PREDICTING AIRBNB PRICES USING LOCATION AND IMAGE DATA

(AIRBNB FIYATLARININ, KONUM VE FOTOGRAF VERILERI KULLANARAK TAHMINLENMESI)

by

ÖZGÜN AKALIN, B.S.

Thesis

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

in the

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

FEBRUARY 2025

ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. Dr. Gülfem Işıklar Alptekin for her continuous guidance and support through this process, my wife for her patience, and the taxpayers of Türkiye for high quality tuition-free higher education.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Until recently, short-term or seasonal accommodation needs were commonly met only through businesses, such as hotels and guesthouses. However, in recent years, with a model pioneered by Airbnb, residences that were offered only for the purpose of meeting long-term accommodation needs have begun to be used by their owners.

There are many features that differentiate this model from traditional accommodation services. First of all, the physical characteristics of each rented property (number of rooms, features offered in the residence, etc.) are different both from traditional services and from each other. Secondly, while traditional services are found in certain parts of a city, it is possible to find residences that can be rented through Airbnb in every location. Finally, the services offered in these residences also differ from hotels and guesthouses. Moreover, all these differences require different solutions in determining the price of this service. Hence, the ability to correctly identify the factors affecting the price and to determine the price using such information is very important both for the profitability of the service provider and for the service recipient to evaluate a fair price offer.

In this context, this research was conducted specifically for Istanbul listings to find out which feature affects an Airbnb price to what extent, and which method is the most successful price estimation method. We believe that methods used in this research, and price determinants that are found to affect the price most, can be extended and used for similar price prediction problems. Within the scope of this research, in addition to the standard information provided in all Airbnb listings, we also utilized factors such as the integration of the listing's location with public transportation, the availability of daily necessities in the area, and an analysis of the first photo of the listing. Price estimations were then conducted using various machine learning techniques. In addition, the impact of each feature on price estimation was calculated individually.

According to the research findings, while location information is not as crucial as the physical attributes of a residence in price estimation, it was still found to be an

important factor. In terms of public transportation, proximity to minibus and pier stops emerged as key influences, whereas for meeting daily needs, cultural activities and educational facilities were significant. The analysis of the listing photo using large language model, however, was found to have no significant impact on the results.

Keywords : machine learning, feature engineering, airbnb, llm, regression

RÉSUMÉ

Jusqu'à récemment, répondre aux besoins d'hébergement à court terme ou saisonnier n'était possible que par le biais d'entreprises telles que les hôtels et les auberges. Cependant, ces dernières années, avec un modèle lancé par Airbnb, les maisons proposées uniquement pour répondre au besoin d'hébergement à long terme ont commencé à être utilisées par leurs propriétaires pour répondre à ce besoin.

De nombreuses caractéristiques différencient ce modèle des services d'hébergement traditionnels. Tout d'abord, les caractéristiques physiques de chaque bien loué (nombre de pièces, prestations proposées au sein de la résidence, etc.) sont différentes tant des prestations traditionnelles que les unes des autres en elles-mêmes. Deuxièmement, même si les services traditionnels sont localisés dans certaines parties d'une ville, il est possible de trouver des logements pouvant être loués via Airbnb dans chaque endroit. Enfin, les services proposés dans ces résidences diffèrent également par rapport aux hôtels et auberges. Toutes ces différences nécessitent des solutions différentes pour déterminer le prix de ce service. Identifier correctement les facteurs affectant le prix et déterminer le prix à l'aide de ces informations est très important à la fois pour la rentabilité du prestataire de services et pour l'évaluation d'une offre de prix équitable par la partie qui reçoit le service.

Dans ce contexte, cette recherche a été menée spécifiquement sur les annonces d'Istanbul pour savoir quelle caractéristique affecte le prix d'un Airbnb et dans quelle mesure, et quelle méthode est la méthode de prévision de prix la plus efficace. Dans le cadre de la recherche, outre les informations telles que le nombre de chambres, les informations sur le quartier et la note moyenne attribuée par les clients à cette annonce, qui sont standard dans toutes les annonces Airbnb, le degré d'intégration de l'emplacement de l'annonce avec le public transport, quels besoins quotidiens peuvent être satisfaits à cet endroit et dans quelle mesure, ainsi que l'analyse de la première photo de la publicité et les prévisions de prix ont été réalisées à l'aide de différentes techniques d'apprentissage automatique. De plus, pour chacune de ces fonctionnalités, il a été calculé dans quelle mesure chaque fonctionnalité affecte l'estimation du prix.

Selon les résultats de la recherche, il a été constaté que les informations obtenues grâce à la localisation sont également importantes pour l'estimation du prix, même si elles ne sont pas aussi importantes que les caractéristiques physiques d'une maison. En termes de transports publics, la proximité des arrêts de minibus et des quais, ainsi que les activités culturelles et éducatives ont été mises en avant pour répondre aux besoins. Il a été observé que l'analyse de la photo publicitaire via la modèle de langage à grande Échelle n'a pas eu d'effet significatif sur les résultats.

Mots Clés : machine learning, feature engineering, airbnb, llm, regression

ÖZET

Kısa veya dönemsel konaklama ihtiyaçları, yakın zamana kadar ağırlıklı olarak otel ve pansiyon gibi işletmeler aracılığıyla karşılanmaktaydı. Ancak son yıllarda, Airbnb'nin öncülük ettiği model sayesinde, uzun süreli barınma amacıyla kullanılan konutlar da bu ihtiyaca yanıt vermek üzere sahipleri tarafından kiralanmaya başlandı.

Bu modeli geleneksel konaklama hizmetlerinden ayıran birçok özellik bulunmaktadır. Öncelikle, kiralanan mülklerin fiziksel özellikleri (örneğin oda sayısı, sunulan olanaklar vb.), hem geleneksel hizmetlerden hem de kendi aralarında önemli farklılıklar göstermektedir. İkinci olarak, geleneksel konaklama hizmetleri genellikle bir şehrin belirli bölgelerinde yoğunlaşırken, Airbnb aracılığıyla kiralanabilen konutlar hemen her konumda bulunabilmektedir. Son olarak, bu konutlarda sunulan hizmetler de otel ve pansiyonlara kıyasla çeşitlilik göstermektedir. Bu farklılıklar, konaklama hizmetlerinin fiyatlandırılmasında farklı yaklaşımlar gerektirmektedir. Fiyatı etkileyen faktörlerin doğru bir şekilde belirlenmesi ve bu bilgiler doğrultusunda adil bir fiyatlandırma yapılması; hem hizmet sağlayıcıların karlılığı hem de hizmet alanların makul bir fiyat teklifi değerlendirmesi açısından büyük önem taşımaktadır.

Bu doğrultuda, İstanbul'daki Airbnb ilanlarını temel alan bu araştırmada, bir konutun fiyatını hangi faktörlerin ne ölçüde etkilediği ve en başarılı fiyat tahmin yönteminin hangisi olduğu analiz edilmiştir. Araştırmada, tüm Airbnb ilanlarında standart olarak sunulan oda sayısı, ilçe bilgisi ve müşteri puanı gibi temel bilgilerle birlikte; ilanın konumunun toplu taşımaya entegrasyonu, günlük ihtiyaçların ne ölçüde karşılanabildiği ve ilanın ilk fotoğrafının analizi gibi unsurlar da değerlendirilmiştir. Fiyat tahminlemesi için farklı makine öğrenmesi teknikleri uygulanmış ve her bir faktörün fiyat üzerindeki etkisi hesaplanmıştır.

Araştırma sonuçlarına göre, bir konutun fiziksel özellikleri kadar olmasa da, konuma dayalı bilgilerin de fiyat tahminlemesinde önemli bir rol oynadığı görülmüştür. Toplu taşıma açısından minibüs ve iskele duraklarına yakınlık, ihtiyaçların karşılanması açısından ise kültürel aktiviteler ve eğitim olanakları öne çıkan faktörler olarak belirlenmiştir. Öte yandan, ilanın ilk fotoğrafının büyük dil modeli aracılığıyla analiz edilmesinin fiyat

tahminine anlamlı bir katkı sağlamadığı tespit edilmiştir.

Anahtar Kelimeler : makine ogrenmesi, yapay zeka, airbnb, llm, regresyon

1 INTRODUCTION

With advances in digital technologies that allow peer-to-peer connections through standardized processes via websites or mobile applications, people can now purchase services from other people instead of established organizations. This change resulted in the emergent of disruptive business models in different areas. The new networked model introduces variety in the services offered, and customers can make their decisions based on their specific needs. With this model, traditional suppliers lose market share to individuals unless they take the necessary measures (Oskam and Boswijk, 2016). Examples of this new business model called the "sharing economy" are the sharing of rental properties and ridesourcing. For example, Uber connects private car owners with passengers (Jin et al., 2018). Similarly, Airbnb connects home owners to people in need of accommodation. These companies provide the digital platforms that allow common search functionalities, reviews, integration with payment systems, dispute handling, localized services, compliance with regulations, and more. Trust, an important factor between both parties, is established through reviews, certifications (Ert and Fleischer, 2019), and special to Airbnb, through listings' images (Hu et al., 2024) and status of the host, such as superhost status (Xie and Mao, 2017).

Airbnb was born in 2007 when two university graduates rented their apartment to attendees of a conference in San Francisco. Since then, Airbnb has reached more than 5 million hosts and 2 billion guest arrivals on their platform [1]. Airbnb allows owners to list their properties with information such as the number of rooms, beds, images of the property, nightly price, etc. Customers can browse these listings and reserve a property for the number of nights of their choosing, based on availability. This allows any property owner to enter the tourism accommodation market, without the need for special advertisements through different channels (Guttentag, 2015).

Fig. 1.1 shows the search user interface of the Airbnb website, where users can browse the rental properties by three methods : Firstly, they can browse a list, where quick information such as the featured image, name of the listing, and the nightly price is shown. Secondly, a filter component allows users to filter by different fields such as price,

1. https ://news.airbnb.com/about-us/

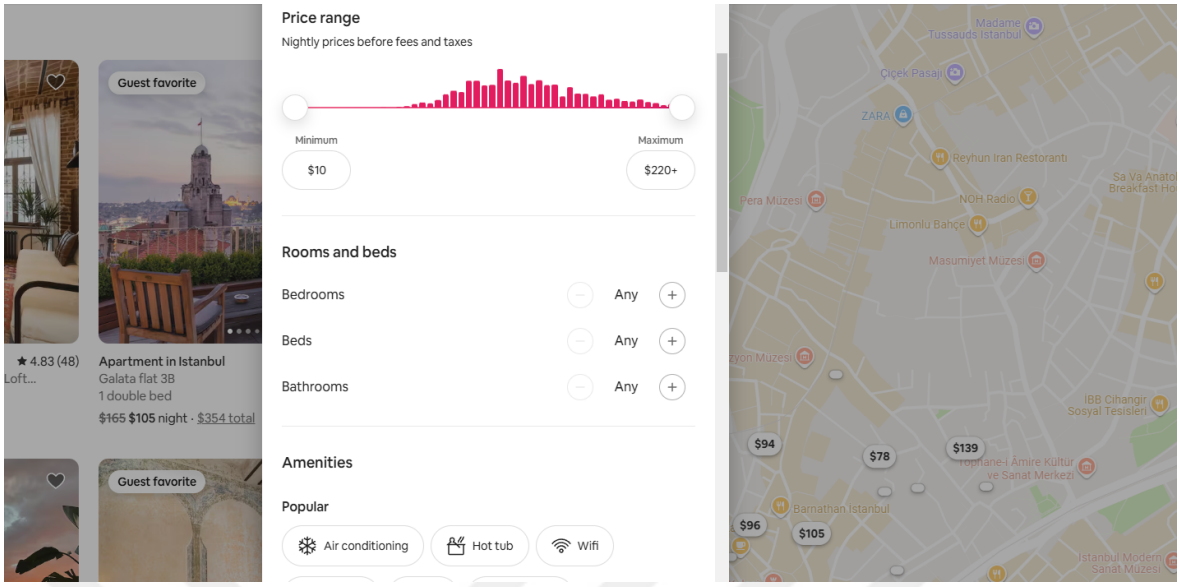

FIGURE 1.1 – Airbnb User Interface

number of beds, and amenities. Lastly, users can use the map to check the listings. The dedicated page of an Airbnb listing shows other details, including scores in different categories, recent reviews, description, and more.

