ÇANKAYA UNIVERSITY
FACULTY OF ENGINEERING
COMPUTER ENGINEERING DEPARTMENT

Project Report

Version 1

CENG 408

Innovative System Design and Development II

202002
AIRBNB PRICE DETERMINANTS

201512019 Rana Gürcü Deliömeroğlu
201514039 Merve Karakaya
201627015 Melisa Yıldız
201611033 Selin Kara

Advisor: Roya Choupani

Table of Contents

Abstract

Airbnb is a worldwide house and property rental platform that provides short-term accommodation services to those who prefer houses to hotels. Airbnb is getting more popular day by day. The platform puts the responsibility of determining the price of the hosts. However, the prices are not determined randomly. As a matter of fact, many features such as location, availability of public transportation, size of the rented house, the existence of shopping areas in the neighborhood and many more play a role in determining the price. In this project, the real data from the official site of Airbnb will be used to determine which attributes are the most decisive ones in determining the price. The model to be developed with this project is to provide an auxiliary service for homeowners to estimate the prices more accurately. In order to work more efficiently within the scope of this project, we are conducting extensive searches and investigations on machine learning, deep learning and data mining, which are the applications of artificial intelligence.

Key words: Airbnb, Determining The Price, Machine Learning, Deep Learning, Data Mining, Artificial Intelligence.

Özet

Airbnb, evi otele tercih edenlere kısa süreli konaklama hizmetleri sunan dünya çapında bir ev ve mülk kiralama platformudur. Airbnb gün geçtikçe daha popüler hale gelmiştir. Platform, ev sahiplerinin fiyatını belirleme sorumluluğunu üstlenir. Ancak fiyatlar tesadüfi olarak belirlenmez, aslında fiyatın belirlenmesinde konum, toplu taşıma kullanılabilirliği, kiralanan evin büyüklüğü, mahallede alışveriş alanlarının varlığı ve daha birçoğu gibi farklı özellikler rol oynamaktadır. Bu projede Airbnb'nin resmi sitesinden alınan gerçek veriler, fiyatın belirlenmesinde hangi özelliklerin en belirleyici olduğunun belirlenmesinde kullanılacaktır. Bu proje ile geliştirilecek model, ev sahiplerinin fiyatı daha doğru tahmin etmeleri için yardımcı bir hizmet sağlayacaktır. Bu proje kapsamında daha verimli çalışabilmek için yapay zeka uygulamaları olan makine öğrenmesi, derin öğrenme ve veri madenciliği konularında kapsamlı araştırmalar ve çalışmalar yapılmıştır.

Anahtar Kelimeler: Airbnb, Fiyat Tahmini, Makine Öğrenmesi, Derin Öğrenme, Veri Madenciliği, Yapay Zeka.

1. Introduction

In recent years, people prefer to use platforms such as Airbnb, which determine the price according to the characteristics of the place they will stay, instead of fixed and generally high-priced hotels during their touristic trips and holidays. Although hosts can compare prices manually with other houses and they can determine a price for their house but this is not a healthy method. In order to get a more accurate result when determining the price, a prediction algorithm should be established according to the characteristics of the house. Features such as number of bedrooms, number of bathrooms, neighborhood, house type (entire house / entire apartment), if it is a room, it can be shared with another person or not are determined for more accurate price prediction. This article aims to develop an accurate price development model using machine learning, so that hosts can more easily price according to the features above.

1.1 Problem Statement

The Airbnb platform does not have a price determinants for homeowners based on the characteristics of their home. Through this project, we aim to provide a platform where homeowners can determine the prices of their homes. It sets an appropriate rental price using the properties found in the house. Such a platform has not been implemented before, but there are many resources on the subject.

1.2 Background or Related Work

In the past years, multiple price setting projects have been carried out for hosts using Airbnb. Various machine learning algorithms and models were used in all of these projects. Machine learning is an area where we can find a lot of work today, so we were able to examine many literature research that could be a reference to our project. In addition, the fact that Airbnb is a widely used application today has been effective. While we are researching, we saw that as the demand for Airbnb has increased, the number of projects done over the years has increased. These projects guided us at the idea stage.

1.3 Solution Statement

After the researches, we examined Regression and Classification models. In addition, we investigated machine learning algorithms which are Logistic Regression, K-Nearest Neighbors, Naive Bayes, Random Forests, Decision Tree etc. These models and algorithms will be used in our project. Some of them will be eliminated according to their efficiency, compatibility and our hardware resources.

1.4 Contribution

We will try to synthesize all our research and all the articles we read and transfer it to our own project. In order to get the best results in our own project, we separated the algorithms and models that worked best in previous projects among algorithms and models. We decided that regression was the best model for our project.

2. Literature Search

In December 2018, V. Raul Perez-Sanchez and his team published an article about this subject. In this paper, they adopt a hedonic price model, thanks to this model, they determine the relation between Airbnb accommodation attributes and price. The model uses OLS to estimate the relationship between features and pricing [1].

OLS model can be explained as follows:

$$\ln ln\ (P_i) = \alpha + \sum_{j=1}^{n} \beta_j X_{ij} + \sum_{k=1}^{m} \gamma_k + D_{ik} + \varepsilon_i \quad [1]$$

$\ln ln\ (P_i)$: The Neperian Logarithm of *daily price* ("i")

α: The fixed component

β_j: Estimated parameter which is related to characteristic "j"

X_{ij}: Continuous variable which considers the characteristic "j" of observation "i"

γ_k: Estimated parameter which is related to characteristic "k"

D_{ik} : The fictitious variable which considers the characteristic "k" of observation "i"

ε_i: Error associated with the observation "i". [1]

QRM overcomes some of the limitations of the OLS model, allowing the modeling of different sizes of the dependent variable.

QRM can be explained as follows:[1]

$$\gamma_i = X_i\beta_\theta + u\theta_i \quad [1]$$

Where

γ_i : The dependent variable
x_i : The matrix of independent variables
β_θ : The vector of parameters to be estimated for quantile θ
$u\theta_i$: The aleatory perturbation that corresponds to quantile . [1]

Studies based on the hedonic regression model in 2019 examined pricing determinants from an algorithmic perspective.

For Airbnb pricing studies, researchers have examined past studies and defined some explanatory variables. The researchers used the semilogarithmic hedonic price model to observe the effect of these explanatory variables on the price [2].

Hedonic price model was improved by Rosen (1974) in order to measure the marginal effects of the properties of a product on the price. The original hedonic price function is defined below:

P **:** Observed price of a product

Z **:** Vector of the product’s attributes or utilities

ε **:** Error Term

The semilog form of hedonic price function was used in the studies, which is indicated as:

ln P : Log form of the observed price of a sample

β_0 : Constant term

X_i : ith explanatory variable

Bt : Semi Elasticity of P with respect to Xi

ε : Error Term

As a result, it has been determined from the application of this model that some of the explanatory variables have a great effect on the price, and some do not. On the other hand, some variables on the price are expected to be clarified in future studies as their results are not reached. In this study covering the year 2020, a new approach type, Multiscale Geographically Weighted Regression (MGWR), was used in the regression model. The main idea in this model is to determine the Airbnb price by considering the spatial differences. Scales should have differences such as local and global depending on the spatial differences so that the pricing study can be done more accurately. According to MGWR, which is the model chosen to analyze in the study, a different scale should be applied to estimate the effect of each feature on the price, namely its coefficient. The mathematical approach of this model is as follows: The here shows the different scale process applied to each feature (j) [3].

