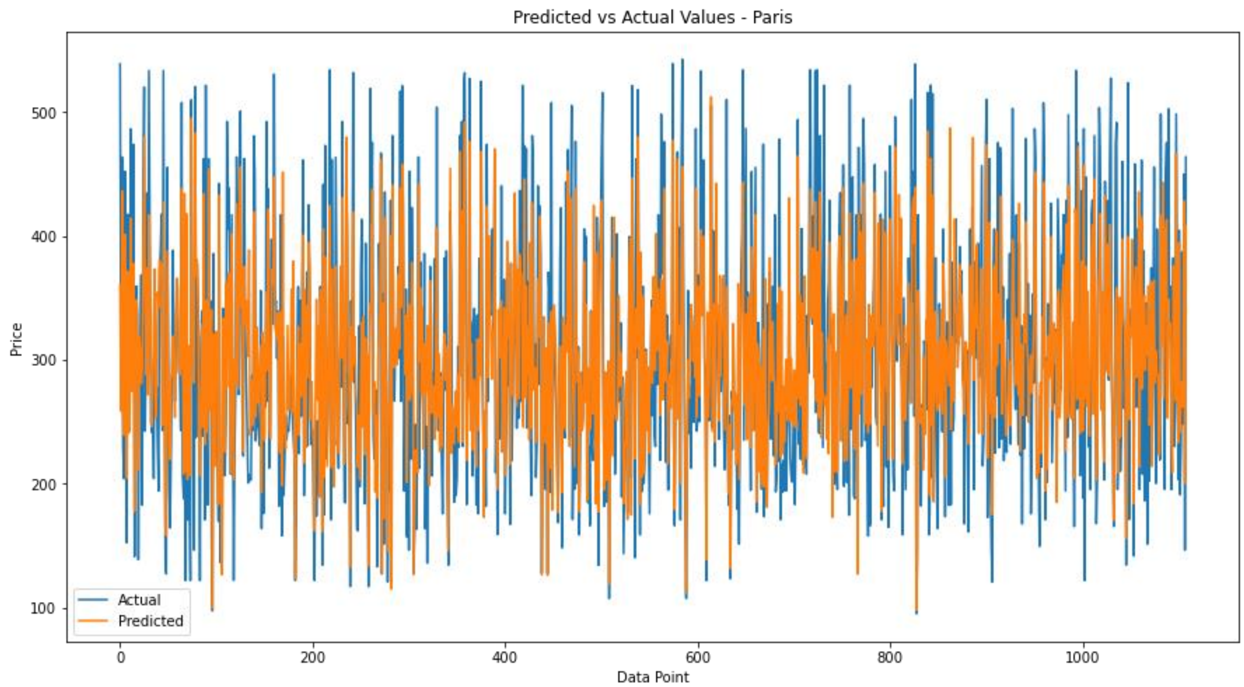

Figure 17: Line chart showing Prediction Vs Actual Values - Paris

Rome

City	Model	R^2	
		Training Performance	***Testing Performance***
Rome	Random Forest	0.96	0.71
	Neural Network	0.41	0.41
	XGBoost	0.87	0.67

Table 15: Coefficient of Determination(R^2) - Rome

In the city of Rome, within the scope of this study's examined predictive models, the **Random Forest** model emerged as the standout performer, boasting an R^2 value of 0.71. Following closely was the XGBoost model, which achieved an R^2 value of 0.67. Conversely, the Neural Network model delivered subpar results in Rome, evident in its comparatively modest R^2 value of 0.41. Similarly, the Random Forest model yielded the lowest MAE and RMSE values compared to the other models, reinforcing its superior accuracy in predicting Airbnb prices in Rome.

Much akin to the other cities, a line chart was employed to vividly illustrate the disparities between predicted and actual prices in Rome, leveraging the Random Forest model—revealing its prowess as the most adept predictor of Airbnb prices in this particular city.

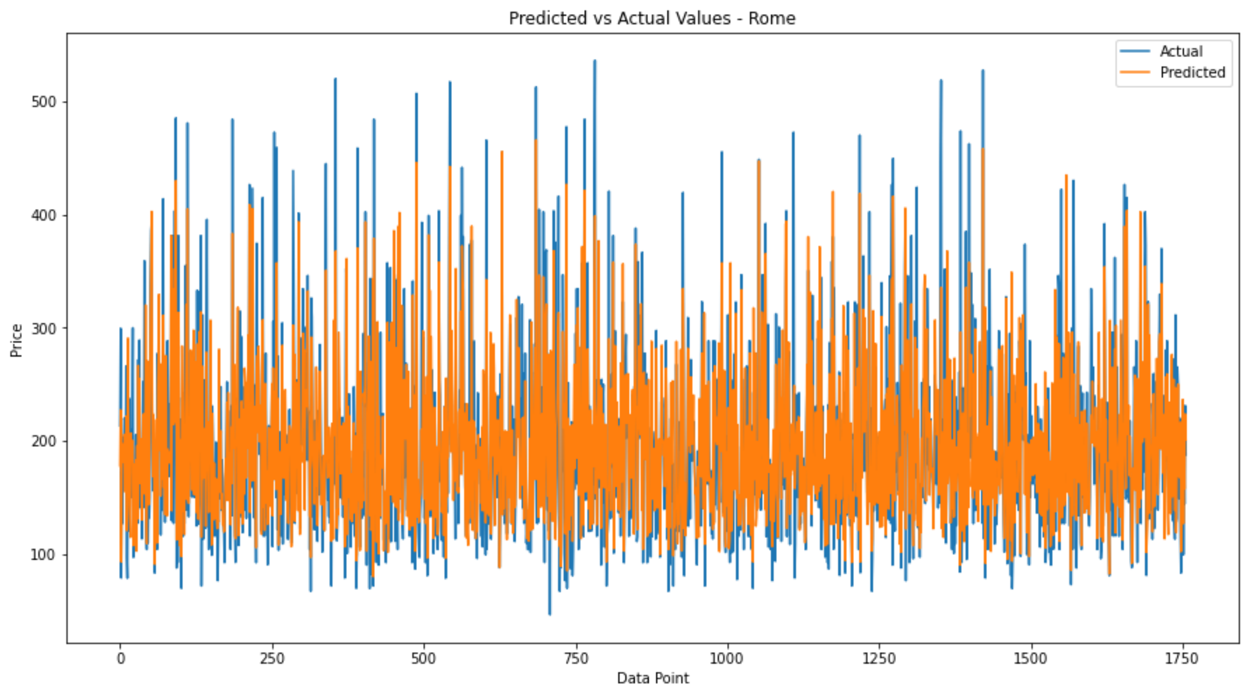

Figure 18: Line chart showing Prediction Vs Actual Values - Rome

Vienna

City	Model	R^2	
		Training Performance	***Testing Performance***
Vienna	Random Forest	0.95	0.65
	Neural Network	0.42	0.42
	XGBoost	0.97	0.63