Unlike hotel rooms, where rooms within the same category of a hotel are nearly identical, each Airbnb listing is unique. Factors such as the number of rooms, furnishings, location, amenities, view, number of reviews, and review scores contribute to the distinct characteristics of each listing. Given this diversity, a detailed analysis is essential to determine the appropriate nightly price.

Outside of variables directly available within an Airbnb listing, location-based characteristics should also be considered, as location is important for lodging units, just as it is for hotels (Yang and Mao, 2020). In addition to classic measurements such as distance to tourist centers, characteristics such as walkability around a listing should also be investigated (Hall and Ram, 2019). Moreover, although each listing has a featured image and many other images within the listing details, few research has made use of the images when predicting the prices of an Airbnb listing. With the increasing performance of Large Language Models (LLM)Touvron et al. (2023), and their capabilities to support multi-modal input Huang et al. (2024), it might be possible to benefit from LLMs when predicting Airbnb prices.

1.1 Research Questions

By analyzing existing listings, this thesis aims to accurately predict a listing's price based on its properties using various machine learning methods. Additionally, the research seeks to determine the impact of each price determinant and address the following research questions :

1- How accurately can machine learning models predict the prices of Airbnb listings ?

2- To what extent do location-based characteristics influence the price of an Airbnb listing ?

3- Can large language models effectively utilize the featured images of an Airbnb listing as input for price prediction models ?

2 LITERATURE REVIEW

In this chapter, an overview of the work related to this research will be presented. The related work is divided into different sections, each section representing each research question's focus.

2.1 Machine Learning Models to Predict Airbnb Listing Price

Price prediction performance of different machine learning models has been analyzed in great detail in the literature. Often, results of multiple models are compared at the end of the work.

Table 2.1 shows some of the recent work that used machine learning methods to predict Airbnb prices. Most of the papers used Kaggle or InsideAirbnb as their data source and used a different number of methods and compared them. According to the results, Random Forest and XGBoost are the leading methods.

TABLE 2.1 – Machine Learning Methods in Recent Literature

Article	Data Source	Cities	Methods	Results
Zhu et al. (2020)	Kaggle	New York	GeneralizedAdditiveModel Linear Regression Deep Neural Network Random Forest XGBoost	XGBoost performs best, closely followed by Random Forest
Wang (2023)	Kaggle	Boston	Linear Regression Random Forest K-NN Gradient Boosting	Gradient Boosting outperformed other methods
Tang et al. (2024)	Inside Airbnb	Sydney	OLS Lasso Regression tree XG-Boost Model stacking RandomForest ExtraTreesRegressor Gradient Boost Regressor LGBMRegressor CatBoostRegressor	Cat Boost Regressor and Gradient Boost Regressor performed highest and similarly.
Lektorov et al. (2023)	Comcast and Kaggle	New York	Random Forest K Neighbors Support Vector Regressor Decision Trees Extra Trees XGB Regressor	Random Forest and Extra Trees achieved highest R2 scores
Camatti et al. (2024a)		Multiple cities in Netherlands	Random forest Decision tree Neural networks Linear regression GLM	Random Forest had the best R2 score
Islam et al. (2022)	Airbnb Scrap	San Jose County	Linear Regression Random Forest XGBoost	XGBoost performed best
Alves (2024)	Inside Airbnb	Lisbon and Porto	Random Forest XGBoost Artificial Neural Networks	XGBoost performed best
Meijer (2022)	Inside Airbnb	Amsterdam	SVR Ridge Regression XGBoost	XGBoost performed best
Gangarapu and Mernedi (2023)	Kaggle	10 European Cities	Decision Tree Regressor Random Forest Regressor Support Vector Machine Gradient Boost Regressor Ridge Regressor Lasso Regressor	Random Forest Regressor significantly scored higher
Camatti et al. (2024b)	AirDNA	Cities in Netherlands	Random Forest Linear Regression Neural Networks	Random Forest reached highest R2 value

2.2 Using Location-Based Features

Enhancing existing Airbnb listing data using the location of listings has been studied using different location-related data sources for different cities.

Camilleri, Gabriella (2023) used OpenStreetMap's API named Overpass Turbo to calculate the distance between listings to popular attraction points such as beaches, nightclubs, and historical sites in the Maltese Islands. The results show that for XGBoost, the following features make it to the top 20 important features based on Shapley values : "walkable_distance_capital" is in 8th order, "walkable_distance_nightclub" is in 11th order, "walkable_distance_historical" is in 14th order, and finally "walkable_distance_beach" is in 19th order. However, the R^2 value of all models which include KNN, Linear Regression, Ridge Regression, XGBoost, and CatBoost showed only minor improvement.

Thackway et al. (2022) also used OpenStreetMap to find the proximity of Airbnb listings to amenities and points of interest. These include beaches, hospitals, universities, public transportation stops, and parks. They have used this information along with the listing information to calculate how Airbnb density affects the sale prices of houses.

Shabrina and Morphet (2022) used two different sources to enhance existing data with location data to predict the rent distribution in London. The first is "Visit Britain", which lists 88 most visited tourist attractions in London, as well as the number of annual visitors to the attractions. The second is the "Open Trip Planner" in order to calculate the travel time to tourist destinations.

Panahandeh et al. (2025) included "FourSquare" data and distance of listing to city center and various destinations under different categories, such as university or hospital, to find the price determinants of an Airbnb listing. Foursquare was used to obtain the density of places, which was interpreted as the proximity to landmarks and services in the city. Among the top 10 significant features, the distance from the city center was in the sixth order, and the FourSquare points count was in the seventh order. Additionally, XGBoost was the model that performed the best and additional features increased the R^2 value of the model by 2%.

Jiang et al. (2023) calculated distance between listings in Shanghai and the closest landmark, transportation stop, mall, and competitor. Among the distance-related variables, distance to landmarks was the most important field.

Using a geographic information system, Chica-Olmo et al. (2020) calculated the distance from each listing to the city center, beaches, and the nearest point of interest. In addition, the density of pedestrian zones and the traffic noise levels around the listing were included. The addition of location variables increased the R^2 value from 0.367 to 0.45.

Peng et al. (2020) added distance to the closest landmark for each listing as a prediction factor for the price and used linear regression, XGBoost, Support Vector Regression, and Deep Neural Network to predict prices. The results show that the XGBoost model had the highest R^2 value, and location information increased R^2 from 46.1 to 47.5.

Schwarzová (2020) merged the Airbnb data with Foursquare data (for distance to nearest restaurant, park, etc.), CzechCrime.org (for criminality index of a location), and "opendata.praha.eu" (for location of public transport stations). The results show

that the features added using location made only a minor difference in the R^2 values.

Alves (2024) added geospatial features such as distance to the nearest public transportation station, airport, city center, grocery stores, restaurants, and tourist attractions. These features had little to no effect on the R^2 values of different models.

Williams (2023) measured the distance between the listings and the airport in Cape Town. The distance from the nearest attraction point was also calculated. These added features did not have a significant effect on the R^2 value.

2.3 Processing Airbnb Images Using Large Language Models

Large Language Models are complex transformer language models with hundreds of billions of parameters and have language modeling and task-solving capabilities (Zhao et al., 2023). Examples of recently released Large Language Models that support image input include models from the GPT, Claude, Gemini, NVLM, and Llama product families (Dai et al., 2024). The uses of multi-modal LLMs include classification of images Wu et al. (2025), analysis of road conditions in traffic de Zarzà et al. (2023), providing diagnosis given patient history and relevant images Panagoulias et al. (2024).

Although the use of images of an Airbnb listing has been studied in the literature, research on the processing of Airbnb images using large language models to predict prices has not been found.

Existing work on Airbnb listing images includes using CNN and MobileNet to process multi-modal input Tan et al. (2024), using artificial intelligence (everypixel.com) to assign an aesthetic score by Lin et al. (2024), using SURF and OpenCV to map listing images to words Tang and Sangani (2015), calculating BRISQUE scores using image processing and finding values such as hue, saturation using OpenCV Garcia (2023).

The lack of use of LLMs in research related to Airbnb price prediction shows that there is a need in the field to evaluate the performance of LLMs in this subject.

To conclude the literature review section, answers to some of the research questions have been explored, and some have not. While different machine learning models have been studied extensively, we see that location-based price determinants are usually limited to proximity to historical centers and city centers, few articles taking other

location-based information into consideration. In this research, further research is done to calculate the effects of more location related variables. Finally, lack of use of LLMs for Airbnb listing images shows that recent advances in multi-modal LLMs can be studied and findings can be contributed to the literature. In this research, three different LLM prompts and their outputs are investigated, assessing their impact in price prediction.

3 DATASETS

3.1 InsideAirbnb Dataset

InsideAirbnb is a website that provides Airbnb listing data publicly. As of December 2024, InsideAirbnb has listing information of 119 cities across the globe. For each city, usually every three months, a new set of data is published. This data is a snapshot of listings at a particular date and does not include historical information, except for a few fields. For this research, Istanbul listings data from 30 June 2024 are used.

Figure 3.1 shows the visualization of the listings in some areas in Istanbul, where red dots represent entire homes/apartments and green dots represent shared rooms.

Table 3.1 lists descriptions of some of the fields in the InsideAirbnb dataset, obtained from InsideAirbnb Data Dictionary. [1]

TABLE 3.1 – Some of InsideAirbnb Data Fields

Field	Description
id	Airbnb's unique identifier for the listing
latitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
longitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
room_type	[Entire home/apt|Private room|Shared room|Hotel]
accommodates	The maximum capacity of the listing
bathrooms	The number of bathrooms in the listing
bedrooms	The number of bedrooms
beds	The number of bed(s)
price	daily price in local currency
minimum_nights	minimum number of night stay for the listing (calendar rules may be different)
number_of_reviews	The number of reviews the listing has

The listing data for Istanbul contains a total of 31 758 entries and 75 columns, where each entry is a different rental property and each column is a property describing the entry.