Consequently of the application of this model, it has been observed that the scales applied to the features selected to examine the effect on pricing vary considerably according to the applied places, even if the same feature is applied.

2.1 Data Set and Features

2.1.1 Data Set

A data set is a structure that holds more than one data information in it. The data contained in the data set are taken from one or more databases. Each column in the table reflects the name of a data variable and each row reflects the data of that variable. In our project, we chose to use data sets for Airbnb listings in Istanbul. insideairbnb.com maintains data sets for users to access Airbnb data collectively. Airbnb data of each city is available on the site.

2.1.2 Features

Features are the main factor for price prediction. It should be informative to the user and relevant to the price. Because of that, we selected our features based on the important requirements for a homeowners. In order to the homeowners to get the most accurate price estimates, we determined the following features according to the their needs:

- *House Type:* Determining the type of house for pricing is the main step. Whether the house is an entire apartment, entire house changes the pricing.
- *Room Type:* After choosing the home type, the user should also indicate whether the room is shared or not, if it is a private/shared room, this criterion also affects the price.
- *Neighborhood:* Which neighborhood the house is located in is also an important factor. For example, the rental price of a house in Beşiktaş would be higher than in Avcılar.
- *Number of Bedrooms:* The number of bedrooms in the house should also be specified. The large number of bedrooms is a factor that increases the price.
- *Number of Bathrooms:* The number of bathrooms in the home is also an important factor in price determination. More than one bathroom increases the price.
- *Extra features:* The features here will be extracted from the dataset with machine learning algorithms and will be determined later. One of the extra features resulting from the model development during the implementation of the project *is Number of Beds*.

2.2 Model and Algorithms

Machine learning tries to estimate the output values as closely as possible according to the given inputs. It establishes relationships between data set. The algorithms it implements optimize processes to improve performance and develop artificial intelligence applications.

In Supervised Machine Learning, the machine learning algorithm is built on a data set according to the given inputs and a desired output. The algorithm makes prediction and the algorithm works until the closest result is reached. Classification and regression are models that work with controlled learning.

The classification model is used to create outputs and classify new observations with what algorithms learn from inputs. Data sets are used to determine target classes and to create conditions for boundaries. After the conditions are determined, the next step is to predict the target class, or output.

Regression model focuses on one dependent and more than one independent variable. Makes output estimates based on these variables. The general purpose of the regression model is to think like a mathematical equation and make a y estimate that gives results according to x variables.

We can easily use classification and regression models for a price determinants system. Algorithms we can use depending on the following methods:

2.2.1 Naive Bayes Algorithm

The Naive Bayes algorithm is one of the most efficient and most used algorithm in machine learning studies. Its logic is based on Bayes' theorem. The algorithm's attributes are independent and easy to implement. Big data entries can be scaled more easily in this algorithm.

2.2.2 Logistic Regression Algorithm

Logistic Regression algorithm is a very useful algorithm to understand the effect of multiple inputs on a single output. It focuses on binary classification. If the problem is multi-categorized, sequential logistic regression is applied. It is especially used in problems such as purchasing and credit analysis.

2.2.3 K Nearest Neighbor Algorithm (KNN)

The KNN algorithm predicts the probability that an input is a member of another set after the inputs are received and divided into sets. Data is set for the reference named k and other data compare with this reference.

2.2.4 Decision Tree Algorithm

In the Decision Tree algorithm, the tree sets rules for classifying data. Separates the inputs into homogeneous clusters according to their differences. Then, it divides the data until the leafs reach the maximum depth and repeats this process, selecting the one with the highest probability of accuracy.

2.2.5 Random Forest Algorithm

It is a classification and regression algorithm derived from Random Forest decision trees. This algorithm works effectively even with large data sets. The algorithm starts with a tree structure, after the first entry is entered, other entries are entered down. It is divided into small clusters according to the relationships among the entries. Additionally, it is a preferred model when the data are continuous values.

Furthermore, it is the algorithm chosen for the implementation part of the project.

3.Summary

3.1 Summary of Conceptual Solution

As a result of this study, we examined similar studies conducted in 2017, 2018, 2019 and 2020 for our Airbnb Price Determinants project. We grasped why, when and for what Airbnb, a Sharing economy business model, is used. We conclude that a price determinants system is required for these platform users. We have previously observed that regression and classification is the most used method in projects in the field of price determination with Airbnb. According to the researches, we examined Logistic Regression, K-Nearest Neighbors, Naive Bayes, Random Forest and Decision Tree algorithms. These models and algorithms can be used easily in a price determinants system. Because of that, we will use these algorithms and models in our project.

3.2 Technology Used

For the coding part of this project; Scikit-Learn, Pandas, NumPy etc. libraries used for machine learning in Python programming language. For Python we also used Anaconda's Jupyter Notebook editor. Additionally, Visual Studio Code and Flask used for the web part of the project.

4. Software Requirements Specification

4.1 Introduction

4.1.1 Purpose

The purpose of this document is to explain the Airbnb pricing platform. This platform will prepare an accurate pricing policy for Airbnb hosts. In this document, the necessary explanations and models will be presented in detail for a better understanding of the project. Additionally, in these models, the features of the users who can access the platform will be explained.

4.1.2 Scope of Project

Airbnb is a platform where homes can be shared for travelers who prefer short-term accommodation and home to hotel. Airbnb is different from other types of accommodation due to its structure. This is why Airbnb pricing strategy requires different approaches. Different methods are used in this pricing strategy. This situation differs from country to country, from city to city and even in regions within the same city.

There are many factors for pricing policy. When we look at the literature studies on this subject, different effects of the factors on pricing policies are observed. For example, some factors have a great effect on price determination, while some do not have a positive or negative effect. Some are still up for debate.

The aim of this project is to develop a necessary subsidiary platform for Airbnb that can help with the difficulties that arise in the price determination policy mentioned above. Therefore, this platform can be easily used by Airbnb hosts. Additionally, the data set used for model training in the project is the data of Airbnb records located in Istanbul, Turkey.

4.1.3 Glossary

Table 1: Glossary of SRS

Host	People sharing their homes on the Airbnb platform
SRS	Software Requirements Specification

4.1.4 References

[1] IEEE Recommended Practice for Software Design Description

[2] IEEE Recommended Practice for Software Requirements Specification

[3] IEEE Standard for Software Project Management Plans

4.1.5 Overview of Document

This document is arranged according to the IEEE Recommended Practice for Software Requirements Specifications [2].

Second part contains a general description of the project. Also, product perspective, user characteristic, constraints, assumptions and dependencies, and apportioning of requirements sections are included.

In the third part, there are external interface requirements, functional requirements, performance requirements, software system attributes and design constraints.

4.2 Overall Description

In this section, the main factors and requirements affecting the project will be described. In order to understand these requirements easily, this section of the SRS will provide the necessary background. More detailed definitions will be included in the third section of the SRS.

4.2.1 Product Perspective

First of all, the necessary data will be obtained from Airbnb's official site for the Airbnb price determination platform that we will create. Later, this data will be classified according to price determination factors and an auxiliary platform will be developed that will determine the price through an appropriate model.

4.2.2 Development Methodology

While developing our project, we tried to use scrum, an agile software development methodology. Basically, our work in this direction is compatible with the Waterfall model, which is one of the software development models. For this, we first conducted a literature review. Then, we will carry out the Software Requirements Specification and Software Design Document processes that provide the documentation required for the development of the project. After these stages, we will make the project a usable platform, so that the operation of the project will proceed more systematically.