Table 16: Coefficient of Determination(R^2) - Vienna

In Vienna, the models' testing performance was evaluated, revealing distinct trends. The **Random Forest** model exhibited a solid R^2 value of 0.65, indicating its ability to explain approximately 65% of price variations. The XGBoost model maintained its effectiveness with a testing R^2 value of 0.63, signifying its capability to elucidate about 63% of price fluctuations. In contrast, the Neural Network model lagged behind with a lower R^2 value of 0.42, suggesting its struggle to capture and explain the complexities of price dynamics. These observations were consistent with the MAE and RMSE values, where the Random Forest model again demonstrated the lowest errors, reinforcing its suitability for precise Airbnb price prediction in Vienna.

As with other cities, a line chart was employed to illustrate the disparity between predicted and actual prices for Vienna. The chart, centered around the Random Forest model, reinforced its status as the most precise predictor for Airbnb prices in the city.

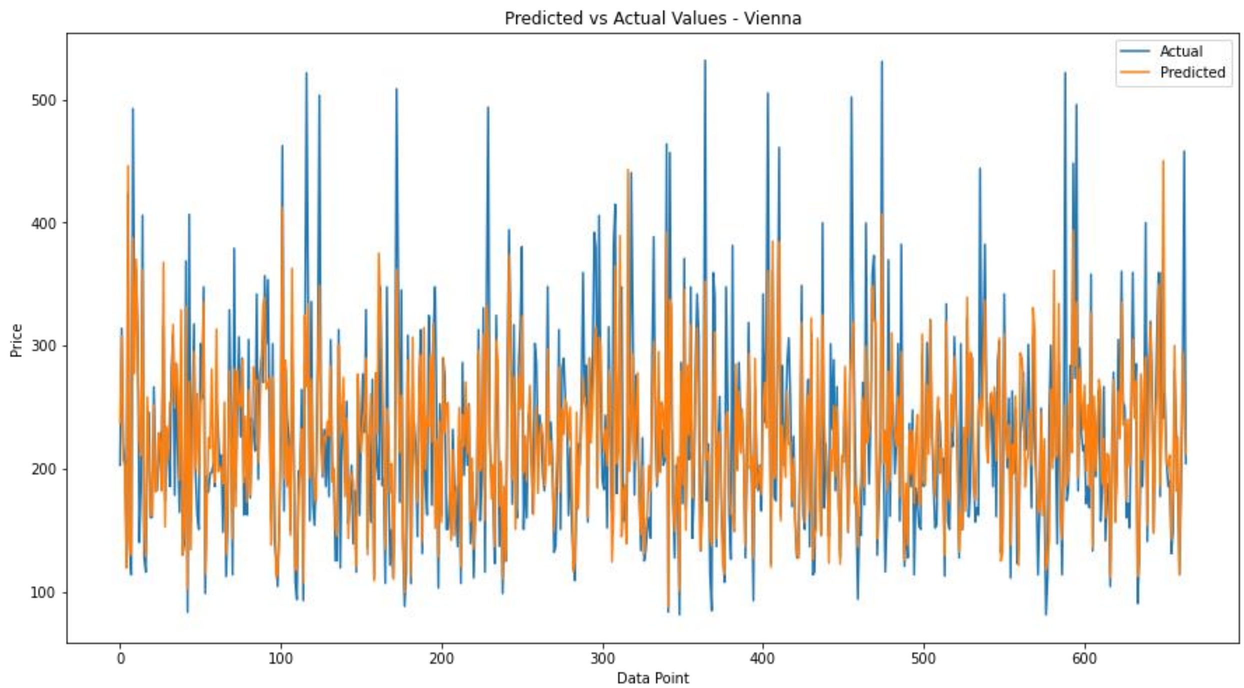

Figure 19: Line chart showing Prediction Vs Actual Values - Vienna

CHAPTER FIVE: SUMMARY OF FINDINGS

5. Summary of Findings

Throughout this analysis, we delved into the performance of three prominent machine learning models—Random Forest, Neural Networks, and XGBoost—across the selected eight European cities. By examining key performance metrics (MAE and RMSE) during both the training and testing phases, we were able to uncover nuanced insights into their predictive abilities, strengths, challenges, and generalization capabilities.

5.1. Predictive Accuracy Strength of Models

Random Forest Model:

The Random Forest model surfaced as a formidable contender in terms of its predictive performance, displaying a remarkable combination of robustness and consistency along with a distinct aptitude for excelling within specific the European Airbnb urban contexts.

- ***City-specific Excellence:*** A standout feature of the Random Forest model was its ability to shine across multiple cities, including Athens, Barcelona, Lisbon, Paris, Rome, and Vienna. In these urban contexts, the model showcased the lowest MAE and RMSE values during testing. Take, for instance, Athens, where it achieved a remarkable MAE of 25.87 and RMSE of 41.44—a testament to its prowess in capturing intricate Airbnb pricing intricacies.
- ***Consistency and Dependability:*** The Random Forest model's standout feature lay in its consistent and dependable performance. The model exhibited negligible fluctuations in error metrics as it transitioned from the training performance to the testing phase. This remarkable steadiness underscored its ability to generalize predictions effectively, even in scenarios involving data patterns it hadn't encountered before especially for the very specific industries like Airbnb where prices are generally given by the host of the apartment, hence, new data point are always generated. Such a high level of consistency reinforces the model's reliability in real-world applications, making it a valuable tool for a wide range of forecasting tasks.

Neural Networks Model:

The Neural Networks model introduced intriguing insights, showcasing its potential and its limitations:

- ***Varied Performance Across Cities:*** While the Neural Networks model exhibited strength in specific instances, such as Rome, Barcelona and Berlin where its ability to effectively perform good to unseen data evidenced by its reduced MAE and RMSE in the testing phase was experienced, its performance was varied across different cities in comparison to the other two models that were adopted in this study. This variance in performance across cities underscores the intricate nature of Airbnb pricing dynamics in Europe. While the Neural Networks model could capture specific nuances in pricing patterns, its inability to consistently yield the lowest errors in any single city suggests that the underlying complexities of the data can vary significantly between different urban contexts. As such, the model's strengths and limitations are contingent upon the specific characteristics of each city's Airbnb price dataset.