1. https ://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4

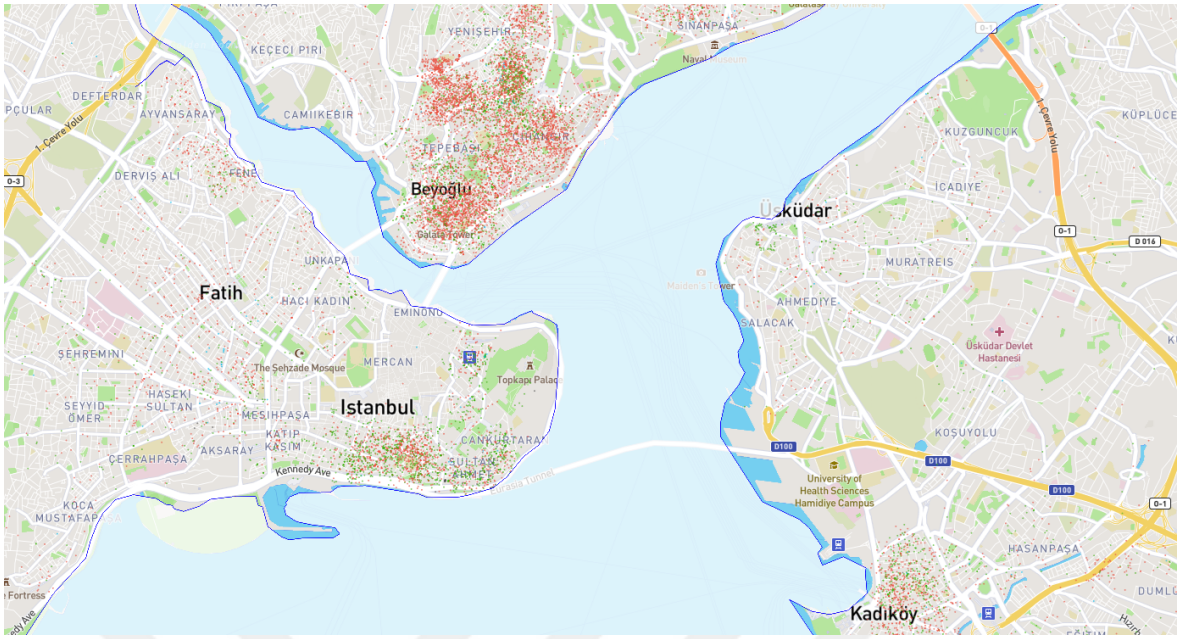

FIGURE 3.1 – InsideAirbnb Istanbul Visualization

The complete list of columns in the airbnb dataset can be found in APPENDIX A.

The 75 columns describe the listing in different aspects. For example, while some columns such as "property_type", "bedrooms" and "room_type" inform about physical properties; "neighbourhood", "latitude" and "longitude" give location-related information, and "review_scores_rating" and "reviews_per_month" provide data about reviews.

The target variable this research aims to predict is the "price" column, which shows the nightly cost of a listing.

3.2 Values Generated by Large Language Model

There are many LLMs that support image input. Some of these models are commercial, and their services are offered via a paid API. Some of them are released publicly and can be downloaded and used with a local computer. The sizes and hard hardware requirements of downloadable models can vary and may require a cluster of machines instead of a single workstation.

For this thesis, the Llama 3.2 Vision 11B version is used to be able to use the model offline and omit the costs of using an LLM API. Llama 3.2 Vision is an LLM released by Meta and supports multi-modal input. It accepts both image and text input and

produces an output by taking the prompt into consideration. It requires 64 GB of RAM and a high-end GPU.

For each listing, the InsideAirbnb dataset contains a field that contains the URL of the featured image. The featured image of a listing is the first image that is displayed on the user interface. For each listing, using a custom application, the URL is visited and the featured image is downloaded locally. The image resolution size of the Llama 3.2 Vision model is 1120 x 1120 pixels. To comply with this size limit, all images were re-scaled.

Since hosts aim to present their properties as attractive as possible, the featured image can be of interior such as the living room, or it can be of exterior such as a photo of the balcony, or an image from outside the rental home's building. In some cases, the featured images are of no relevance to properties, and are photos from touristic places in Istanbul.

For this part, three different prompts are given to the LLM and three different outputs are saved.

— You are given the featured image for a rental home listing. As a real estate specialist, produce a score to assess the price score of the rental unit between 1 and 10, 1 for the lowest score and 10 for the highest score.
— You are given the featured image for a rental home listing. Assess whether the image is of the interior or exterior of a home. For interior, output 1, and for exterior, output 0.
— You are given the featured image for a rental home listing. Produce a score to assess the luxury score of furniture and decoration between 1 and 10, 1 for the lowest score, and 10 for the highest score.

In order to be able to use the price score in the regression models, an integer output is requested from LLM. But behind the scenes, LLM reasoning is also checked for sample images.

An example of the reasoning of the LLM for figure 3.2 is as follows : "This rental unit

would likely receive a score of 2 out of 10 because it lacks sufficient lighting, furniture, and amenities. The room is sparsely furnished with only two pillows on the bed and there are no other furnishings or decor to make the space feel welcoming or functional.

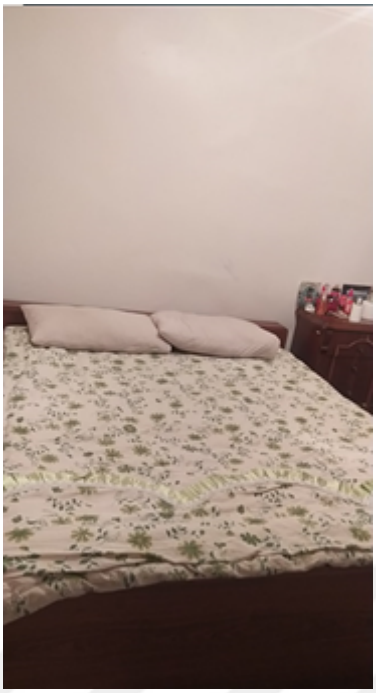

FIGURE 3.2 – Example Featured Airbnb Image

Additionally, the lack of natural light makes the room feel dark and uninviting, which would negatively impact the overall comfort and livability of the space."

Generating image-related data using LLM took around 3 seconds for each image. For a total of 31103 images, it took 26 hours to create results for a single prompt. For three different prompts, 78 hours were needed.

3.3 34 Minutes Istanbul Dataset

Istanbul Metropolitan Municipality provides a dataset showing how each location in Istanbul meets the different daily needs of urban life [2]. These needs are represented by different categories. The names and brief criteria can be found below.

- **Shelter :** Housing type diversity, urban density, pollution levels, police stations
- **Work :** Business centers, offices, plazas
- **Meeting Your Needs :** Grocery stores, markets, banks, local services
- **Cultural Activity :** Cultural centers, social facilities, theaters, cinemas
- **Learning :** Schools, libraries, nursing centers

2. https ://data.ibb.gov.tr/en/dataset/34-dakika-istanbul-cesitlilik-indeksi/resource/f4884364-c87a-4987-810d-2686537741ac

— **Health :** Family health centers, hospitals, pharmacies

— **Transport :** Public transportation stops, mobility score

— **Spending Time :** Green spaces, squares, places of worship, gyms, cafes, restaurants

— **Affordability :** Affordability of the housing market

— **Walkability :** connections between transportation types, street texture quality, road network connectivity

For each index, a separate file is provided by Istanbul Metropolitan Municipality open data portal, where an index score between 0 and 100 is assigned to a hexagon within Istanbul.

For each Airbnb listing, using its latitude and longitude information, its index value is found by finding hexagon to which the Airbnb listing belongs. These values are then added to the dataset as new variables.

The complete list of new features for each listing can be found in APPENDIX B.

3.4 Istanbul Public Transportation Dataset

Istanbul Metropolitan Municipality also releases public transportation data through its open data portal. This data includes locations of stops for different transportation methods such as taxis, subways, and ferries.

The files for each transportation method lists all stations for that method in JSON format, where coordinates are also available, in addition to names of the stations. For this thesis, for each Airbnb listing, number of stops for each different transportation method within different distances are calculated. These distances are 0.5 kilometer, 1 kilometer, 3 kilometers, and 5 kilometers.

After this feature enrichment, each Airbnb listing also has the number of stops within different distances for different transportation methods, showing how well that rental property is integrated into the Istanbul public transportation network.

The complete list of variables added at this step can be found in APPENDIX C.

4 EXPLORATORY DATA ANALYSIS AND PRE-PROCESSING

This chapter describes the pre-processing steps and analysis of Airbnb data in detail.

4.1 Special Characters

Some columns of the data are formatted in a way that contains special characters. To be able to use them in regression models, they need to be converted to number format.

The first of such columns is the price variable. An example value from the dataset is "$4,108.00". First, the "$" character is removed from the data. Then, the "," character is removed, and the resulting value is converted to float type.

Additionally, "host_response_rate" and "host_acceptance_rate" columns hold percentage values, and '%' symbol is removed from values.

4.2 Technical Variables

Among the 75 columns in the dataset, some are about technical information or elements that can be used in user interfaces. Since these do not contribute to the price of a listing, they are removed.

The list of removed columns due to this reason are as following : "id", "listing_url", "scrape_id", "last_scraped", "source", "picture_url", "host_id", "host_url", "host_name", "host_thumbnail_url", "host_picture_url", "calendar_updated", "calendar_last_scraped".

4.3 Text Data

As textual features of a listing such as "description" are not used in this research, the following features are deleted : "name", "description", "host_about", "bathrooms_text", "neighborhood_overview".

4.4 Features Unrelated to Price

"host_location", "host_neighbourhood", and "license" contain information about the host or regulations that are not relevant to the price of a listing. Additionally,

- "maximum_nights_avg_ntm"
- "minimum_nights_avg_ntm"
- "minimum_minimum_nights"
- "maximum_minimum_nights"
- "minimum_maximum_nights"
- "maximum_maximum_nights"

columns provide historical values of minimum, maximum, or average number of days a listing can be rented in the last year, which are not of relevance to a current price.

4.5 Date Columns

In the dataset, three columns contain date information, and these are "host_since", "first_review" and "last_review" columns. These columns are replaced by three new columns named

- "host_account_age_in_days"
- "days_since_first_review"
- "days_since_last_review"

These hold the number of elapsed days since the date till June 30, 2024 (airbnb listings file's content generation date).

4.6 Boolean Values

For each column that takes boolean values, its values are converted to 1 and 0.

Columns with boolean values are : "host_is_superhost", "host_has_profile_pic", "host_identity_verified", "has_availability", "instant_bookable".

4.7 Categorical Variables

For "host_response_time" and "room_type", which have 4 distinct values each, one-hot encoding is used. Similarly, for "neighbourhood_cleansed", which can take 39 unique values, is one-hot encoded.