Thanks to the regular meetings held every week with our team and our advisor, we draw up plans and ensure the systematic progress of our project by taking the deadline into consideration.

4.2.3 User Characteristics

4.2.3.1 Host

- This type of user is the person who will benefit from the platform.
- Host will be able to filter according to the price determination features on the platform.
- Host will be able to do price research in accordance with price determination features.
- Host will be able to get the most optimized Airbnb price.

4.2.3.2 Admin

- Admin will be able to enter price determination features on the platform.
- Admin will be able to access the dataset and model.
- The admin user type will be able to change the interface of the platform.

4.2.4 Constraints

- There are two types of users on the platform, host and admin. Host is the person who can use this platform and the admin is the person who can manage the platform.
- The types of housing where this platform can be used are entire house, entire apartment, or just a room. When the single room option is selected among the housing types, the issue of whether it can be shared or private.
- There are also restrictions on price determination, restrictions are information of neighborhood, the number of bathrooms and bedrooms, and extra features such as number of beds in the house, entire villa and private room in apartment for house type, entire home/apt.

4.2.5 Assumptions and Dependencies

Any user with the appropriate hardware and software specified in the requirements specification which is the third section can use the platform.

4.2.6 Apportioning of Requirements

The dataset used by the Airbnb price determination platform we have created is dynamic and can be updated. Meanly, the platform can be a source of inspiration for future studies.

4.3 Requirements Specification

4.3.1 External Interface Requirements

4.3.1.1 User Interfaces

- This system will be able to work actively on all platforms with python 3.6 installed. In the interface of this system, which has two types of users, the host who is the first user, can select the house type such as an entire house, entire apartment, or room, if house-type is room, can also select the room type. Besides these, the host can select the neighborhood in which the house is in, the number of bathrooms and bedrooms, and extra features such as number of beds in the house, entire villa and private room in apartment for house type, entire home/apt. Admins who are the second user can edit the system.

4.3.1.2 Hardware Interfaces

The computer to be used for this system should be on a minimum 32 bit based. Also, this system is usable for all operating systems.

4.3.1.3 Software Interfaces

The computer to be used for this system must contain Anaconda for Python and Visual Studio Code.

4.3.1.4 Communications Interfaces

Internet connection is required to operate the system.

4.3.2 Functional Requirements

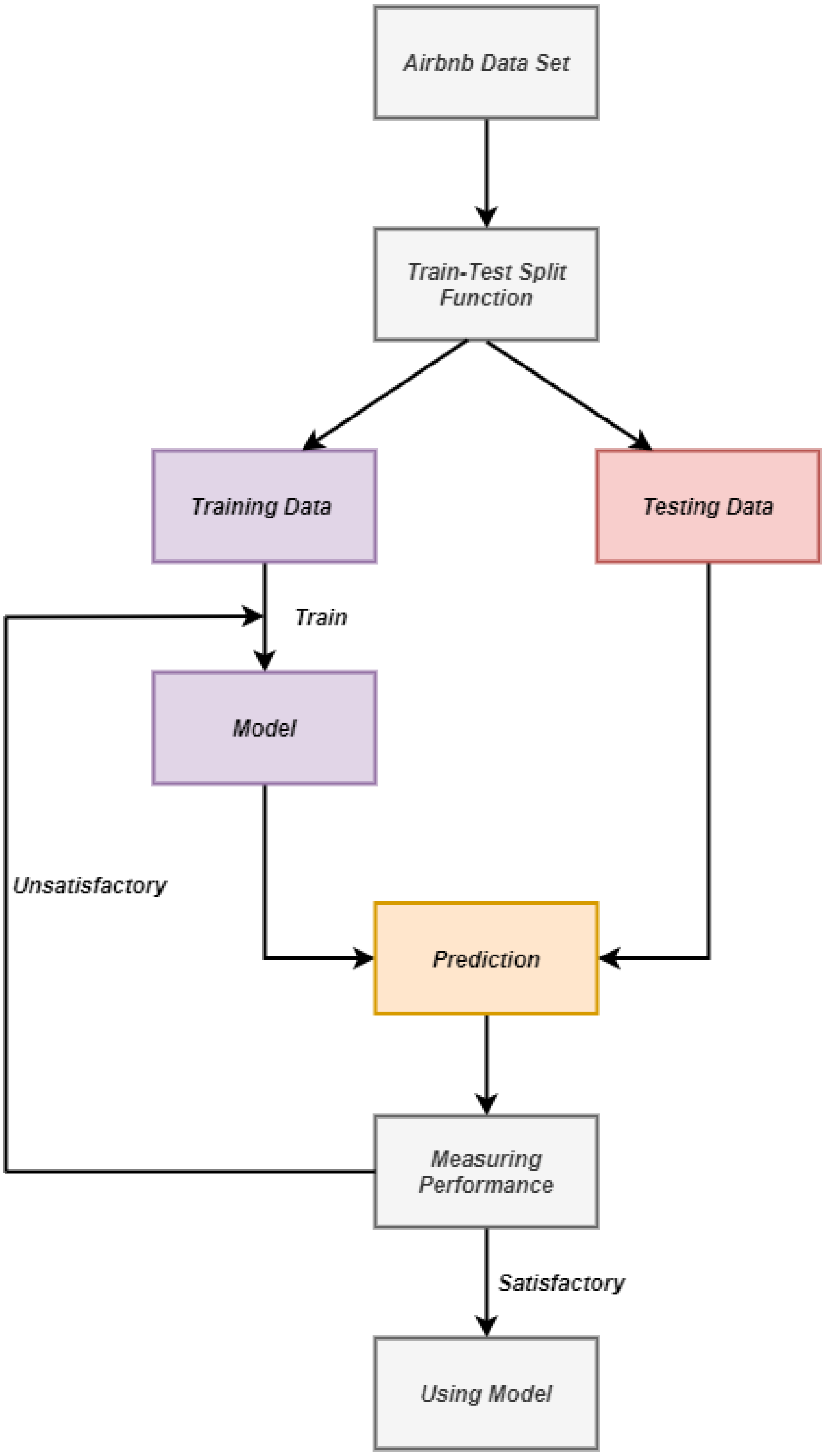

Figure 1: System Flow Diagram

The system flow diagram of the machine learning model that will work for price determination on this platform is as in Figure 1. Accordingly, data from Airbnb is primarily divided into train and test. It then uses the data reserved for training to develop the model. Test data are used to observe the prediction performance of this model. The model can be used if it gives satisfactory estimates, and if it does not, the model should be developed again.

4.3.2.1 Profile Management Use Case

Use Case:

- Start
- Edit
- Exit

Diagram:

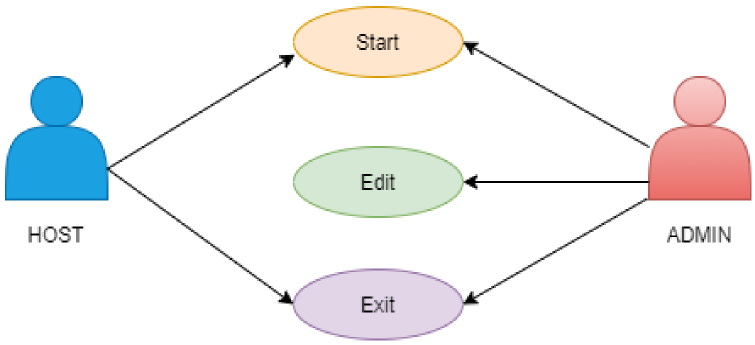

Figure 2: Profile Management Use Case

The profile management diagram describes the basic operations for host and admin that they will use in the system. It uses common exit operation as host and admin. The host and admin can enter the system without login. Also, admin has edit authority.