- ***Localized Mastery and Sensitivity***: It is worth highlighting the Neural Networks model's adeptness in capturing intricate pricing patterns that are specific to particular cities, such as Athens. In these instances, the model demonstrated a heightened level of proficiency, effectively capturing the subtle fluctuations that define the local Airbnb price dynamics. This localized mastery showcases the model's potential for precision, particularly when it comes to understanding the intricate factors that influence pricing in distinct urban contexts. However, the model's performance variability became evident when assessed across the various cities included in the study. Despite its prowess in certain cities, the model's ability to consistently achieve the lowest errors proved challenging. This performance inconsistency suggests that the model is sensitive to the intricate interplay of factors that are unique to each individual urban environment. In essence, the Neural Networks model's effectiveness is contingent upon the specific characteristics of the city it is applied to, and its sensitiveness to the diverse dynamics that underpin each urban setting. This nuanced interplay between the model's localized strengths and its sensitivity to different urban contexts underscores the complexity of predicting Airbnb prices across diverse European cities.

XGBoost Model:

The XGBoost model added another layer of insight to our findings.

- ***Strengths in Specific European Cities***: Similar to the Neural Networks model, the XGBoost model displayed strengths in particular cities like Berlin and Amsterdam, where it demonstrated the lowest MAE and RMSE values even lower than the Random Forest model. These instances showcased its ability to effectively discern and capture distinct pricing dynamics unique to those urban landscapes***.***
- ***Consistency Challenges:*** However, like the Neural Networks model, the XGBoost model encountered challenges in maintaining uniform accuracy across the range of cities studied. This challenge underscored the inherent complexity of predicting Airbnb prices across diverse European cities.

Comparative Insights

By drawing parallels between the models, several distinct comparative insights emerged:

- The Random Forest model's strength lay in its consistent excellence, prominently showcased in cities like Athens, Barcelona, Lisbon, Paris, Rome, and Vienna, where it consistently achieved the lowest MAE and RMSE values during testing.
- The Neural Networks model provided localized insights and captured nuances but faced challenges in consistently producing the lowest errors for predicting Airbnb prices across diverse European urban region.
- The XGBoost model demonstrated strengths in deciphering unique Region pricing dynamics, yet its uniform performance faced hurdles similar to the Neural Networks model.

In conclusion, the choice of model hinges on the specific priorities of the prediction task and the intricate dynamics intrinsic to each city's Airbnb pricing data. While the Neural Networks and XGBoost models offered distinctive perspectives, the Random Forest model's unwavering excellence in cities like Athens, Barcelona, Lisbon, Paris, Rome, and Vienna positions it as a reliable option. Its ability to consistently yield the lowest MAE and RMSE values across diverse urban contexts underscores its efficacy in delivering precise Airbnb price predictions—a testament to its practicality and reliability in various city scenarios.

5.2. Effectiveness of Predictors for Airbnb Prices in European Cities

This study delves into the intricate relationships between various predictors and Airbnb prices across the spectrum of European cities. The correlation matrix in Figure 10 unveils the dynamics that shape pricing in this context. This section provides a comprehensive elaboration of the derived insights:

The impact of room types on pricing is evident, with distinct categories revealing a discernible influence. This nuanced categorization leads to specific room types being closely associated with lower prices, contributing to a diverse pricing spectrum that caters to varying traveler preferences. Moreover, the influence of person capacity on pricing dynamics becomes apparent, unveiling a compelling connection between the size of accommodations and their associated prices. Larger lodgings, capable of accommodating more individuals, exert an upward pull-on prices, tapping into the demand for expansive stays and enhancing the perception of value. In the realm of host attributes, the influence of superhost status on pricing is revealed to be subtly impactful. While superhosting remains a valuable factor in influencing guest preferences, its effect on pricing is measured, indicating a delicate balance between perceived value and cost considerations. This study also delves into the pivotal role of qualitative aspects in shaping pricing patterns. Higher cleanliness ratings and guest satisfaction scores are found to correlate positively with prices, underscoring the intrinsic value attributed to exceptional guest experiences. The spatial dimension further amplifies these insights, as accommodations offering multiple bedrooms demonstrate a direct connection with higher pricing. This correlation reinforces the notion that spatial considerations are central in influencing pricing decisions, aligning with intuitive expectations. Interestingly, the analysis of proximity to the city center and metro connections reveals a modest yet discernible impact on pricing. This spatial factor, while not the sole determinant, contributes to the delicate interplay of location and pricing dynamics. In the realm of strategic pricing, listings with multiple rooms are observed to subtly adjust their pricing strategy. By offering slightly lower prices, these listings entice travelers seeking expansive accommodations, striking a balance between supply and demand dynamics. Lastly, the temporal dimension, captured through the day of booking, presents a minimal influence on pricing. This finding underscores the stability of pricing decisions, irrespective of the day of booking, highlighting a balanced approach to temporal considerations.

In conclusion, this comprehensive analysis orchestrates a symphony of insights, illuminating the intricate relationships that dictate Airbnb pricing across diverse European cities. The empirical nuances uncovered hold profound implications for both hosts and travelers, contributing to a panoramic understanding of the intricate fabric that weaves the cost of stay across a dynamic urban landscape.

CHAPTER SIX: CONCLUSIONS

6. Conclusion

In our exploration of predicting Airbnb prices in the diverse European cityscape, this study has delved into data analysis, model creation, and insightful revelations. This section serves to reflect on our discoveries, acknowledge limitations, and consider future steps. Here, we provide a comprehensive overview of the insights we've gathered, while stating the challenges we've encountered, and outline potential avenues for further research.

6.1. Recommendations

The practical implications and managerial recommendations stemming from the analysis extend beyond the mere development of predictive models for Airbnb pricing in European cities. The insights unveiled and findings derived from this study offer a multi-faceted perspective that has profound implications for both hosts, travelers and Airbnb owners navigating the intricate landscape of Airbnb pricing dynamics. The culmination of this comprehensive comparative analysis underscores the importance of refining predictive accuracy, fostering better understanding, and providing actionable insights within the realm of Airbnb pricing. The predictive models evaluated in this study provide a lens through which the complex interplay of various factors shaping pricing can be deciphered. The insights derived not only contribute to the optimization of pricing strategies but also offer a window into the broader tourism and hospitality landscape of these European cities.

As cities vary in their cultural, economic, and regulatory landscapes, the findings accentuate the necessity of tailored approaches. Leveraging machine learning for predictive pricing especially in cities like Athens, Berlin and Lisbon where we can see a good performance accuracy of the machine learning models, this will afford the opportunity to capture localized trends, behavior patterns, and market dynamics that are often obscured by conventional methods. By harnessing the predictive power of machine learning, hosts can calibrate their pricing strategies in alignment with the nuances of each city, maximizing revenue potential, it may be more essential to also include several other factors that may affect pricing has highlighted in the study earlier.