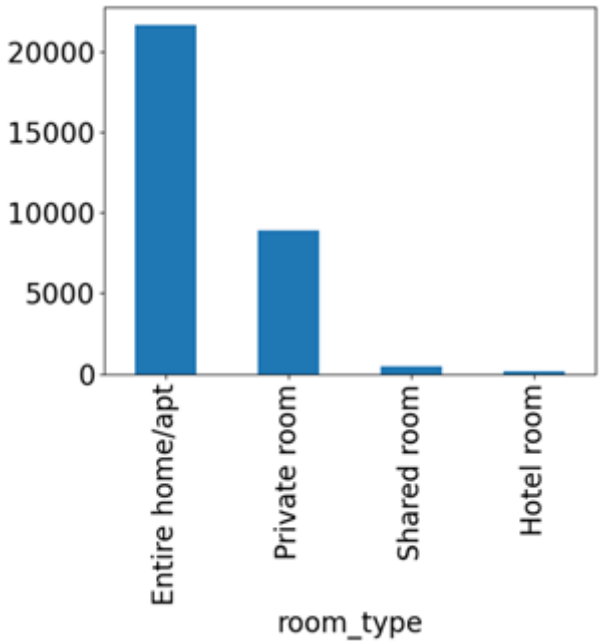

FIGURE 4.1 – Room Type Values - Histogram

Lastly, "host_verifications" feature stores an array of different host verifications of the listing, and the number of different verification methods for the listing is calculated and stored in "host_verifications_method_count" column.

4.8 Amenities

For "amenities", which keeps an array of items for each Airbnb listing, number of amenities is calculated and stored in "amenities_count" feature.

Additionally, amenities for the listings are further analyzed. For Istanbul Airbnb dataset, there are a total of 5413 unique amenities. However, most of these unique values are variations of each other. For example, there are different amenity names for each different connection speed of Wifi. Among the 5413 distinct amenities, 264 belong to Wifi category and 1442 belong to HDTV category. Although different Internet connection speeds and different TV screen sizes with different subscription combinations can

offer a different experience during the stay, they are mapped to general amenity names such as Wifi or HDTV, indicating the availability of such service.

Table 4.1 shows some example variations for some of the amenities.

Table 4.1 – Example Amenity Variations

Example Amenity Variations For Wifi	**Example Amenity Variations For HDTV**
Wifi	HDTV
Wifi – 22 Mbps	HDTV with Netflix
Wifi – 24 Mbps	HDTV with standard cable
Wifi – 25 Mbps	32" HDTV with standard cable
Wifi – 48 Mbps	32" HDTV
Fast wifi – 93 Mbps	HDTV with premium cable
Fast wifi – 50 Mbps	55" HDTV with Netflix

Following the same approach for similar amenities reduced the number of unique amenities from 5413 to 147, allowing the use of one-hot encoding in a meaningful way. Table 4.2 shows new mapping of amenities that contain either of the keywords. Furthermore,

if an amenity contains either of of the following keywords, it is mapped to the keyword, removing other characters, and hence, removing variations : 'oven', 'TV', 'coffee maker', 'stove', 'alarm', 'game console', 'pool', 'sound', 'housekeeping', 'fireplace', 'gym', 'washer', 'high chair', 'hot tub', 'exercise equipment', 'toys', 'clothing storage', 'dryer', 'ski-out', 'beach access', 'crib', 'backyard', 'games', 'ev charger', 'kitchen', 'balcony', 'changing table', 'heating'. This method turns both "Bosch oven" and "Siemens oven" amenities into 'Oven' amenity, for example.

Table 4.2 – Amenity Generalization

Amenities That Contain Keywords	**Mapped New Amenity**
wifi, mbps	Wifi
shampoo, conditioner, soap, shower gel	Booklet
parking, garage, carport	Parking
refrigerator, fridge	Refrigerator
bbq, grill	BBQ
baby, guards, child	child friendly

Additionally, amenities are further investigated to estimate the total daily benefit to the customer. For this, the following algorithm is used.

Algorithm 1: Amenity Cost Algorithm

$totalAmenityCost \leftarrow 0$;
while $length(listingAmenities) \neq 0$ **do**
    $currentAmenity \leftarrow listingAmenities.getFirstElement()$;
    $totalAmenityCost \leftarrow totalAmenityCost + getDailyCost(currentAmenity)$;
    $listingAmenities \leftarrow listingAmenities.remove(currentAmenity)$;
end

The algorithm uses the method "getDailyCost()", which returns the daily cost of the amenity. The daily cost is calculated as follows :

— For subscription-based services, such as Wifi, divide the monthly cost by 30.

— For electronic devices, such as TV, assume 5-year lifetime and divide the sale price by 1825

— For single use items, such as soap, use sale price.

— For household items used daily, such as glasses and pillows, assume a lifetime of 6 months and divide the sale price by 182.

— For furniture such as tables and chairs, assume a 5-year lifespan and divide the sale price by 1825.

— For amenities where the cost cannot be computed in a generalized way, such as "Marina View" or "Host greets you", use 0.

The sale prices of amenities, where applicable, are found by getting the lowest price of search results' first page after searching for the name of the amenity in Turkish at an e-commerce website based in Turkey. The complete list of sale prices and the calculated daily costs of each amenity subject to cost calculation can be found in APPENDIX D.

4.9 Empty Variables

Variable "neighbourhood_group_cleansed" contains only null values for the whole dataset and is removed.

4.10 Other

"neighbourhood" column contains city and country information, which is not required as we are limiting the dataset to Istanbul listings. It also includes province and street data, but the order categories for each row do not follow the same pattern and are unstructured. "neighbourhood_cleansed" column is preferred over this. "room_type" is a cleansed version of data in "property_type" column. In summary, "property_type" and "neighbourhood" are removed.

4.11 Target Variable

The "price" column values is null for 2526 listings in the Airbnb dataset. Since this is the target column, rows with null values in this field are deleted.

TABLE 4.3 – Target variable "Price" statistics

Field	Value
count	29232.00
mean	3947.35
std	35136.45
min	80.00
25%	1258.00
50%	2000.00
75%	3286.00
max	3310226.00

For the target variable, the distribution shows that there are outliers present. In order to remove outliers, listings with price larger than 8000 are removed from the dataset. This removes 3.9% of the data.

Log transformation is also applied on the target variable to handle nonlinear relationship between dependent and independent variables (Benoit, 2011). Figure 4.2 shows the price distribution after this step.

The correlation between features and the target variable can be found in Figure 4.3. In this figure, features with a correlation less than 0.15 are omitted. The correlation map shows that the physical features and amenities of a listing have the highest correlation, followed by public transport proximity. Location indexes and image features are not

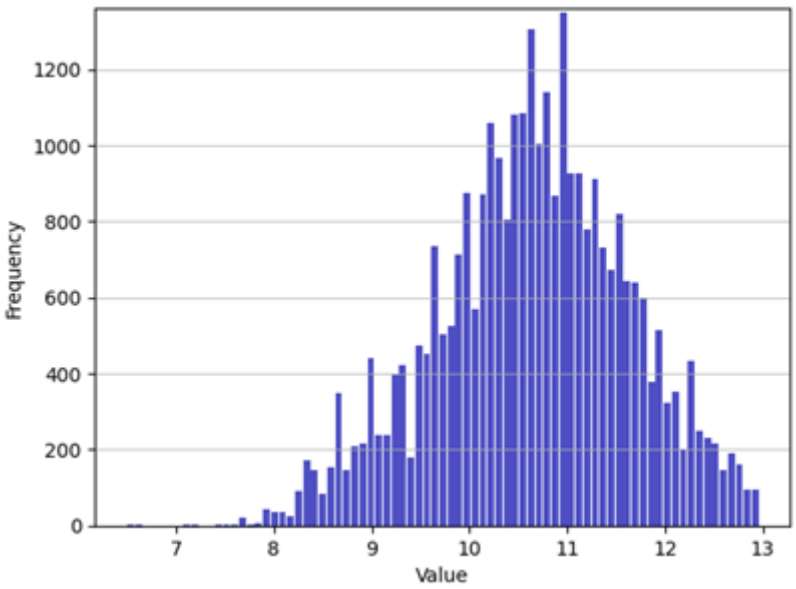

FIGURE 4.2 – Price Distribution

shown on the figure as their correlation is less than 0.15.

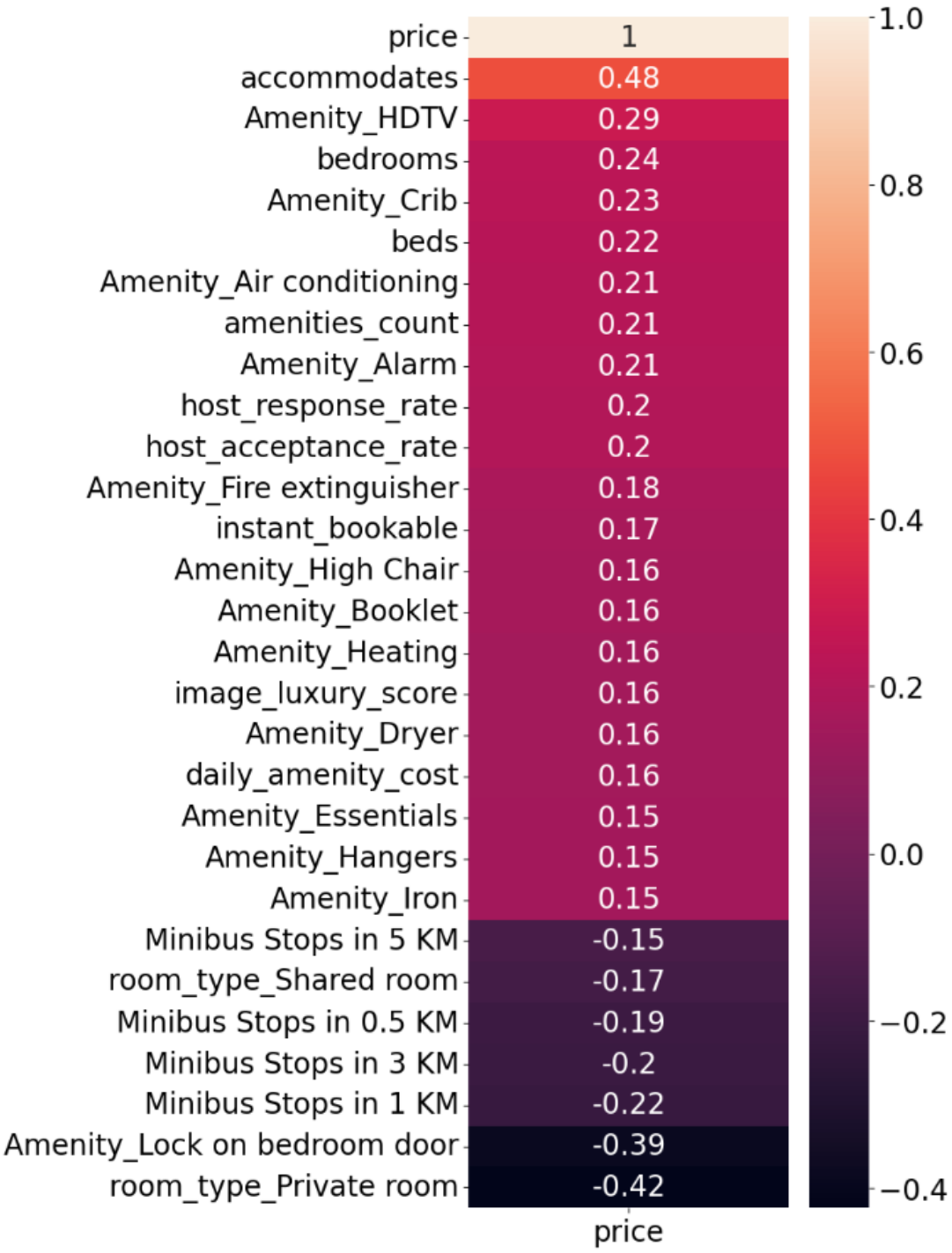

FIGURE 4.3 – Feature Correlation With Price

5 BACKGROUND

This chapter describes the machine learning methods and evaluation metrics used in the research.