Initial Step-by-Step Description:

- Host can enter the system without Login.
- Admin can enter the system without Login.
- Admin can edit the system.
- Admin and Host can exit the system unconditionally.

4.3.2.2 Host Use Case

Use Case:

- Start
- Edit
- Exit

Diagram:

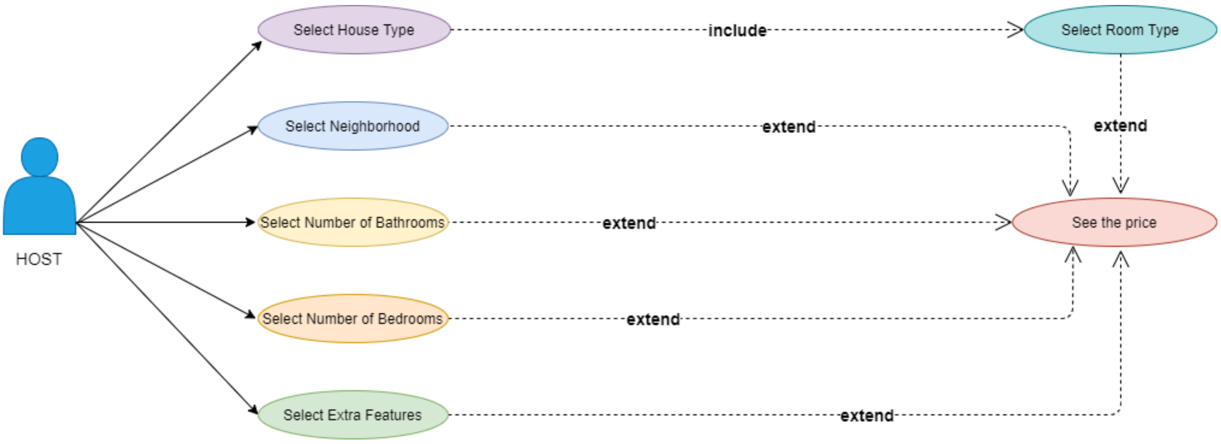

Figure 3: Host Use Case

The Host Use Case Diagram explains the operations that hosts can do. First of all, the hosts should reach this created platform via any internet browser. Hosts should select the house type, room type, neighborhood, number of bathrooms, number of bedrooms, and extra features such as number of beds, entire villa and private room in apartment for house type, entire home/apt to get an estimate of the price on my site. These selecting operations in the system start with the house type. After selecting the house type, if it is a room, the room type should also be selected. After the house type, the neighborhood where the house is located, the number of bathrooms, bedrooms and beds in the house should be selected. After entering this information, the host will see the price, which is a numerical value that the Airbnb pricing platform will give it.

According to the researches in the literature, Airbnb price determinants vary considerably according to many situations. However, the information entered by the host in the use case for this platform is chosen in this way especially because it is the basic price determination criteria that can fit many situations.

Initial Step-by-Step Description:

- Host should select their house type. There are four options for this, an entire house, an entire apartment, entire villa and private room in apartment.
- If the Host selects the room, they should choose their room type. The room which is selected can be a shared room, private room or entire home/apt.
- The Host should select the neighborhood of the house. Options that are listed here will be include string values which are neighborhoods in İstanbul.
- The Host should select the number of bathrooms in the house. Options that are listed will be integer values.
- The Host should select the number of bedrooms in the house. Options that are listed will be integer values.
- The Host should select the number of beds in the house. Options that are listed will be integer values.
- After selecting this information, they can see their estimated house rent price value.

4.3.3 Performance Requirements

Airbnb price determinants platform is a web application. There is no need for a specific system to run this application, minimum conditions will be sufficient. For example, the processor must be at least an Intel® Core ™ i3 processor or AMD Phenom X4. The operating system used must be Linux, macOS, or Windows.

4.3.4 Design Constraints

In addition to the constraints section, this platform can be used by all web browsers. The data to be used for the price prediction platform will be accurate data from Airbnb's official site. It can be used by all operating systems.

4.3.5 Software System Attributes

4.3.5.1 Portability

Data for this web application will be pulled from Airbnb via Python libraries. The Anaconda resource containing IDEs that will allow us to easily develop Python will be used. Also, Flask will be used for the web development part.

4.3.5.2 Maintainability

Since Airbnb is a platform that is constantly open to new data entry, the dataset that this application will use is updated depending on the data on Airbnb. There will be no change in the interface of the platform that will fundamentally affect the flow of the project, but the data used for price determination in the background should constantly change.

4.3.5.3 Usability

When using this application, the host should select the house type, room type (if the house type is a single room), neighborhood, number of bathrooms, number of bedrooms, number of beds to determine the price of the house.

4.3.5.4 Adaptability

This platform works in integration with data on Airbnb. Therefore, as dataset on Airbnb may change over time, the prices set by the platform may change accordingly.

4.3.5.5 Scalability

The criteria used by the system for the price determinants are sufficient for now, but these criteria can be increased according to the demands of the users.

5. Software Design Document

5.1 Introduction

5.1.1 Purpose

The purpose of the software design description document is to make the explanations of the steps to design the software required for a system and the details such as what will be built and how to design these steps.

5.1.2 Scope

The scope of the software design description document is system oriented. It allows the information and documents required for this system to be gathered, edited and changed when necessary. In short, it helps to create the basics of the software to be developed before the system becomes a product with explanations and visuals.

5.1.3 Glossary

Table 2: Glossary of SDD

Host	People sharing their homes on the Airbnb platform
SRS	Software Requirements Specification
SDD	Software Design Description
ER Diagram	Entity Relationship Diagram

5.1.4 References

[1] IEEE Recommended Practice for Software Design Description

[2] IEEE Recommended Practice for Software Requirements Specification

[3] https://wireframe.cc

[4] https://www.draw.io

5.1.5 Overview of Document

The overview of the software design description document will be as follows:

Introduction, design explanations, necessary software and hardware architectures, interfaces and designs.

5.2 Design Consideration

This section contains the terms and concepts required for design in the context of software design description. Accordingly, this part of the document contains approaches, tools, assumptions, dependencies and restrictions.

5.2.1 Approach

Our basic approach is to show in detail the patterns, structural styles, frame templates required for design.

5.2.2 Tools Used

The data required for this web platform will be extracted from Airbnb's official site. This data will be processed with python. Anaconda source will be used for this. In addition, Flask which is the web framework of python will be used for the web part of the project.

5.2.3 Constraints

The restrictions for the system are available. There are two types of users. One of them, the host, will use this platform for price estimation and admin is the person who develops and manages the platform. Housing types are entire house, entire apartment, entire villa and private room in apartment. If there is only one room, it should be stated that this room will be shared, private or entire home/apt. There are also restrictions on the features that determine the price. These features are selected as follows:

Selecting number of bathrooms, bedrooms and beds.

5.2.4 Assumptions and Dependencies

There is no need for a special login to use the platform. Users with basic software and hardware requirements can access and use this platform.