For travelers, the insights garnered from this study offer a transparent glimpse into the forces that shape Airbnb prices in European cities. The ability to anticipate pricing trends based on an understanding of variables such as room type, location, and seasonal fluctuations will empower travelers to make more informed decisions. Armed with this knowledge, travelers can navigate the booking process with confidence, optimizing their budget due to the transparency while ensuring a satisfying stay.

Furthermore, the implications of this analysis extend beyond individual cities. The broader implications extend to the evolution of the entire hospitality industry and the tourism industry at large, where data-driven insights are transforming the way pricing is approached. The adoption of machine learning techniques for predictive pricing models paves the way for a more responsive and agile pricing strategy that adapts to real-time demand and supply fluctuations.

In conclusion, the comparative analysis of predictive models for Airbnb pricing in European cities not only enhances predictive accuracy but also contributes to a deeper understanding of the multifaceted dynamics at play. This analysis equips hosts and travelers alike with valuable insights to navigate the ever-evolving landscape of Airbnb pricing. As the hospitality industry continues to evolve, the integration of machine learning into pricing strategies stands as a testament to innovation and adaptability.

6.2. Limitations and Scope for future Research

Undoubtedly, a project of this magnitude comes with its inherent set of limitations that warrant attention. One primary limitation encountered resides in the dataset employed for this study. While the dataset provides a robust foundation, its inherent capture of dates is worth considering. The absence of futuristic data points restricts the study's ability to perform predictive analyses beyond the dataset's temporal scope. Moreover, an exploration of the broader contextual factors shaping pricing dynamics remains an untapped dimension. This limitation becomes particularly evident in the absence of variables capturing local regulations, cultural preferences, and economic conditions. Incorporating these variables could offer a more comprehensive understanding of the intricate interplay that molds pricing trends particularly across diverse European cities because of each European cities' distinct features.

In addition, while this study skillfully deploys individual machine learning models to discern patterns, an unexplored horizon lies in the realm of model synergy. An avenue for future research could involve investigating the potential benefits of amalgamating different machine learning models. By combining the strengths of various models, the predictive accuracy could potentially be heightened, unlocking a deeper level of insights for both hosts and guests alike.

However, it is important to acknowledge that this research, like any endeavor, is not devoid of limitations. The meticulous exploration of the dataset, predictive models, and factors shaping pricing dynamics provides an invaluable foundation. Nonetheless, the limitations present underscore the evolving nature of research and beckon for continuous refinement and expansion in pursuit of comprehending the complex tapestry of Airbnb pricing in European cities.

6.3. Future Steps

Considering the insights garnered and limitations encountered throughout this research, a multitude of promising steps can be taken to improve our model and future exploration as well. These steps hold the potential to not only bolster our comprehension of Airbnb pricing dynamics in European cities but also pave the way for enhanced predictive models.

Addressing the limitation of futuristic predictions could be achieved through the implementation of advanced time series forecasting techniques. By integrating these methods, the project could potentially harness the predictive power of trends over time which would involve leveraging historical data patterns to anticipate future pricing trends, offering a more robust predictive capability.

Additionally, we should consider including more factors that affect prices, such as local rules, what people prefer culturally, and the economic situation. These aspects all interact in complex ways and factoring them in could give us a better overall view of how Airbnb pricing works in Europe, especially considering the diverse cultures in different cities. Exploring the correlation between user behavior and pricing trends could shed light on the decision-making process of both hosts and guests. Better understanding why prices, change the way they do, we could explore how people behave when they decide on prices. By looking at things like how often rooms get booked, and what features guests prefer, we can shed light on the decision-making process of both hosts and guests. This would unveil the intricate interplay between external forces and pricing dynamics, providing a holistic perspective on Airbnb pricing.

Segmenting our dataset based on specific criteria, like the size of a city, tourism patterns, and economic trends, could give us more detailed insights. This could help us uncover pricing patterns that are unique to different groups, providing a clearer picture of how prices change and would unveil the pricing nuances in different markets.

As a final improvement that could be adopted in future studies, the development of predictive models for guest satisfaction, centered on pricing and accommodation attributes, holds the potential to enrich our comprehension of the entire user experience. This endeavor would enable us to delve beyond mere numerical analysis and delve into the tangible quality of a stay, effectively closing the gap between data-driven insights and the tangible guest journey. By doing so, the predictability of the Airbnb pricing model could be significantly enhanced, leading to more accurate and refined predictions. Furthermore, the implementation of dynamic pricing models that respond to real-time shifts in demand and supply emerges as a strategic step. This is particularly pertinent considering the variations in pricing that occur across different seasons and the evolving landscape of remote work opportunities. As companies increasingly adopt the "Work from Anywhere" paradigm, the demand for short-term rentals could see fluctuations tied to new working trends. Recognizing these dynamics becomes crucial not only in the context of contemporary business practices but also in the accurate prediction of Airbnb prices. Implementing such dynamic pricing models presents a forward-looking strategy for optimizing pricing strategies, acknowledging the dynamic nature of market forces and ensuring that the pricing remains aligned with the evolving demand patterns.

As research in the field of predictive Airbnb pricing models continues to evolve, these future steps present avenues for enhancing accuracy, broadening understanding, and providing actionable insights for both hosts and travelers navigating the intricate landscape of Airbnb pricing in European cities.

-- END --

References:

Adebiyi, A.A., Adewumi, A.O. and Ayo, C.K. (2014). *Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction*. [online] Journal of Applied Mathematics. Available at: https://www.hindawi.com/journals/jam/2014/614342/.

Airbnb (2023a). *About Us* . [online] Airbnb Newsroom. Available at: https://news.airbnb.com/about-us/.

Airbnb (2023b). *Airbnb Hosts share their vision for future EU STR rules*. [online] Airbnb Newsroom. Available at: https://news.airbnb.com/airbnb-hosts-share-their-vision-for-future-eu-str-rules/ [Accessed 17 Jul. 2023].

Airbnb (2023c). *More EU Hosts rely on Airbnb income, popularity of rooms grows*. [online] Airbnb Newsroom. Available at: https://news.airbnb.com/more-eu-hosts-rely-on-airbnb-income-popularity-of-rooms-grows/ [Accessed 17 Jul. 2023].

AllTheRooms. (2022). *Average Airbnb Prices by City [2022]*. [online] Available at: https://www.alltherooms.com/resources/articles/average-airbnb-prices-by-city/.