5.1 Machine Learning Methods

Random Forest Regression : Suggested by Breiman (2001), ensemble decision trees are used to predict continuous variables.

Gradient Boost : A robust algorithm in which weak learners are merged to create a strong learner (Friedman, 2001).

XGBoost : Implemented as a variation of the gradient boost algorithm, XGBoost utilizes decision trees and runs in parallel with optimizations(Chen and Guestrin, 2016).

Linear Regression : A model to estimate the relationship between a dependent variable and independent variables. It is used in this research because it is the most widely used model in the existing Airbnb price prediction literature.

5.2 Evaluation Metrics

To compare the performance of different models; Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Adjusted R^2 are used.

Mean Absolute Error (MAE) : MAE, as described by Sammut and Webb (2011), is the average of absolute errors $|y_i - \widehat{y_i}|$, where y_i is the actual value and $\widehat{y_i}$ is the predicted value.

Root Mean Squared Error (RMSE) : RMSE is the standard deviation of errors between predicted and actual values (Chai and Draxler, 2014).

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}}$$

Mean Squared Error (MSE) : MSE is the average of the square of errors between actual and predicted values (Willmott and Matsuura, 2005).

$$\text{MSE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}$$

Adjusted $\mathbf{R^2}$: Adjusted R^2 Miles (2005), an extension of the coefficient of determination R^2 proposed by Wright (1921), is calculated as :

$$R^2_{adj} = 1 - \left(1 - R^2\right) \frac{N - 1}{N - k - 1}$$

where N is the sample size and k is the number of predictor variables. R^2 measures how well the model fits the data.

$$R^2 = 1 - \frac{\sum \left(y_i - \hat{y}_i\right)^2}{\sum \left(y_i - \bar{y}\right)^2}$$

Adjusted R^2 is preferred to R^2 because it accounts for the number of predictor variables in the model. This adjustment is particularly important when comparing models trained with different number of features, such as those that include location-based variables and those that use only base Airbnb variables. The adjusted R^2 penalizes the inclusion of additional predictor variables, addressing the problem highlighted by Miles (2005), where R^2 always increases with each added variable, potentially misleading the evaluation of model performance.

6 SOLUTION METHODOLOGY

Figure 6.1 shows the steps of the methodology followed. InsideAirbnb listing dataset is preprocessed and merged with Istanbul index data and Istanbul public transportation data. In addition, images from the listings are downloaded, processed, and fed to the LLAMA 3.2 Large Language Model to generate price scores, which are then merged with previous data to create a final data set. The final data set contains Airbnb listing properties, index scores, proximity to public transport, and price scores. This final data is then used to train and test machine learning models. The following machine learning algorithms are used.

- Linear Regression
- XGBoost
- Gradient Boost
- Random Forest Regression

The following sections in this chapter explain the model training approach, hyperparameter tuning and technology stack used.

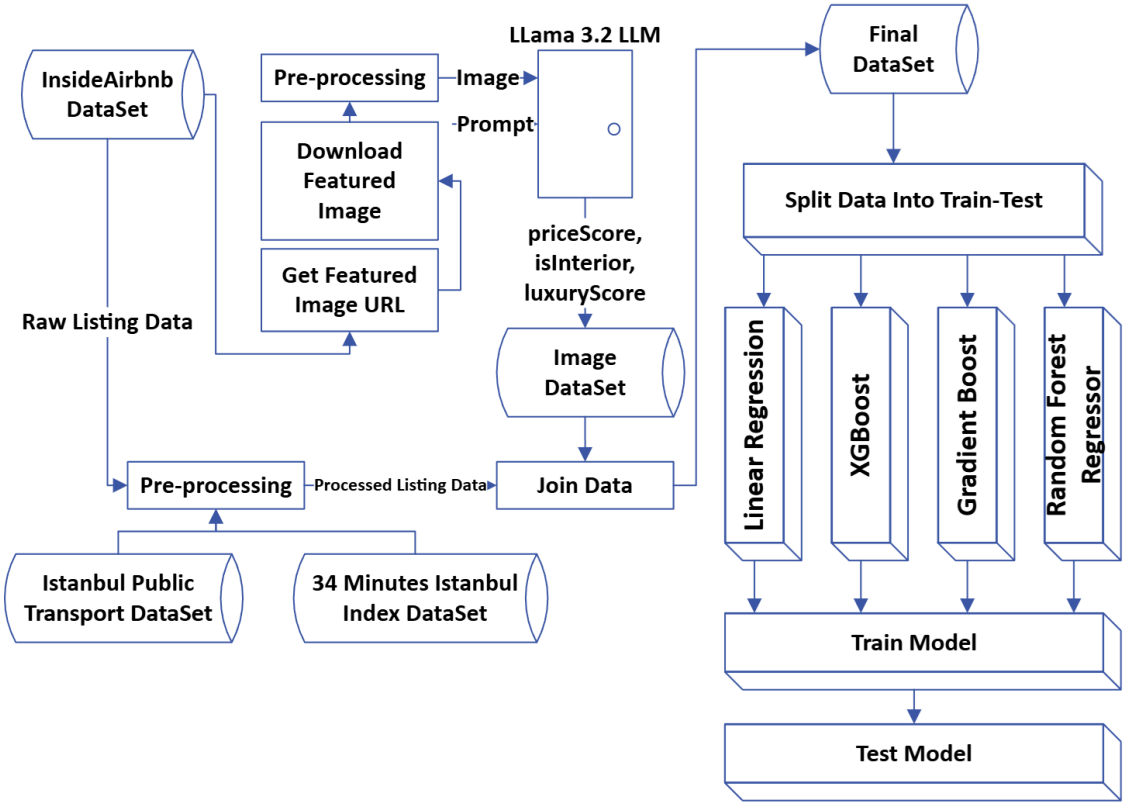

FIGURE 6.1 – The Proposed Methodology and Its Components

6.1 Model Training

The final dataset, after the pre-processing steps, is split into two categories, where 80% of the data is used for model training and 20% of the data is used for model testing. Furthermore, subsets of both train and test sets are created to train and test the models with different feature combinations of data.

Using seven different combinations of features and four different machine learning methods, a total of 28 models are created and tested.

TABLE 6.1 – Features Used for Evaluation

Model Name	**Features**
Base	Features obtained by InsideAirbnb listing data
Base&LocIndex	11 location indexes obtained from Istanbul Metropolitan Municipality added to base data
Base&Transport	Public transport proximity information added to base data
Base&Image	Image features obtained via LLM added to base data
Base&Amenity	One hot encoded amenities (where similar amenities are grouped) and daily amenity cost in addition to 'Base' model features
Base&Loc&Img	Base features enhanced with location based data and image features
Base&Loc&Img&Amenity	One hot encoded amenities (where similar amenities are grouped) and daily amenity cost in addition to 'Base&Loc&Img' features

6.2 Hyperparameter Tuning

The hyperparameter tuning step of the research was performed using Random Search, with R^2 as the score measure. Due to the number of models and possible hyperparameter combinations, Grid Search was not chosen. For each model, the hyperparameter grid values in the below tables are used.

TABLE 6.2 – Hyperparameter Grid for XGBoost

Parameter	**Value**
learning_rate	[0.1, 0.01, 0.001]
n_estimators	[1000, 3000, 5000]
max_depth	[1, 3, 7]

TABLE 6.3 – Hyperparameter Grid for Gradient Boost Regressor

Parameter	**Value**
learning_rate	[0.1, 0.01, 0.001]
n_estimators	[1000, 3000, 5000]
max_depth	[1, 3, 7]

Best parameters for XGBoost were :

- n_estimators : 5000
- max_depth : 7
- learning_rate : 0.01

Best parameters for Gradient Boost Regressor were :

- n_estimators : 3000

TABLE 6.4 – Hyperparameter Grid for Random Forest Regressor

Parameter	Value
n_estimators	[1000, 3000, 5000]
max_depth	[1, 3, 7]

— max_depth : 7

— learning_rate : 0.01

Best parameters for Random Forest Regressor were :

— n_estimators : 3000

— max_depth : 7

6.3 Hardware and Software

For the experiment, a computer with below hardware was used :

— Intel i7-13700K 16 core CPU

— 64GB DDR5 RAM

— NVIDIA RTX 4070 12GB VRAM GPU

The following software components were used for processing the data, using the LLM and training and running the models :

— Python 3.12

— Ollama, Langchain

— Pandas, Numpy

— Sklearn, Xgboost

— Matplotlib, Seaborn

7 EXPERIMENT RESULTS

Using four different regression methods and seven different sets of features, a total of 28 experiments are performed to evaluate the performance of different machine learning models and different variables.

The results show that for all sets of features, XGBoost performed the best. The base dataset, which contains only pre-processed InsideAirbnb data without one-hot encoded amenities and daily amenities cost, had an adjusted R^2 value of 0.62. Adding location-based index and public transportation proximity data increased this value to 0.642 and 0.654, respectively. The LLM image price scores had almost no effect on the performance of XGBoost. When the model is trained with all the features and amenities, it reached Adj. R^2 value of 0.678.

Gradient Boost followed XGBoost in performance, reaching maximum Adj. R^2 value of 0.593 when amenities and all features were used, compared to 0.527 with the base features. Linear Regression performed came third, and Random Forest Regressor was the worst performing model.

The results also show that the best increase in terms of percentage was achieved with Linear Regression.