5.3 Architecture Design

5.3.1 Software Architecture

Airbnb Price Determinants platform has a data-based software architecture. Thanks to this database, meanly data set, users can take advantage of the platform and access the value corresponding to the price up to date. The representation of the software architecture is as follows:

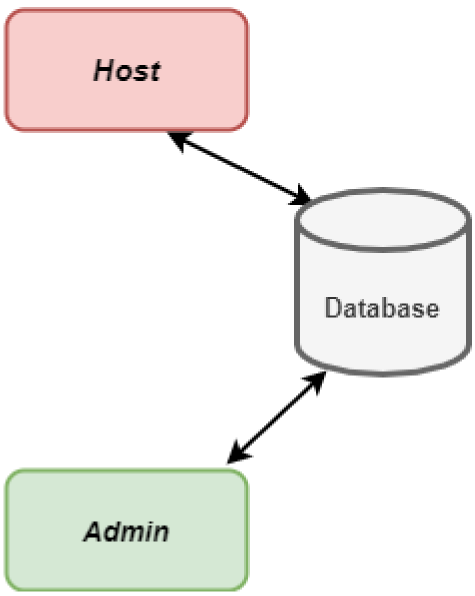

Figure 4: Software Architecture of Platform

5.3.2 Hardware Architecture

A user layer, a server layer and a database layer will be needed for this platform because users need a user layer to access the platform, while a server layer is required to use the platform. Also, a database will be required to keep the platform up to date. In short, although there is no restriction in the user-side, the access to the server-side is limited to the admin type user.

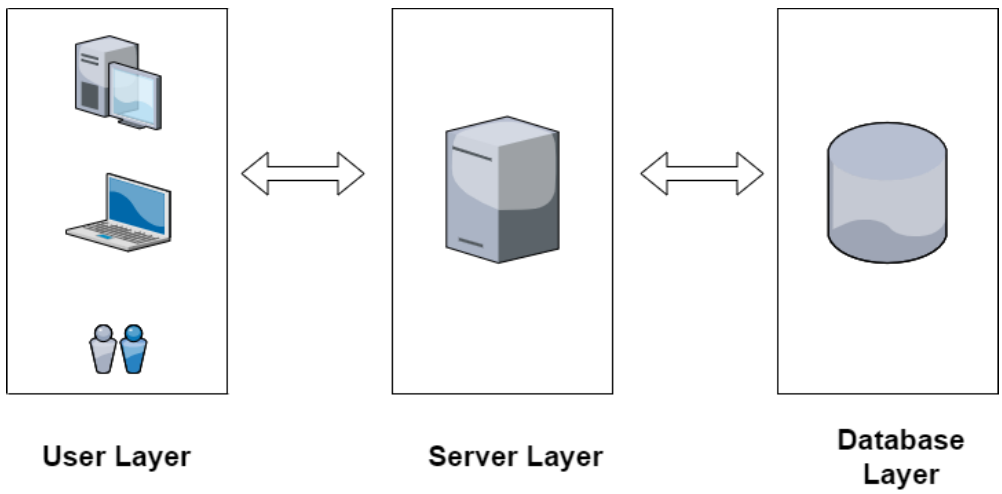

Figure 5: Hardware Architecture of Platform

5.4 System Interfaces

5.4.1 Software Interfaces

There is no need for a specially defined operating system for the computer to use this platform.

5.4.2 Hardware Interfaces

There is no need for a specially defined system to run this platform. This platform can be easily run on any computer with basic requirements.

5.5 User Interface Design

The interface perspective provides a visual design of the product to be created for designers, programmers and testers and provides a better understanding. The interface design presented in this part of the document is designed in accordance with the use cases and scenarios in the SRS document.

5.5.1 Home Page

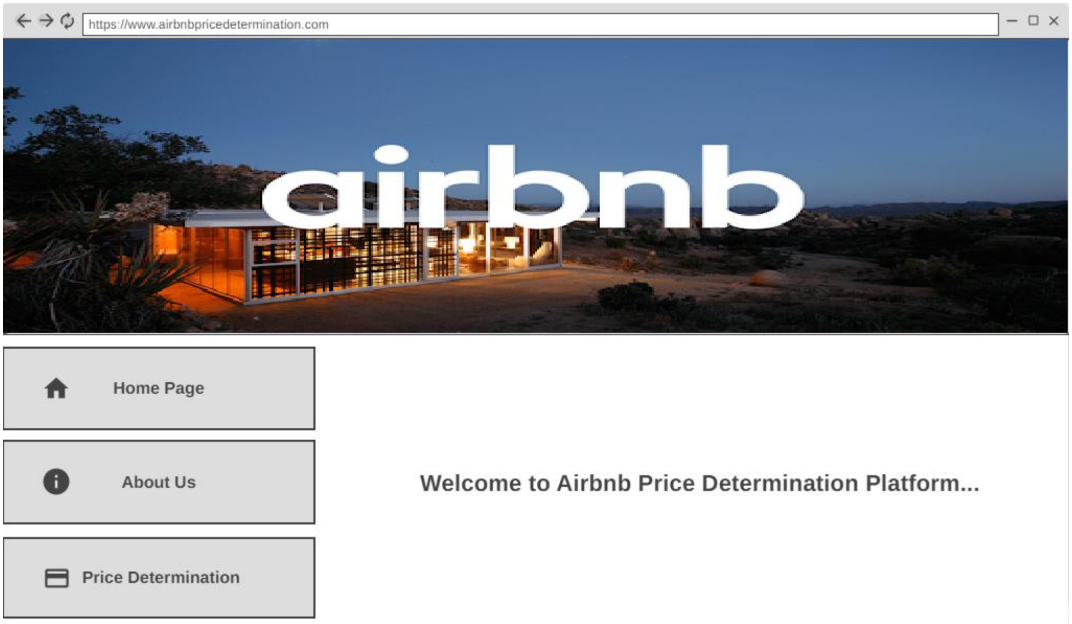

Figure 6: Home Page

5.5.2 About Us Page

When the Airbnb Price Determination platform is opened, this page is accessed when the *About Us* button is clicked on the main page.

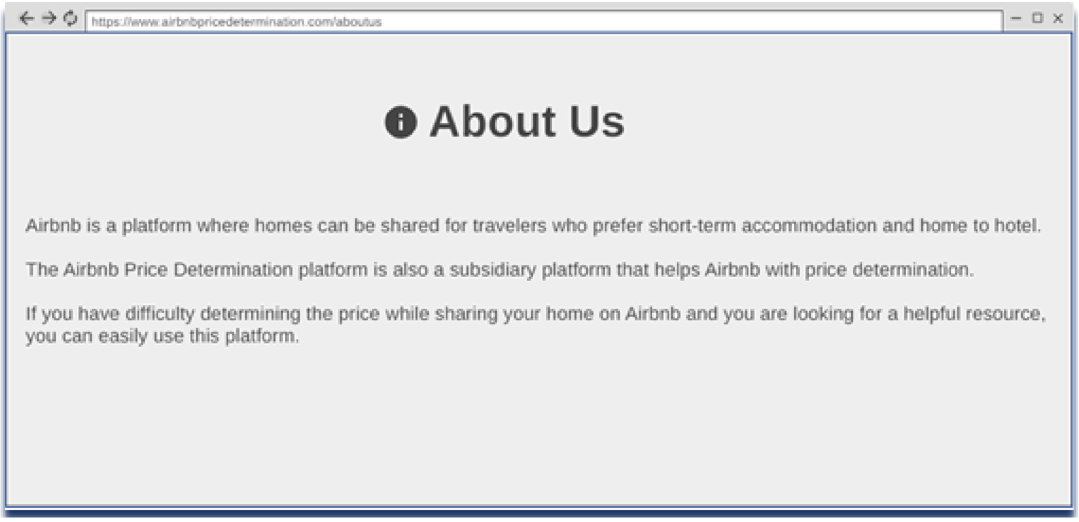

Figure 7: About Us Page

5.5.3 Price Determination Page

When the Airbnb Price Determination platform is opened, this page is accessed when the *Price Determination* button on the main page is clicked. On this page, there are features that will determine the price. The user must fill in these features according to the property to be placed on Airbnb. Features such as house type, room type, neighborhood, number of bathrooms and bedrooms are general price determination criteria for Airbnb. However, there will be additions according to the features that will appear as a result of the model to be used in the section specified as extra features.