Alqahtani, M., Gupta, D., Kwon, B. H., & Chung, T. (2021). Predicting Airbnb Rental Prices Using Machine Learning Techniques. IEEE Access, 9, 2813-2821.

Barnes, S.J. and Kirshner, S.N. (2021). Understanding the impact of host facial characteristics on Airbnb pricing: Integrating facial image analytics into tourism research. *Tourism Management*, 83, p.104235. doi:https://doi.org/10.1016/j.tourman.2020.104235.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), pp.197–227. doi:https://doi.org/10.1007/s11749-016-0481-7.

Blal, I., Singal, M. and Templin, J. (2018). Airbnb's effect on hotel sales growth. *International Journal of Hospitality Management*, 73, pp.85–92. doi:https://doi.org/10.1016/j.ijhm.2018.02.006.

Bowers, B. (2017). *Bucknell Digital Commons Global Manager Abroad Student Scholarship Airbnb in Paris AIRBNB EXPANSION EFFORTS IN PARIS, FRANCE*. [online] Available at: https://digitalcommons.bucknell.edu/cgi/viewcontent.cgi?article=1004&context=glbm400.

Bull, A.O. (1994), "Pricing a Motel's Location", International Journal of Contemporary Hospitality Management, Vol. 6 No. 6, pp. 10-15. https://doi.org/10.1108/09596119410070422

Castro, C. and Ferreira, F.A. (2018). Online hotel ratings and its influence on hotel room rates: the case of Lisbon, Portugal. *Tourism & Management Studies*, 14(SI1), pp.63–72. doi:https://doi.org/10.18089/tms.2018.14si107.

Chattopadhyay, M. and Mitra, S.K. (2019). Do airbnb host listing attributes influence room pricing homogenously? *International Journal of Hospitality Management*, 81, pp.54–64. doi:https://doi.org/10.1016/j.ijhm.2019.03.008.

Dhillon, J., Eluri, N.P., Kaur, D., Chhipa, A., Gadupudi, A., Eravi, R.C. and Pirouz, M. (2021). *Analysis of Airbnb Prices using Machine Learning Techniques*. [online] IEEE Xplore. doi:https://doi.org/10.1109/CCWC51732.2021.9376144.

Fitrianingsih, Rahayu, D.A. and Zazila, F.R. (2022). *Dynamic Pricing Analytic of Airbnb Amsterdam Using K-Means Clustering*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICIC56845.2022.10006966.

Guttentag, D. (2019). The Airbnb effect: Disruptive innovation and the rise of an informal tourism accommodation sector. Journal of Tourism Futures, 5(2), 118-130.

Hartmans, A. (2017). *Airbnb now has more listings worldwide than the top five hotel brands combined - Business Insider*. [online] Business Insider. Available at: https://www.businessinsider.com/airbnb-total-worldwide-listings-2017-8?r=US&IR=T.

Hodson, T.O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 15(14), pp.5481–5487. doi:https://doi.org/10.5194/gmd-15-5481-2022.

Kanakaris, N. and Karacapilidis, N. (2023). Predicting prices of Airbnb listings via Graph Neural Networks and Document Embeddings: The case of the island of Santorini. *Procedia Computer Science*, [online] 219, pp.705–712. doi:https://doi.org/10.1016/j.procs.2023.01.342.

Kushlev, K., & Dunn, E. W. (2015). Checking email less frequently reduces stress. Computers in Human Behavior, 43, 220-228.

Kwan, M. P., Lee, J., & Lee, J. H. (2018). Measuring spatial dependence and heterogeneity of Airbnb nightly rental prices in the United States. Applied Geography, 97, 93-104.

Lampinen, A. and Cheshire, C. (2016). Hosting via Airbnb. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. doi:https://doi.org/10.1145/2858036.2858092.

Li, Y., Pan, Q., Yang, T. and Guo, L. (2016). Reasonable price recommendation on Airbnb using Multi-Scale clustering. *2016 35th Chinese Control Conference (CCC)*. [online] doi:https://doi.org/10.1109/CHICC.2016.7554467.

Liu, Y. (2021). Airbnb Pricing Based on Statistical Machine Learning Models. *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*. doi:https://doi.org/10.1109/conf-spml54095.2021.00042.

Lonely Planet. (n.d.). *Travelling in Europe? Here is how much an average Airbnb costs in each country*. [online] Available at: https://www.lonelyplanet.com/news/average-airbnb-costs-europe [Accessed 20 May 2023].

Luo, Y., Zhou, X. and Zhou, Y. (2019). *Predicting Airbnb Listing Price Across Different Cities*. [online] Available at: https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647491.pdf.

Mahyoub, M., Ataby, A.A., Upadhyay, Y. and Mustafina, J. (2023). *AIRBNB Price Prediction Using Machine Learning*. [online] IEEE Xplore. doi:https://doi.org/10.1109/DeSE58274.2023.10099909.

McNeil, B. (2020). Price Prediction in the Sharing Economy: A Case Study with Airbnb data. *Honors Theses and Capstones*. [online] Available at: https://scholars.unh.edu/honors/504/.

Nafisa Teshaboyeva (2023). THE IMPORTANCE OF TOURISM IN PRESENT DAY. Журнал иностранных языков и лингвистики, [online] 5(5). Available at: https://fll.jdpu.uz/index.php/fll/article/view/7811

Peng, N., Li, K. and Qin, Y. (2020). *Leveraging Multi-Modality Data to Airbnb Price Prediction*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICEMME51517.2020.00215.

Rezazadeh Kalehbasti, P., Nikolenko, L. and Rezaei, H. (2021). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. *Lecture Notes in Computer Science*, pp.173–184. doi:https://doi.org/10.1007/978-3-030-84060-0_11.

Schwarzová, L. (2020). *Predicting Airbnb Prices with Neighborhood Characteristics: Machine Learning Approach*. [online] Available at: http://arno.uvt.nl/show.cgi?fid=156281.

Soler, I.P. and Gemar, G. (2018). Hedonic price models with geographically weighted regression: An application to hospitality. *Journal of Destination Marketing & Management*, 9, pp.126–137. doi:https://doi.org/10.1016/j.jdmm.2017.12.001.

Thessen, A.E. (2016). *Adoption of machine learning techniques in Ecology and Earth Science*. [online] peerj.com. Available at: https://peerj.com/preprints/1720/ [Accessed 10 Jul. 2023].

Toader, V., Negrușa, A.L., Bode, O.R. and Rus, R.V. (2021). Analysis of price determinants in the case of Airbnb listings. *Economic Research-Ekonomska Istraživanja*, pp.1–17. doi:https://doi.org/10.1080/1331677x.2021.1962380.