TABLE 7.1 – Experiment Results

	Linear Regression			
	Adj R2	MAE	MSE	RMSE
Base	0.443	0.582	0.554	0.744
Base&LocationIndex	0.47	0.565	0.527	0.726
Base&PublicTransport	0.486	0.553	0.511	0.714
Base&Image	0.455	0.577	0.542	0.736
Base&Amenity	0.54	0.523	0.454	0.674
Base&Loc&Img	0.501	0.544	0.495	0.704
Base&Loc&Img&Amenity	0.57	0.501	0.425	0.652
	XGBoost			
	Adj R2	MAE	MSE	RMSE
Base	0.62	0.465	0.378	0.614
Base&LocationIndex	0.642	0.446	0.356	0.596
Base&PublicTransport	0.654	0.437	0.343	0.586
Base&Image	0.626	0.462	0.371	0.609
Base&Amenity	0.654	0.437	0.343	0.585
Base&Loc&Img	0.654	0.439	0.343	0.586
Base&Loc&Img&Amenity	**0.678**	0.418	0.318	0.564
	GradientBoostingRegressor			
	Adj R2	MAE	MSE	RMSE
Base	0.527	0.537	0.47	0.686
Base&LocationIndex	0.552	0.52	0.445	0.667
Base&PublicTransport	0.558	0.516	0.438	0.662
Base&Image	0.528	0.536	0.469	0.685
Base&Amenity	0.565	0.512	0.431	0.657
Base&Loc&Img	0.561	0.512	0.435	0.660
Base&Loc&Img&Amenity	0.593	0.490	0.402	0.635
	RandomForestRegressor			
	Adj R2	MAE	MSE	RMSE
Base	0.309	0.65	0.687	0.829
Base&LocationIndex	0.309	0.65	0.687	0.829
Base&PublicTransport	0.31	0.65	0.686	0.828
Base&Image	0.309	0.65	0.687	0.829
Base&Amenity	0.307	0.65	0.687	0.829
Base&Loc&Img	0.309	0.65	0.686	0.828
Base&Loc&Img&Amenity	0.306	0.65	0.686	0.828

7.1 Permutation Importance

Permutation importance is a method to calculate the importance of a feature to a model (Altmann et al., 2010). In this section, the permutation importance of features are calculated using XGBoost Regressor and is presented below under four different categories. The first category contains the preprocessed base features (where values very close to zero are omitted) that come with the InsideAirbnb dataset. The second category includes the different index features. The third category shows the public transportation proximity variables. The last category contains only the image-related variables.

For the base features (with amenity count for amenities only), Figure 7.1 (where low impact base features are removed) shows that accommodation is the most important feature that affects the fit of the model, followed by the host response rate, whether the listing is an entire home and whether the listing is a private room.

Location index features with highest effect on the R^2 value are Meeting Your Needs Index, Learning Index, and Work Index.

When it comes to proximity to public transportation stops, variables related to minibus stops are the leading features in their category.

Considering the permutation importance of the image features, only the luxury score had a negligible effect on the R^2 value, while other two features had importance close to zero.