Figure 8: Price Determination Page

Figure 9: Price Determination Page

5.5.4 Estimated Price

After the features on the price determination page are filled, the estimated price value will appear on this page by clicking the *See Price* button.

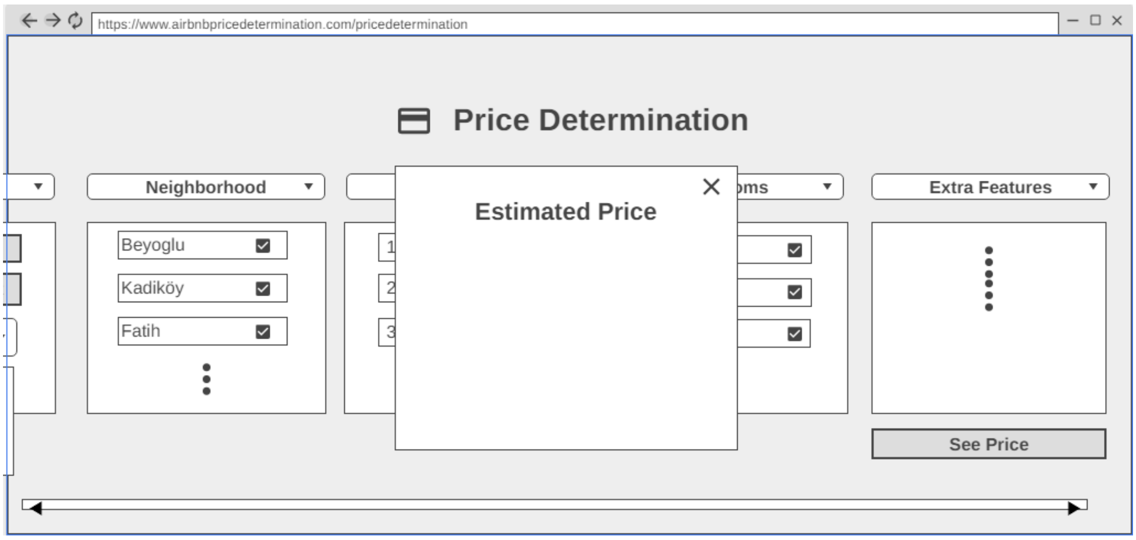

Figure 10: See The Price

5.6 Process Design

5.6.1 Use Cases

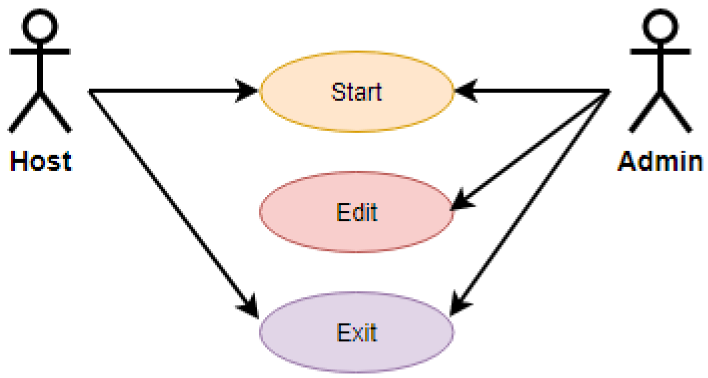

Figure 11: Profile Management Use Case

Admin: This type of user can develop and manage the platform.

Host: This type of user is people who share their homes on the Airbnb platform. Meanly, he/she is the owner.

Table 3: Actions for Admin User Type

Actor	Action	
Admin	Start	Indicates access to the platform
	Edit	It is the act of performing operations such as the database of the platform and interface design
	Exit	Indicates leaving the platform

Table 4: Actions for Host User Type

Actor	Action	
Host	Start	Indicates access to the platform
	Exit	Indicates leaving the platform

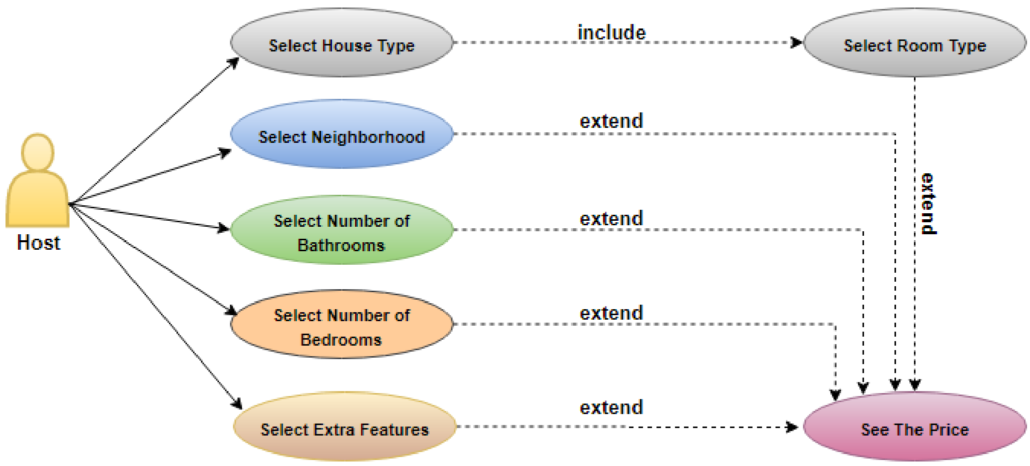

Figure 12: Host Use Case

Table 5: Explanation of Host Use Case

Actor	Action	
Host	Select House Type	There are two types of house types, entire house and entire apartment
	Select Room Type	There are two types of room types, shared and private.
	Select Neighborhood	It specifies the selection from the list of neighborhood according to the location of Airbnb
	Select Number of Bathrooms	Indicates the selection of the number of bathrooms in house types
	Select Number of Bedrooms	Indicates the selection of the number of bedrooms in house types
	Select Extra Features	It is the selection of the criteria of the price determination features to be added according to the result of the model developed according to the data.

5.6.2 Sequence Diagram

In this section, the actions that the host user can do are specified using the diagram. Accordingly, the host must first access the platform, then open the price determination page on the platform and select the features in the appropriate way. Finally, when you click the see price button, it should be able to access the estimated price.