Truong, Q., Nguyen, M., Dang, H. and Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, 174, pp.433–442. doi:https://doi.org/10.1016/j.procs.2020.06.111.

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174, 433-442. ISSN 1877-0509

Wang, D. and Nicolau, J.L. (2017). Price determinants of sharing economy-based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management*, 62, pp.120–131. doi:https://doi.org/10.1016/j.ijhm.2016.12.007.

Willmott, C. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, [online] 30(1), pp.79–82. doi:https://doi.org/10.3354/cr030079.

Woodward, M. (2022). *Airbnb Statistics [2023]: User & Market Growth Data*. [online] www.searchlogistics.com. Available at: https://www.searchlogistics.com/learn/statistics/airbnb-statistics/.

Yang, S. (2021). Learning-based Airbnb Price Prediction Model. *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*. doi:https://doi.org/10.1109/ecit52743.2021.00068.

Zhu, A., Li, R. and Xie, Z. (2020). *Machine Learning Prediction of New York Airbnb Prices*. [online] IEEE Xplore. doi:https://doi.org/10.1109/AI4I49448.2020.00007.

Appendix

Appendix 1: Hyperparameters Adjustments on Each Cities Best Model.

Athens		**Testing Performance**	
Model	Hyperparameters	MAE	RMSE
Random Forest	Default	27.56	43.23
	Maximum Tree Depth - 6	30.23	52.36
	Stopping Rules Minimum- 0.7	26.61	42.75
	Maximum Tree Depth - 8	25.87	41.44
	Maximum Tree Depth - 9	29.62	45.93
Amsterdam		**Testing Performance**	
Model	Hyperparameters	MAE	RMSE
XGBoost	Default	40.52	72.34
	Maximum Depth-7	43.02	79.62
	reg_lambda- 1.4	39.21	53.69
	reg_alpha- 1.2	39.87	54.61
	Trees estimator- 600	38.94	49.7
Barcelona		**Testing Performance**	
Model	Hyperparameters	MAE	RMSE
Random Forest	Default	40.12	59.97
	Maximum Tree Depth - 7	37.77	54.67
	Stopping Rules Minimum- 0.8	38.24	55.49
	Maximum Tree Depth - 8	39.64	58.03
	Maximum Tree Depth - 9	39.51	57.54
Berlin		**Testing Performance**	
Model	Hyperparameters	MAE	RMSE
XGBoost	Default	35.61	53.74
	Maximum Depth-7	32.93	51.68
	reg_lambda- 1.4	33.75	52.29
	reg_alpha- 1.2	36.12	54.94
	Trees estimator- 700	33.93	52.31
Lisbon		**Testing Performance**	
Model	Hyperparameters	MAE	RMSE
Random Forest	Default	31.75	45.64
	Maximum Tree Depth - 6	34.02	46.39
	Stopping Rules Minimum- 0.6	33.71	47.73
	Maximum Tree Depth - 7	35.24	49.52
	Maximum Tree Depth - 8	34.36	48.21

Paris		**Testing Performance**	
Model	Hyperparameters	MAE	RMSE
Random Forest	Default	46.29	65.92
	Maximum Tree Depth - 7	48.34	72.41
	Stopping Rules Minimum- 0.8	45.29	63.45
	Maximum Tree Depth - 8	44.62	60.87
	Maximum Tree Depth - 9	44.98	61.13
Rome		**Testing Performance**	
Model	Hyperparameters	MAE	RMSE
Random Forest	Default	36.41	51.23
	Maximum Tree Depth - 6	35.87	48.51
	Stopping Rules Minimum- 0.8	31.25	44.39
	Maximum Tree Depth - 7	33.78	47.92
	Maximum Tree Depth - 8	30.03	43.83
Vienna		**Testing Performance**	
Model	Hyperparameters	MAE	RMSE
Random Forest	Default	34.06	52.34
	Maximum Tree Depth - 6	41.51	55.87
	Stopping Rules Minimum- 0.7	35.61	53.67
	Maximum Tree Depth - 8	33.55	51.04
	Maximum Tree Depth - 9	34.97	52.99

Appendix 2: Best Performing Python Model Output

Amsterdam

In [36]:

```
def evaluate_xgboost(x_train, y_train, x_test, y_test):
    # Create the XGBoost regression model
    xgb_model = xgb.XGBRegressor()
    xgb_model.fit(x_train, y_train)

    # Predict the target variable for both train and test datasets
    y_train_pred = xgb_model.predict(x_train)
    y_test_pred = xgb_model.predict(x_test)

    # Calculate accuracy measures for train and test datasets
    train_mae = mean_absolute_error(y_train, y_train_pred)
    test_mae = mean_absolute_error(y_test, y_test_pred)

    train_rmse = np.sqrt(mean_squared_error(y_train, y_train_pred))
    test_rmse = np.sqrt(mean_squared_error(y_test, y_test_pred))

    train_mse = mean_squared_error(y_train, y_train_pred)
    test_mse = mean_squared_error(y_test, y_test_pred)

    train_mape = np.mean(np.abs((y_train - y_train_pred) / y_train)) * 100
    test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100

    # Calculate R-squared for train and test datasets
    train_r2 = r2_score(y_train, y_train_pred)
    test_r2 = r2_score(y_test, y_test_pred)

    # Create a dataframe to store the accuracy measures
    accuracy_df = pd.DataFrame({
        'MAE': [train_mae, test_mae],
        'RMSE': [train_rmse, test_rmse],
        'MSE': [train_mse, test_mse],
        'MAPE': [train_mape, test_mape],
        'R-squared': [train_r2, test_r2]
    }, index=['Train', 'Test'])

    return accuracy_df

# Assuming you have your x_train, y_train, x_test, and y_test datasets
accuracy_df = evaluate_xgboost(x_train, y_train, x_test, y_test)
print(accuracy_df)
```

```
             MAE       RMSE          MSE       MAPE  R-squared
Train   7.183738  10.105918   102.129578   2.123451   0.988259
Test   38.043856  49.695807  2469.673238  11.093021   0.692610
```