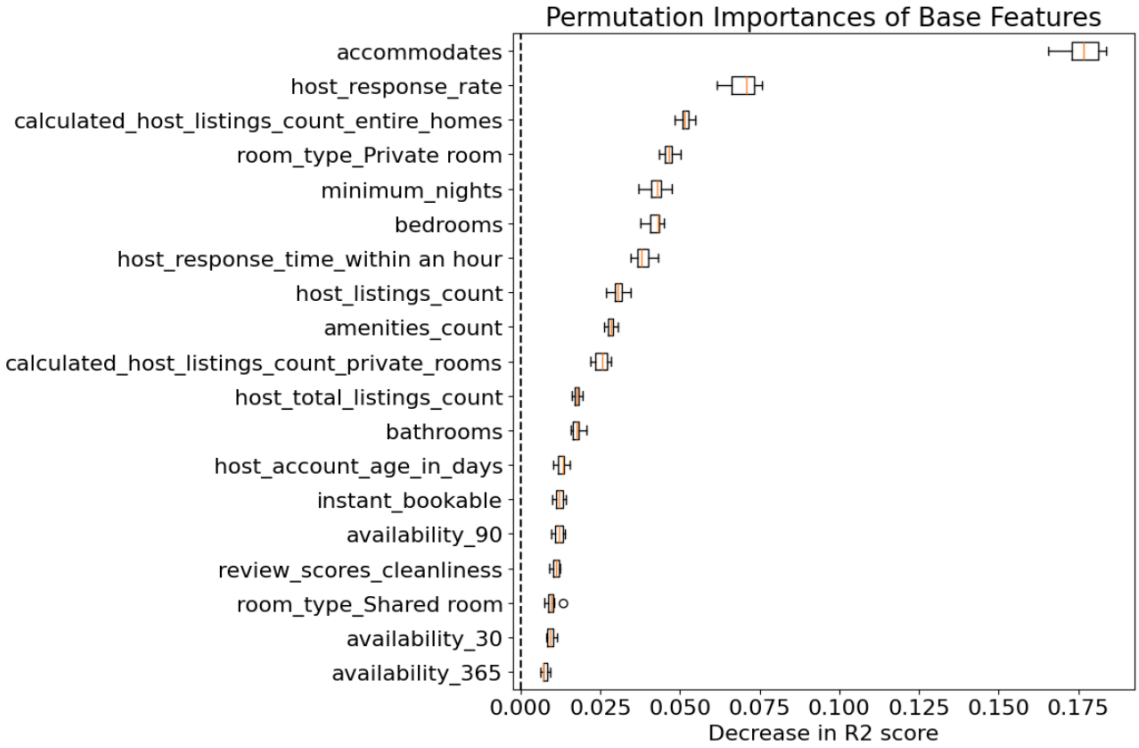

FIGURE 7.1 – Permutation Importance of Base Features

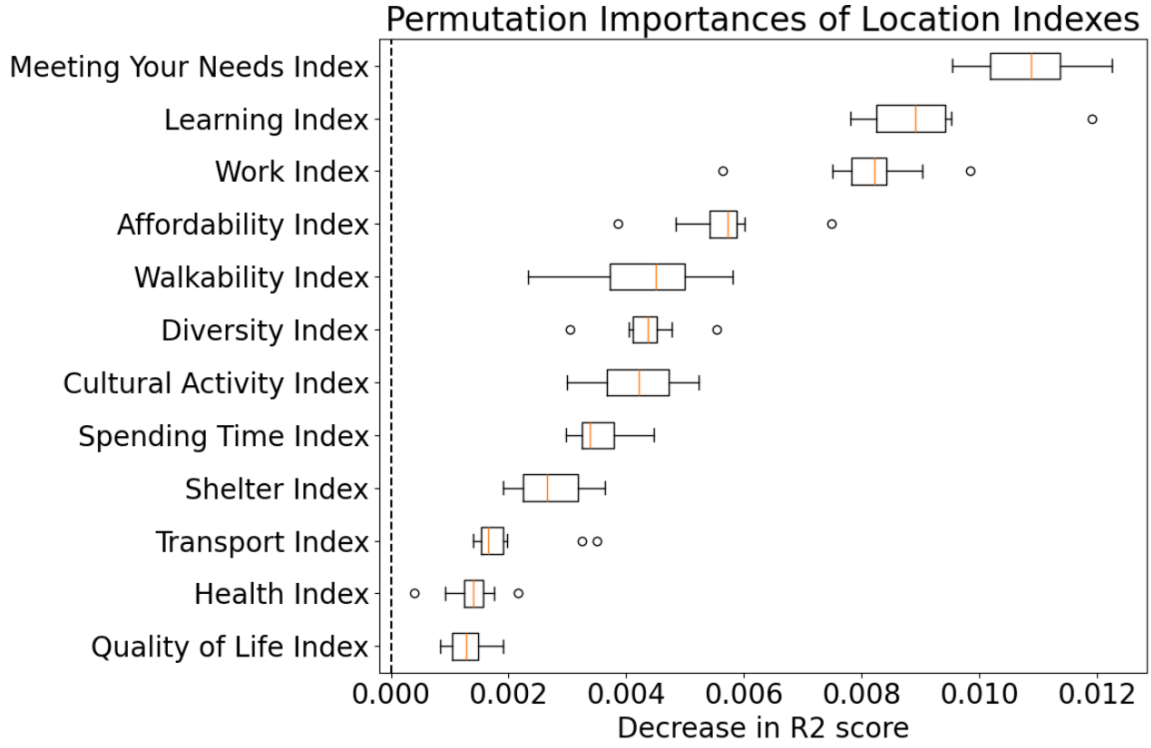

FIGURE 7.2 – Permutation Importance of Index Features

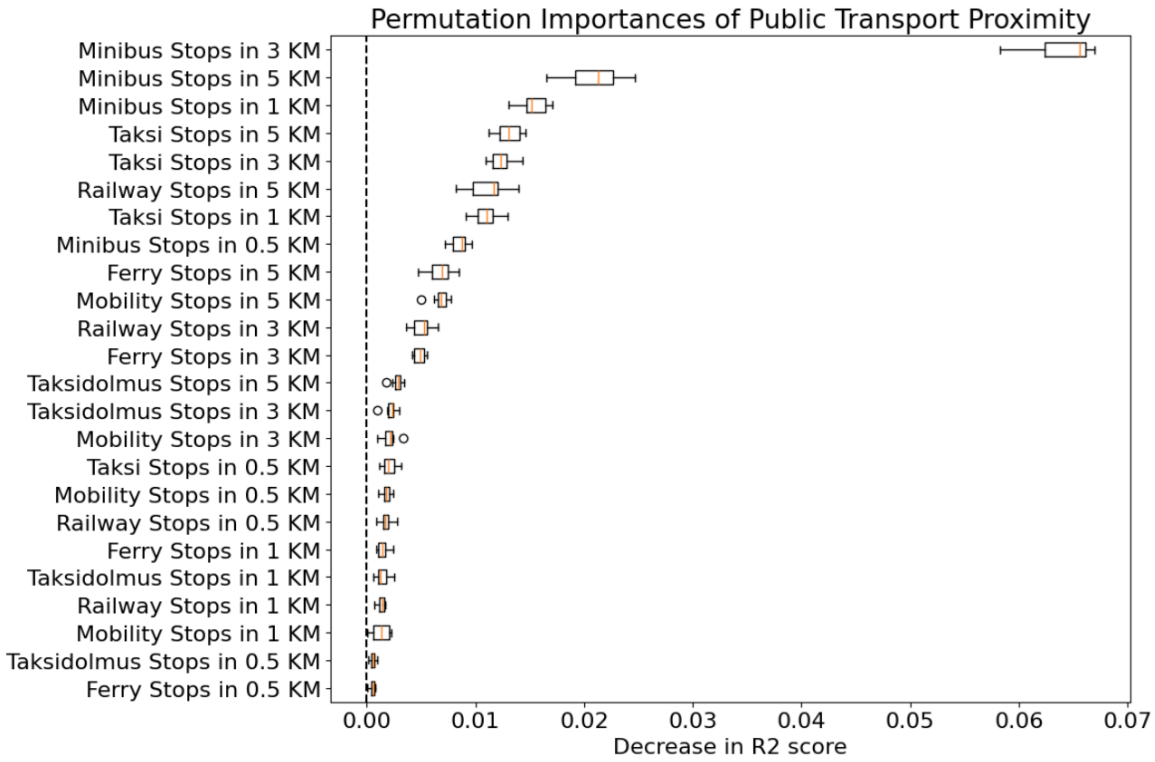

FIGURE 7.3 – Permutation Importance of Public Transportation Features

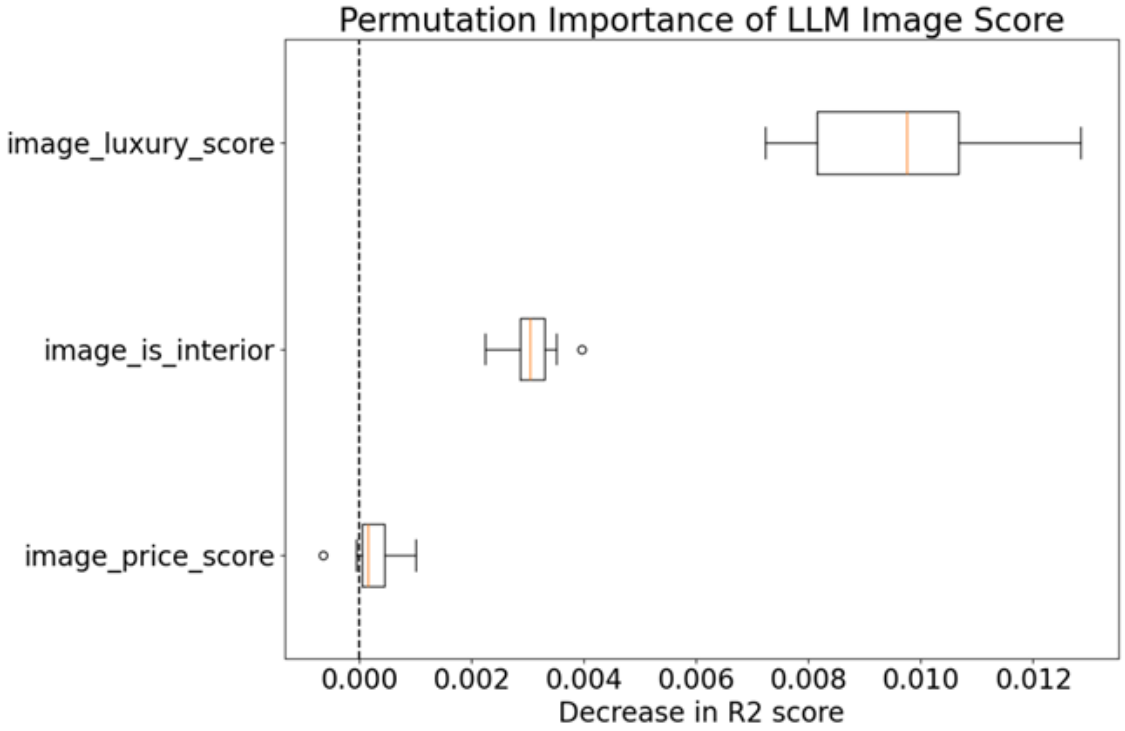

FIGURE 7.4 – Permutation Importance of Image Features

8 CONCLUSION & DISCUSSION

In this thesis, the objective was to answer the following research questions.

1- How accurately can machine learning models predict the prices of Airbnb listings ? To answer this question, four different machine learning models were trained using seven different datasets and different model evaluation metrics were used. It is observed that choosing the right machine learning model for the price prediction task makes the most difference, since the choice of model always had more impact on the fit of the model than added features. Similar to findings in other research mentioned in the Literature Review chapter, XGBoost was the best model for this task, followed by the Random Forest Regressor.

2- To what extent do location-based characteristics influence the price of an Airbnb listing ? To extend the base data obtained from InsideAirbnb, coordinates of the listings were used to add location-based information to each listing. These variables can be grouped into two categories. Firstly, different index values measuring how much the different daily needs can be met at that location were used. Secondly, how well the location is integrated with different public transport systems was calculated using the number of stops within a distance. The results show that the set of variables in both categories increased the performance of the models in a similar and noticeable way. The importance of each variable was calculated using permutation importance. Results showed that physical properties of a listing, such as the number of accommodates it can host, room type or amenities were always the most important price determinants. However, location-based features added in this research also had impact on model performance. Among the public transportation variables, the number of minibus and taxi stops around the Airbnb listing improved the fit of the model most. And among the different location indexes, Meeting Your Needs, Learning and Work indexes had the most contribution.

3- Can large language models effectively utilize the featured images of an Airbnb listing as input for price prediction models ? In this thesis, the possible contribution of using large language models was explored by asking the LLM to assign a price score, interior flag, and luxury score to the featured image of each Airbnb listing.

Although the reasoning made sense when sample outputs were manually checked, the scores assigned to all listings did not improve the performance of any of the machine learning models noticeably.

8.1 Threats To Validity

This thesis used Airbnb listing data in Istanbul and location-based information was added using publicly provided data by the Istanbul Metropolitan Municipality. Therefore, the scope of the work is limited to one city. Additionally, location-based information cannot be extended to other cities as similar data in a similar form is not provided by municipalities of other cities.

Moreover, the listing data obtained through InsideAirbnb is a snapshot of a certain date, and changes in price of individual listings cannot be tracked. The price of listings at a certain location might be different on the snapshot day than their averages due to seasonality or local events.

Finally, when measuring the cost of provided amenities, lifespans of amenities were assumed and actual lifespans of items may vary. Also, due to the generalization of variations in amenities, the exact cost of amenities might be different than the actual cost of amenities.

8.2 Impact of the Work

For location-based information, most of the existing work, as summarized in the literature review chapter, focuses on the distance of a listing to major tourist sites and airports. However, as Hill et al. (2023) states, there is a correlation between Airbnb listing prices and long-term residency rental prices. This shows that location-related price determinants should not be limited to proximity to tourist attractions, and price determinants that affect long-term rental prices should also be included when predicting Airbnb prices. In this thesis, 10 different indexes such as spending time, health, learning, and work are used to predict the prices, and their individual importance is measured. We believe that this will help increase the focus on inclusion and measurement of similar location-based price determinants.

Furthermore, in the existing literature, Airbnb listing images are less used when predicting prices. We showed, at least for Istanbul, that using the LLama 3.2 Large Language Model to create three different variables did not help with model performance. However, the impact of LLMs can be further studied using different Large Language Models, cities, or prompts.

REFERENCES

Altmann, A., Toloşi, L., Sander, O. and Lengauer, T. (2010). Permutation importance : a corrected feature importance measure, *Bioinformatics* **26**(10) : 1340–1347.

Alves, C. J. (2024). *Airbnb price prediction in lisbon and porto : A machine learning approach with multi-feature integration*, Master's thesis, Universidade NOVA de Lisboa (Portugal).

Benoit, K. (2011). Linear regression models with logarithmic transformations, *London School of Economics, London* **22**(1) : 23–36.

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

Camatti, N., di Tollo, G., Filograsso, G. and Ghilardi, S. (2024a). Predicting airbnb pricing : a comparative analysis of artificial intelligence and traditional approaches, *Computational Management Science* **21**(1) : 30.

Camatti, N., Di Tollo, G., Filograsso, G. and Ghilardi, S. (2024b). Predicting airbnb pricing : a comparative analysis of artificial intelligence and traditional approaches, *Computational Management Science* **21**(1) : 30.

Camilleri, Gabriella (2023). Modelling Airbnb Prices in the Maltese Islands. Student Paper.

Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae) ?–arguments against avoiding rmse in the literature, *Geoscientific model development* **7**(3) : 1247–1250.

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

Chica-Olmo, J., González-Morales, J. G. and Zafra-Gómez, J. L. (2020). Effects of location on airbnb apartment pricing in málaga, *Tourism Management* **77** : 103981.

Dai, W., Lee, N., Wang, B., Yang, Z., Liu, Z., Barker, J., Rintamaki, T., Shoeybi, M.,

Catanzaro, B. and Ping, W. (2024). Nvlm : Open frontier-class multimodal llms, *arXiv preprint arXiv :2409.11402* .

de Zarzà, I., de Curtò, J., Roig, G. and Calafate, C. T. (2023). Llm multimodal traffic accident forecasting, *Sensors* **23**(22) : 9225.

Ert, E. and Fleischer, A. (2019). The evolution of trust in airbnb : A case of home rental, *Annals of Tourism Research* **75** : 279–287.

Friedman, J. H. (2001). Greedy function approximation : a gradient boosting machine, *Annals of statistics* pp. 1189–1232.

Gangarapu, S. and Mernedi, V. S. A. (2023). Predicting airbnb prices in european cities using machine learning.

Garcia, S. G. (2023). *Evaluating the impact of image features on airbnb price predictions : A machine learning approach to hedonic pricing*, Master's thesis.

Guttentag, D. (2015). Airbnb : disruptive innovation and the rise of an informal tourism accommodation sector, *Current Issues in Tourism* **18**(12) : 1192–1217.

Hall, C. M. and Ram, Y. (2019). Measuring the relationship between tourism and walkability ? walk score and english tourist attractions, *Journal of Sustainable Tourism* **27**(2) : 223–240.

Hill, R. J., Pfeifer, N. and Steurer, M. (2023). The airbnb rent premium and the crowding-out of long-term rentals, *Journal of Housing Economics* **61** : 101935.

Hu, M., Lin, L., Liu, M. and Ma, S. (2024). Images' features and airbnb listing price : the mediation effect of visual aesthetic perception, *Tourism Review* **79**(5) : 1182–1195.

Huang, D., Yan, C., Li, Q. and Peng, X. (2024). From large language models to large multimodal models : A literature review, *Applied Sciences* **14**(12) : 5068.

Islam, M. D., Li, B., Islam, K. S., Ahasan, R., Mia, M. R. and Haque, M. E. (2022). Airbnb rental price modeling based on latent dirichlet allocation and mesf-xgboost composite model, *Machine Learning with Applications* **7** : 100208.

Jiang, Y., Zhang, H., Cao, X., Wei, G. and Yang, Y. (2023). How to better incorporate geographic variation in airbnb price modeling ?, *Tourism Economics* **29**(5) : 1181–1203.

Jin, S. T., Kong, H., Wu, R. and Sui, D. Z. (2018). Ridesourcing, the sharing economy, and the future of cities, *Cities* **76** : 96–104.

Lektorov, A., Abdelfattah, E. and Joshi, S. (2023). Airbnb rental price prediction using machine learning models, *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, Las Vegas, NV, USA, p. 0339–0344.

Lin, L., Hu, M., Li, H. and Liu, M. (2024). First impressions on sharing accommodation market platforms : The association between cover image type and property listing price, *International Journal of Hospitality Management* **123** : 103919.

Meijer, N. (2022). Improving airbnb listing price prediction with sentiment analysis, review recency and topic modelling.

Miles, J. (2005). R-squared, adjusted r-squared, *Encyclopedia of statistics in behavioral science* .

Oskam, J. and Boswijk, A. (2016). Airbnb : the future of networked hospitality businesses, *Journal of Tourism Futures* **2**(1) : 22–42.

Panagoulias, D. P., Virvou, M. and Tsihrintzis, G. A. (2024). Evaluating llm–generated multimodal diagnosis from medical images and symptom analysis, *arXiv preprint arXiv :2402.01730* .

Panahandeh, A., Rabiei-Dastjerdi, H., Goktas, P. and McArdle, G. (2025). Answering new urban questions : Using explainable ai-driven analysis to identify determinants of airbnb price in dublin, *Expert Systems with Applications* **260** : 125360.