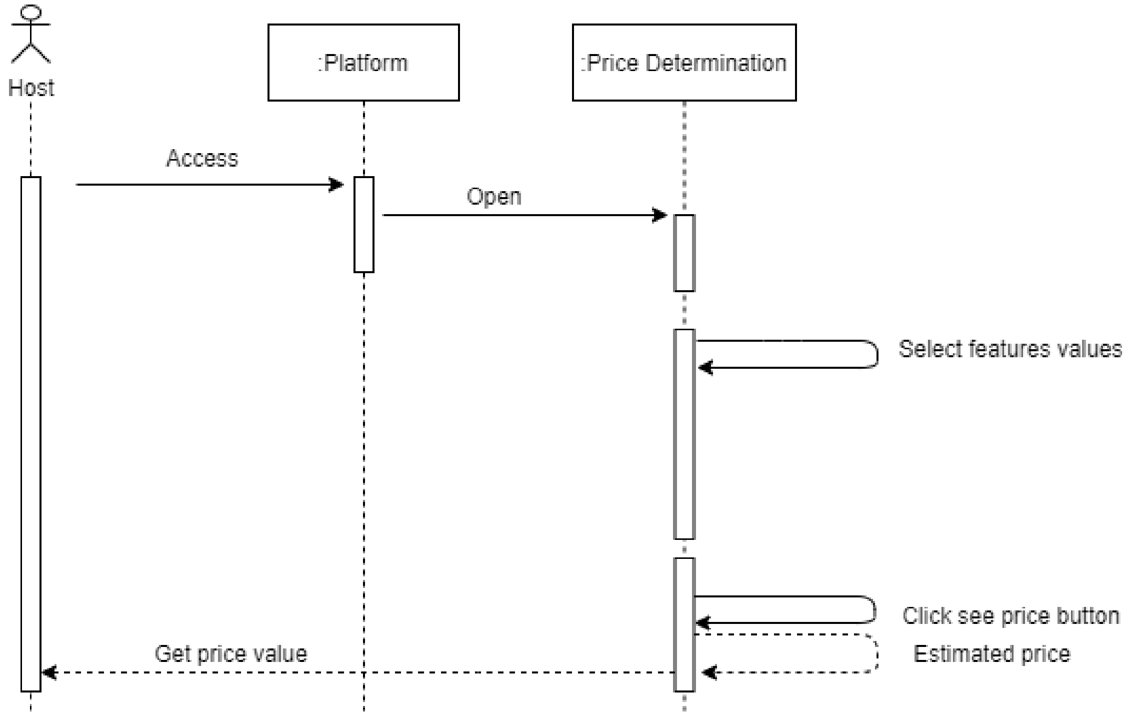

Figure 13: Sequence Diagram of Airbnb Price Determination Platform

5.7 Database Design

5.7.1 ER Diagram

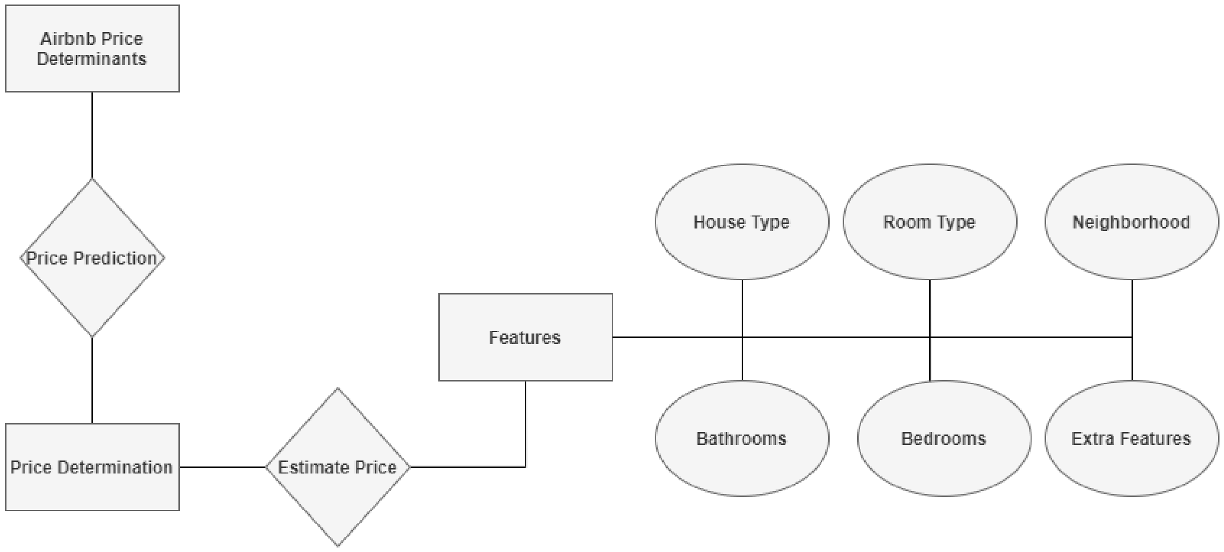

Figure 14: Entity Relationship Diagram

6. Test Plan

6.1 Introduction

6.1.1 Version Control

Table 6: Version Control Table

Version No	Description of Changes	Date
1.0	First Version	March 22, 2021
1.1	Second Version	May 28, 2021
1.2	Third Version	May 29, 2021

6.1.2 Overview

This document is for testing the working product of Airbnb Price Determinants application functionality. Detailed explanations for these tests will be shared in the following sections of the document. These sections are generally as follows:

- Features To Be Tested
- Features Not To Be Tested
- Item Pass/Fail Criteria
- Test Design Specifications
- Detailed Test Cases

6.1.3 Scope

This test plan only covers features specified in the SRS and SDD documentation created for the Airbnb Price Determinants platform. Therefore, all test scenarios will be covered in detail in this document.

6.1.4 Terminology

Table 7: Terminology Table

Acronym	Definition
SRS	Software Requirements Specification
SDD	Software Design Description

6.2 Features To Be Tested

In this project, the price determination page, which is the main purpose of the application, should be tested in detail.

6.2.1 Price Determination

In order to get the price on this page correctly, the combo boxes and buttons with the necessary features should be controlled. The scenarios that may occur for these controls are discussed in the test cases section.

6.3 Features Not To Be Tested

There is no need for detailed testing of sections other than the price determination page in the application. Because these pages do not directly affect pricing, they only provide the general flow of the platform. These sections are as follows:

- Home Page
- About Us

6.4 Items Pass/Fail Criteria

If the model behind the project works correctly; If it gives the outputs specified in SRS and SDD, the test will be passed successfully. But if this does not work according to SRS and SDD, the test will fail.

6.4.1 Exit Criteria

The conditions required for the platform to be considered successful or unsuccessful during the test phase will be as follows:

- 100% of test scenarios must be executed
- Must pass 95% of test scenarios
- All High and Medium Priority test cases are deemed passed
- All low priority test cases are deemed unsuccessful

6.5 References

[1] CENG407_202002_SRS_V1.0, November 05, 2020

[2] CENG407_202002_SDD_V1.0, December 25, 2020

6.6 Test Design Specifications

Test design specifications for the sections specified in features to be tested will be written as follows.

6.6.1 Price Determination (PD)

6.6.1.1 Subfeatures To Be Tested

6.6.1.1.1 House Type (PD.HT)

With this subfeature, the users can select that the room where they will welcome their guests is the entire house or the entire apartment to get the estimated price.

6.6.1.1.2 Room Type (PD.RT)

With this subfeature, the users select whether the room to accommodate their guests is a shared room or a private room to get the estimated price.

6.6.1.1.3 Neighborhood (PD.ND)

With this subfeature, the user can select which neighborhood of Istanbul the house is located in to get the estimated price.

6.6.1.1.4 Bathrooms (PD.BA)

With this subfeature, the users can select how many bathrooms they have in their residence to get the estimated price.

6.6.1.1.5 Bedrooms (PD.BR)

With this subfeature, the users can select how many bedrooms they have in their residence to get the estimated price.