Athens

In [13]:
```python
def evaluate_xgboost(x_train, y_train, x_test, y_test):
    # Create the XGBoost regression model
    xgb_model = xgb.XGBRegressor()
    xgb_model.fit(x_train, y_train)

    # Predict the target variable for both train and test datasets
    y_train_pred = xgb_model.predict(x_train)
    y_test_pred = xgb_model.predict(x_test)

    # Calculate accuracy measures for train and test datasets
    train_mae = mean_absolute_error(y_train, y_train_pred)
    test_mae = mean_absolute_error(y_test, y_test_pred)

    train_rmse = np.sqrt(mean_squared_error(y_train, y_train_pred))
    test_rmse = np.sqrt(mean_squared_error(y_test, y_test_pred))

    train_mse = mean_squared_error(y_train, y_train_pred)
    test_mse = mean_squared_error(y_test, y_test_pred)

    train_mape = np.mean(np.abs((y_train - y_train_pred) / y_train)) * 100
    test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100

    # Calculate R-squared for train and test datasets
    train_r2 = r2_score(y_train, y_train_pred)
    test_r2 = r2_score(y_test, y_test_pred)

    # Create a dataframe to store the accuracy measures
    accuracy_df = pd.DataFrame({
        'MAE': [train_mae, test_mae],
        'RMSE': [train_rmse, test_rmse],
        'MSE': [train_mse, test_mse],
        'MAPE': [train_mape, test_mape],
        'R-squared': [train_r2, test_r2]
    }, index=['Train', 'Test'])

    return accuracy_df

# Assuming you have your x_train, y_train, x_test, and y_test datasets
accuracy_df = evaluate_xgboost(x_train, y_train, x_test, y_test)
print(accuracy_df)
```

```
             MAE       RMSE          MSE       MAPE  R-squared
Train  14.776323  20.750028   430.563655  11.173429   0.907530
Test   27.965229  43.115518  1858.947863  19.717968   0.626325
```

Barcelona

In [20]:
```
def evaluate_random_forest(x_train, y_train, x_test, y_test):
    # Create and fit the Random Forest regression model
    random_forest_model = RandomForestRegressor()
    random_forest_model.fit(x_train, y_train)

    # Predict the target variable for both train and test datasets
    y_train_pred = random_forest_model.predict(x_train)
    y_test_pred = random_forest_model.predict(x_test)

    # Calculate accuracy measures for train and test datasets
    train_mae = mean_absolute_error(y_train, y_train_pred)
    test_mae = mean_absolute_error(y_test, y_test_pred)

    train_rmse = np.sqrt(mean_squared_error(y_train, y_train_pred))
    test_rmse = np.sqrt(mean_squared_error(y_test, y_test_pred))

    train_mse = mean_squared_error(y_train, y_train_pred)
    test_mse = mean_squared_error(y_test, y_test_pred)

    train_mape = np.mean(np.abs((y_train - y_train_pred) / y_train)) * 100
    test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100

    # Calculate R-squared for train and test datasets
    train_r2 = r2_score(y_train, y_train_pred)
    test_r2 = r2_score(y_test, y_test_pred)

    # Create a dataframe to store the accuracy measures
    accuracy_df = pd.DataFrame({
        'MAE': [train_mae, test_mae],
        'RMSE': [train_rmse, test_rmse],
        'MSE': [train_mse, test_mse],
        'MAPE': [train_mape, test_mape],
        'R-squared': [train_r2, test_r2]
    }, index=['Train', 'Test'])

    return accuracy_df

# Assuming you have your x_train, y_train, x_test, and y_test datasets
accuracy_df = evaluate_random_forest(x_train, y_train, x_test, y_test)
print(accuracy_df)
```

```
             MAE       RMSE          MSE       MAPE  R-squared
Train  14.928219  21.625487   467.661686   6.969832   0.958976
Test   37.772259  54.671248  2988.945357  18.385699   0.731394
```

Berlin

```
In [27]: def evaluate_random_forest(x_train, y_train, x_test, y_test):
    # Create and fit the Random Forest regression model
    random_forest_model = RandomForestRegressor()
    random_forest_model.fit(x_train, y_train)

    # Predict the target variable for both train and test datasets
    y_train_pred = random_forest_model.predict(x_train)
    y_test_pred = random_forest_model.predict(x_test)

    # Calculate accuracy measures for train and test datasets
    train_mae = mean_absolute_error(y_train, y_train_pred)
    test_mae = mean_absolute_error(y_test, y_test_pred)

    train_rmse = np.sqrt(mean_squared_error(y_train, y_train_pred))
    test_rmse = np.sqrt(mean_squared_error(y_test, y_test_pred))

    train_mse = mean_squared_error(y_train, y_train_pred)
    test_mse = mean_squared_error(y_test, y_test_pred)

    train_mape = np.mean(np.abs((y_train - y_train_pred) / y_train)) * 100
    test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100

    # Calculate R-squared for train and test datasets
    train_r2 = r2_score(y_train, y_train_pred)
    test_r2 = r2_score(y_test, y_test_pred)

    # Create a dataframe to store the accuracy measures
    accuracy_df = pd.DataFrame({
        'MAE': [train_mae, test_mae],
        'RMSE': [train_rmse, test_rmse],
        'MSE': [train_mse, test_mse],
        'MAPE': [train_mape, test_mape],
        'R-squared': [train_r2, test_r2]
    }, index=['Train', 'Test'])

    return accuracy_df

# Assuming you have your x_train, y_train, x_test, and y_test datasets
accuracy_df = evaluate_random_forest(x_train, y_train, x_test, y_test)
print(accuracy_df)
```

```
             MAE       RMSE          MSE       MAPE  R-squared
Train  13.695880  20.505424   420.472430   6.552999   0.955988
Test   34.249716  52.997189  2808.702045  16.930331   0.711893
```

Lisbon

In [26]:
```
def evaluate_random_forest(x_train, y_train, x_test, y_test):
    # Create and fit the Random Forest regression model
    random_forest_model = RandomForestRegressor()
    random_forest_model.fit(x_train, y_train)

    # Predict the target variable for both train and test datasets
    y_train_pred = random_forest_model.predict(x_train)
    y_test_pred = random_forest_model.predict(x_test)

    # Calculate accuracy measures for train and test datasets
    train_mae = mean_absolute_error(y_train, y_train_pred)
    test_mae = mean_absolute_error(y_test, y_test_pred)

    train_rmse = np.sqrt(mean_squared_error(y_train, y_train_pred))
    test_rmse = np.sqrt(mean_squared_error(y_test, y_test_pred))

    train_mse = mean_squared_error(y_train, y_train_pred)
    test_mse = mean_squared_error(y_test, y_test_pred)

    train_mape = np.mean(np.abs((y_train - y_train_pred) / y_train)) * 100
    test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100

    # Calculate R-squared for train and test datasets
    train_r2 = r2_score(y_train, y_train_pred)
    test_r2 = r2_score(y_test, y_test_pred)

    # Create a dataframe to store the accuracy measures
    accuracy_df = pd.DataFrame({
        'MAE': [train_mae, test_mae],
        'RMSE': [train_rmse, test_rmse],
        'MSE': [train_mse, test_mse],
        'MAPE': [train_mape, test_mape],
        'R-squared': [train_r2, test_r2]
    }, index=['Train', 'Test'])

    return accuracy_df

# Assuming you have your x_train, y_train, x_test, and y_test datasets
accuracy_df = evaluate_random_forest(x_train, y_train, x_test, y_test)
print(accuracy_df)
```

```
             MAE       RMSE          MSE       MAPE  R-squared
Train  11.576949  17.207614   296.101995   5.148007   0.964687
Test   31.824175  45.843358  2101.613437  13.800139   0.760852
```