Peng, N., Li, K. and Qin, Y. (2020). Leveraging multi-modality data to airbnb price prediction, *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*, pp. 1066–1071.

Sammut, C. and Webb, G. I. (2011). *Encyclopedia of machine learning*, Springer Science & Business Media.

Schwarzová, L. (2020). *Predicting Airbnb Prices with Neighborhood Characteristics : Machine Learning Approach*, PhD thesis, Tilburg University.

Shabrina, Z. and Morphet, R. (2022). Understanding patterns and competitions of short- and long-term rental markets : Evidence from london, *Transactions in GIS* **26**(7) : 2914–2931.

Tan, H., Su, T., Wu, X., Cheng, P. and Zheng, T. (2024). A sustainable rental price prediction model based on multimodal input and deep learning—evidence from airbnb, *Sustainability* **16**(15) : 6384.

Tang, E. and Sangani, K. (2015). Neighborhood and price prediction for san francisco airbnb listings, *Departments of Computer science, Psychology, economics–Stanford University* pp. 021–01.

Tang, J., Cheng, J. and Zhang, M. (2024). Forecasting airbnb prices through machine learning, *Managerial and Decision Economics* **45**(1) : 148–160.

Thackway, W. T., Ng, M. K. M., Lee, C.-L., Shi, V. and Pettit, C. J. (2022). Spatial variability of the 'airbnb effect' : A spatially explicit analysis of airbnb's impact on housing prices in sydney, *ISPRS International Journal of Geo-Information* **11**(1) : 65.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et al. (2023). Llama : Open and efficient foundation language models, *arXiv preprint arXiv :2302.13971* .

Wang, H. (2023). Predicting airbnb listing price with different models, *Highlights in Science, Engineering and Technology* **47** : 79–86.

Williams, C. (2023). *Cape town airbnb price prediction : an exploration of spatial statistic and machine learning methods*, Master's thesis, University of Cape Town.

Willmott, C. J. 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* **30**(1) : 79–82.

Wright, S. (1921). Correlation and causation, *Journal of agricultural research* **20**(7) : 557.

Wu, J., Tang, X., Yang, Z., Hao, K., Lai, L. and Liu, Y. (2025). An experimental evaluation of llm on image classification, *in* T. Chen, Y. Cao, Q. V. H. Nguyen and T. T. Nguyen (eds), *Databases Theory and Applications*, Springer Nature Singapore, Singapore, pp. 506–518.

Xie, K. and Mao, Z. (2017). The impacts of quality and quantity attributes of airbnb hosts on listing performance, *International Journal of Contemporary Hospitality Management* **29**(9) : 2240–2260.

Yang, Y. and Mao, Z. (2020). Location advantages of lodging properties : A comparison between hotels and airbnb units in an urban environment, *Annals of Tourism Research* **81** : 102861.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. et al. (2023). A survey of large language models, *arXiv preprint arXiv :2303.18223* .

Zhu, A., Li, R. and Xie, Z. (2020). Machine learning prediction of new york airbnb prices, *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, pp. 1–5.

APPENDIX A INSIDEAIRBNB

Table APPENDIX A.1 – inside airbnb listing data - Page 1

#	**Column**	**Non-Null Count**	**Dtype**
0	id	31758 non-null	int64
1	listing_url	31758 non-null	object
2	scrape_id	31758 non-null	int64
3	last_scraped	31758 non-null	object
4	source	31758 non-null	object
5	name	31758 non-null	object
6	description	30633 non-null	object
7	neighborhood_overview	11040 non-null	object
8	picture_url	31758 non-null	object
9	host_id	31758 non-null	int64
10	host_url	31758 non-null	object
11	host_name	31758 non-null	object
12	host_since	31758 non-null	object
13	host_location	20955 non-null	object
14	host_about	12817 non-null	object
15	host_response_time	27284 non-null	object
16	host_response_rate	27284 non-null	object
17	host_acceptance_rate	26738 non-null	object
18	host_is_superhost	31185 non-null	object
19	host_thumbnail_url	31758 non-null	object
20	host_picture_url	31758 non-null	object
21	host_neighbourhood	3823 non-null	object
22	host_listings_count	31758 non-null	int64
23	host_total_listings_count	31758 non-null	int64
24	host_verifications	31758 non-null	object
25	host_has_profile_pic	31758 non-null	object
26	host_identity_verified	31758 non-null	object
27	neighbourhood	11040 non-null	object
28	neighbourhood_cleansed	31758 non-null	object
29	neighbourhood_group_cleansed	0 non-null	float64
30	latitude	31758 non-null	float64
31	longitude	31758 non-null	float64
32	property_type	31758 non-null	object
33	room_type	31758 non-null	object
34	accommodates	31758 non-null	int64
35	bathrooms	29291 non-null	float64

Table APPENDIX A.2 – Inside Airbnb Listing Data - Page 2

#	**Column**	**Non-Null Count**	**Dtype**
36	bathrooms_text	31676 non-null	object
37	bedrooms	31438 non-null	float64
38	beds	29247 non-null	float64
39	amenities	31758 non-null	object
40	price	29232 non-null	object
41	minimum_nights	31758 non-null	int64
42	maximum_nights	31758 non-null	int64
43	minimum_minimum_nights	31757 non-null	float64
44	maximum_minimum_nights	31757 non-null	float64
45	minimum_maximum_nights	31757 non-null	float64
46	maximum_maximum_nights	31757 non-null	float64
47	minimum_nights_avg_ntm	31757 non-null	float64
48	maximum_nights_avg_ntm	31757 non-null	float64
49	calendar_updated	0 non-null	float64
50	has_availability	31020 non-null	object
51	availability_30	31758 non-null	int64
52	availability_60	31758 non-null	int64
53	availability_90	31758 non-null	int64
54	availability_365	31758 non-null	int64
55	calendar_last_scraped	31758 non-null	object
56	number_of_reviews	31758 non-null	int64
57	number_of_reviews_ltm	31758 non-null	int64
58	number_of_reviews_l30d	31758 non-null	int64
59	first_review	19402 non-null	object
60	last_review	19402 non-null	object
61	review_scores_rating	19402 non-null	float64
62	review_scores_accuracy	19377 non-null	float64
63	review_scores_cleanliness	19379 non-null	float64
64	review_scores_checkin	19375 non-null	float64
65	review_scores_communication	19379 non-null	float64
66	review_scores_location	19377 non-null	float64
67	review_scores_value	19376 non-null	float64
68	license	10338 non-null	object
69	instant_bookable	31758 non-null	object
70	calculated_host_listings_count	31758 non-null	int64
71	calculated_host_listings_count_entire_homes	31758 non-null	int64
72	calculated_host_listings_count_private_rooms	31758 non-null	int64
73	calculated_host_listings_count_shared_rooms	31758 non-null	int64
74	reviews_per_month	19402 non-null	float64

APPENDIX B LOCATION INDEXES

Table APPENDIX B.1 – 34 Minutes Istanbul Indexes

Index Name	**Data Type**
Shelter Index	float64
Affordability Index	float64
Spending Time Index	float64
Walkability Index	float64
Quality of Life Index	float64
Meeting Your Needs Index	float64
Work Index	float64
Diversity Index	float64
Cultural Activity Index	float64
Learning Index	float64
Health Index	float64
Transport Index	float64

APPENDIX C PUBLIC TRANSPORT

Table APPENDIX C.1 – public transportation columns

Public Transportation Column	Data Type
Mobility Stops in 0.5 KM	int64
Mobility Stops in 1 KM	int64
Mobility Stops in 3 KM	int64
Mobility Stops in 5 KM	int64
Ferry Stops in 0.5 KM	int64
Ferry Stops in 1 KM	int64
Ferry Stops in 3 KM	int64
Ferry Stops in 5 KM	int64
Taksi Stops in 0.5 KM	int64
Taksi Stops in 1 KM	int64
Taksi Stops in 3 KM	int64
Taksi Stops in 5 KM	int64
Taksidolmus Stops in 0.5 KM	int64
Taksidolmus Stops in 1 KM	int64
Taksidolmus Stops in 3 KM	int64
Taksidolmus Stops in 5 KM	int64
Minibus Stops in 0.5 KM	int64
Minibus Stops in 1 KM	int64
Minibus Stops in 3 KM	int64
Minibus Stops in 5 KM	int64
Railway Stops in 0.5 KM	int64
Railway Stops in 1 KM	int64
Railway Stops in 3 KM	int64
Railway Stops in 5 KM	int64

APPENDIX D DAILY AMENITY COST

Table APPENDIX D.1 – amenity sale prices and daily costs part I

Amenity	**Lifetime in Days**	**Sale Price**	**Daily Cost**
Booklet	1	100	100
Washer	1825	10700	5
Wifi	30	590	19
Dryer	1825	9600	5
HDTV	1825	17524	9
Alarm	1825	1500	0
Refrigerator	1825	12149	6
Iron	1825	849	0
Hangers	1825	189	0
Dishes and silverware	182	2250	12
Cooking basics	182	2659	14
Bed linens	182	679	3
Fire extinguisher	1460	607	0
Hot water kettle	1825	329	0
Extra pillows and blankets	182	1100	6
Stove	1825	1000	0
Cleaning products	7	700	100
Dining table	1825	1307	0
Room-darkening shades	1825	490	0
Drying rack for clothing	1825	485	0
Coffee maker	1825	1284	0
Oven	1825	8500	4
Wine glasses	182	195	1

Table APPENDIX D.2 – amenity sale prices and daily costs part II

Amenity	**Lifetime in Days**	**Sale Price**	**Daily Cost**
Crib	1825	749	0
Coffee	1	50	50
Microwave	1825	2749	1
Baking sheet	1	7	7
Mosquito net	1825	3300	1
Portable fans	1825	1240	0
Safe	1825	1020	0
High Chair	1825	599	0
Toaster	1825	759	0
Blender	1825	425	0
Outdoor furniture	1825	2939	1
Sound system	1825	1884	1
Outlet covers	1825	75	0
Bread maker	1825	3799	2
Ceiling fan	1825	794	0
Smart lock	1825	3384	1
Game console	1825	14493	7
Hammock	1825	165	0
Sun loungers	1825	3000	1
Changing table	1825	1791	0
Piano	1825	9000	4
Record player	1825	2391	1
Beach essentials	182	230	1
Baby monitor	1825	1049	0

BIOGRAPHICAL SKETCH

Education

— B.Sc. in Computer Engineering, Middle East Technical University (2014)

Experience

— Central Bank of the Republic of Türkiye (2020-Ongoing)

— General Electric Aviation (2019-2020)

— Anadolu Sigorta (2017-2019)

— Huawei (2016-2017)

— ISIS Information Technologies (2015-2016)

— Universidad Rey Juan Carlos (2014-2015)

PUBLICATIONS

— Ö. Akalın and G. I. Alptekin, "Enhancing Airbnb Price Predictions with Location-Based Data: A Case Study of Istanbul," 2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS), Belgrade, Serbia, 2024, pp. 207-212, doi: 10.15439/2024F7603.