Paris

```
In [15]: def evaluate_random_forest(x_train, y_train, x_test, y_test):
    # Create and fit the Random Forest regression model
    random_forest_model = RandomForestRegressor()
    random_forest_model.fit(x_train, y_train)

    # Predict the target variable for both train and test datasets
    y_train_pred = random_forest_model.predict(x_train)
    y_test_pred = random_forest_model.predict(x_test)

    # Calculate accuracy measures for train and test datasets
    train_mae = mean_absolute_error(y_train, y_train_pred)
    test_mae = mean_absolute_error(y_test, y_test_pred)

    train_rmse = np.sqrt(mean_squared_error(y_train, y_train_pred))
    test_rmse = np.sqrt(mean_squared_error(y_test, y_test_pred))

    train_mse = mean_squared_error(y_train, y_train_pred)
    test_mse = mean_squared_error(y_test, y_test_pred)

    train_mape = np.mean(np.abs((y_train - y_train_pred) / y_train)) * 100
    test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100

    # Calculate R-squared for train and test datasets
    train_r2 = r2_score(y_train, y_train_pred)
    test_r2 = r2_score(y_test, y_test_pred)

    # Create a dataframe to store the accuracy measures
    accuracy_df = pd.DataFrame({
        'MAE': [train_mae, test_mae],
        'RMSE': [train_rmse, test_rmse],
        'MSE': [train_mse, test_mse],
        'MAPE': [train_mape, test_mape],
        'R-squared': [train_r2, test_r2]
    }, index=['Train', 'Test'])

    return accuracy_df

# Assuming you have your x_train, y_train, x_test, and y_test datasets
accuracy_df = evaluate_random_forest(x_train, y_train, x_test, y_test)
print(accuracy_df)
```

```
             MAE       RMSE          MSE       MAPE  R-squared
Train  17.711682  24.205596   585.910901   6.399474   0.944347
Test   44.621676  60.869772  3705.129083  16.116401   0.642637
```

Rome

In [19]:
```
def evaluate_random_forest(x_train, y_train, x_test, y_test):
    # Create and fit the Random Forest regression model
    random_forest_model = RandomForestRegressor()
    random_forest_model.fit(x_train, y_train)

    # Predict the target variable for both train and test datasets
    y_train_pred = random_forest_model.predict(x_train)
    y_test_pred = random_forest_model.predict(x_test)

    # Calculate accuracy measures for train and test datasets
    train_mae = mean_absolute_error(y_train, y_train_pred)
    test_mae = mean_absolute_error(y_test, y_test_pred)

    train_rmse = np.sqrt(mean_squared_error(y_train, y_train_pred))
    test_rmse = np.sqrt(mean_squared_error(y_test, y_test_pred))

    train_mse = mean_squared_error(y_train, y_train_pred)
    test_mse = mean_squared_error(y_test, y_test_pred)

    train_mape = np.mean(np.abs((y_train - y_train_pred) / y_train)) * 100
    test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100

    # Calculate R-squared for train and test datasets
    train_r2 = r2_score(y_train, y_train_pred)
    test_r2 = r2_score(y_test, y_test_pred)

    # Create a dataframe to store the accuracy measures
    accuracy_df = pd.DataFrame({
        'MAE': [train_mae, test_mae],
        'RMSE': [train_rmse, test_rmse],
        'MSE': [train_mse, test_mse],
        'MAPE': [train_mape, test_mape],
        'R-squared': [train_r2, test_r2]
    }, index=['Train', 'Test'])

    return accuracy_df

# Assuming you have your x_train, y_train, x_test, and y_test datasets
accuracy_df = evaluate_random_forest(x_train, y_train, x_test, y_test)
print(accuracy_df)
```

	MAE	RMSE	MSE	MAPE	R-squared
Train	11.687445	17.100223	292.417611	6.247305	0.960091
Test	30.036942	43.825161	1920.644775	16.055724	0.708525

Vienna

```
In [18]: def evaluate_random_forest(x_train, y_train, x_test, y_test):
    # Create and fit the Random Forest regression model
    random_forest_model = RandomForestRegressor()
    random_forest_model.fit(x_train, y_train)

    # Predict the target variable for both train and test datasets
    y_train_pred = random_forest_model.predict(x_train)
    y_test_pred = random_forest_model.predict(x_test)

    # Calculate accuracy measures for train and test datasets
    train_mae = mean_absolute_error(y_train, y_train_pred)
    test_mae = mean_absolute_error(y_test, y_test_pred)

    train_rmse = np.sqrt(mean_squared_error(y_train, y_train_pred))
    test_rmse = np.sqrt(mean_squared_error(y_test, y_test_pred))

    train_mse = mean_squared_error(y_train, y_train_pred)
    test_mse = mean_squared_error(y_test, y_test_pred)

    train_mape = np.mean(np.abs((y_train - y_train_pred) / y_train)) * 100
    test_mape = np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100

    # Calculate R-squared for train and test datasets
    train_r2 = r2_score(y_train, y_train_pred)
    test_r2 = r2_score(y_test, y_test_pred)

    # Create a dataframe to store the accuracy measures
    accuracy_df = pd.DataFrame({
        'MAE': [train_mae, test_mae],
        'RMSE': [train_rmse, test_rmse],
        'MSE': [train_mse, test_mse],
        'MAPE': [train_mape, test_mape],
        'R-squared': [train_r2, test_r2]
    }, index=['Train', 'Test'])

    return accuracy_df

# Assuming you have your x_train, y_train, x_test, and y_test datasets
accuracy_df = evaluate_random_forest(x_train, y_train, x_test, y_test)
print(accuracy_df)
```

```
             MAE       RMSE          MSE       MAPE  R-squared
Train  12.413285  17.717781   313.919755   5.812448   0.955075
Test   33.550442  51.035775  2604.650342  14.958600   0.635887
```