6.6.1.1.6 Beds (PD.BE)

With this subfeature, the users can select how many beds they have in their residence to get the estimated price.

6.6.2 Test Cases

Test cases related to the above features are given below:

Table 8: Table of Price Determination Test Cases

TC ID	Requirements	Priority	Scenario Description
PD.HT	6.6.1	High	Select house type which is the entire house or the entire apartment
PD.RT	6.6.1	High	If room is selected in house type, Select room type which is shared or private room
PD.ND	6.6.1	High	Select neighborhood of İstanbul which is the house is located in
PD.BA	6.6.1	High	Select number of bathrooms in the house
PD.BR	6.6.1	High	Select number of bedrooms in the house
PD.BE	6.6.1	High	Select number of beds in the house

6.6.3 Estimated Price (EP)

6.6.3.1 Subfeatures To Be Tested

6.6.3.1.1 Get Price (EP.GP)

This subfeature calculates the price according to the other subfeatures such as house type, room type, neighborhood, bathrooms, bedrooms and beds selected by the user

6.6.4 Test Cases

Test cases related to the above feature are given below:

Table 9: Table of Estimated Price Test Cases

TC ID	Requirements	Priority	Scenario Description
EP.GP	6.6.3	High	Calculate the price according to the other subfeatures.

6.7 Detailed Test Cases

6.7.1 PD.HT

Table 10: Table of House Type Detailed Test Cases

TC_ID	PD.HT
Purpose	Select House Type
Requirements	6.6.1
Priority	High.
Estimated Time Needed	10-15 Seconds
Dependency	The simulation is executed.
Setup	No setup installation needed.
Procedure	Select House type from house type combo box.
Cleanup	No cleanup function.

6.7.2 PD.RT

Table 11: Table of Room Type Detailed Test Cases

TC_ID	PD.RT
Purpose	Select Room Type
Requirements	6.6.1
Priority	High.
Estimated Time Needed	10-15 Seconds
Dependency	The simulation is executed.
Setup	No setup requirement needed.
Procedure	Select room type from room type combo box.
Cleanup	No cleanup function

6.7.3 PD.ND

Table 12: Table of Neighborhood Detailed Test Cases

TC_ID	PD.ND
Purpose	Select Neighborhood.
Requirements	6.6.1
Priority	High.
Estimated Time Needed	10-15 Seconds
Dependency	The simulation is executed.
Setup	No setup installation needed.
Procedure	Select the neighborhood from neighborhood combo box.
Cleanup	No cleanup function.

6.7.4 PD.BA

Table 13: Table of Bathrooms Detailed Test Cases

TC_ID	PD.BA
Purpose	Select number of bathrooms
Requirements	6.6.1
Priority	High.
Estimated Time Needed	10-15 Seconds
Dependency	The simulation is executed.
Setup	No setup installation needed.
Procedure	Select the number of bathrooms from the combo box.
Cleanup	No cleanup function

6.7.5 PD.BR

Table 14: Table of Bedrooms Detailed Test Cases

TC_ID	PD.BR
Purpose	Select number of bedrooms
Requirements	6.6.1
Priority	High.
Estimated Time Needed	10-15 Seconds
Dependency	The simulation is executed.
Setup	No setup installation needed.
Procedure	Select the number of bedrooms from the combo box.
Cleanup	No cleanup function.

6.7.6 PD.BE

Table 15: Table of Beds Detailed Test Cases

TC_ID	PD.BE
Purpose	Select number of beds
Requirements	6.6.1
Priority	High.
Estimated Time Needed	10-15 Seconds
Dependency	The simulation is executed.
Setup	No setup installation needed.
Procedure	Select the number of beds from the combo box.
Cleanup	No cleanup function.

6.7.7 EP.GP

Table 16: Table of Get Price Detailed Test Cases

TC_ID	EP.GP
Purpose	See estimated price
Requirements	6.6.3
Priority	High.
Estimated Time Needed	20-25 Seconds
Dependency	The simulation is executed.
Setup	No setup installation needed.
Procedure	See the estimated price in screen
Cleanup	Go back to previous page.

7. Test Results

7.1 Individual Test Results

Table 17: Table of Individual Test Results

TC ID	Priority	Date Run	Result	Explanation
PD.HT	H	19.05.2021	Pass	The user can select the house type.
PD.RT	H	19.05.2021	Pass	The user can select the room type.
PD.ND	H	19.05.2021	Pass	The user can select the neighborhood.
PD.BR	H	19.05.2021	Pass	The user can select the number of bedrooms.
PD.BA	H	19.05.2021	Pass	The user can select the number of bathrooms.
PD.BE	H	19.05.2021	Pass	The user can select the number of beds.
EP.GP	H	19.05.2021	Pass	The user can see the estimated price for their house.

7.2 Summary of Test Results

We have executed 7 test cases and 7 test cases are passed. Also, 7 high-priority test cases are passed. Exit criteria are met.

Table 18: Table of Summary Test Results

Priority	Number of TCs	Executed	Passed
H	7	7	7
M	0	0	0
L	0	0	0
Total	7	7	7

7.3 Exit Criteria

We have executed all test cases and 100% of test cases are passed. Also, 100% of high-priority test cases are passed. Exit criteria are met.

Table 19: Table of Exit Criteria

Criteria	Met or Not
100% of the test cases are executed.	M
100% of the test cases passed.	M
100% of High Priority test cases passed	M
No high priority or severe bugs are left outstanding.	M

7.4 Model Test Results

The random forests (RF) approach which we used in our model training involves producing multiple regression trees, which are then combined to make a single consensus prediction for a given observation. We used the mean squared error (abbreviated MSE) as a measure of the prediction accuracy of the RF model.

For model training, 25% of the data set is reserved for testing and the rest for training. The feature selection process was applied to examine the effect of the features in the data set on the price. During this process, some of the features that had a low feature selection coefficient and were not suitable for the main purpose of the project were eliminated. In the elimination process, attention was paid to the continuous decrease of the mean squared error value in order to improve the estimation of the model.

Airbnb Istanbul dataset which we used in our model has 24519 samples(rows) and 74 features(columns) but There are too many null samples and unnecessary features. Because of that, we eliminated some of them to predict price correctly. Then, we had 65 features that have data inside.

With 65 features, Our mean squared error was equal to 580.419. This error represents the result of the comparison of test data and data from prediction. We tried to reduce errors with feature selection.

Our 65 features before feature selection:

['bathrooms_text', 'bedrooms','beds','price', 'Adalar', 'Arnavutkoy', 'Atasehir', 'Avcilar', 'Bagcilar', 'Bahcelievler', 'Bakirkoy', 'Basaksehir', 'Bayrampasa', 'Besiktas', 'Beykoz', 'Beylikduzu', 'Beyoglu','Buyukcekmece', 'Catalca', 'Cekmekoy', 'Esenler', 'Esenyurt', 'Eyup', 'Fatih','Gaziosmanpasa','Gungoren',’'Kadikoy','Kagithane','Kartal','Kucukcekmece','Maltepe',' Pendik',’'Sancaktepe', 'Sariyer', 'Sile', 'Silivri', 'Sisli', 'Sultanbeyli', 'Sultangazi','Tuzla','Umraniye', 'Uskudar','Zeytinburnu', 'Entire apartment','Entire condominium', 'Entire house', 'Entire loft', 'Entire serviced apartment', 'Entire townhouse', 'Entire villa', 'Private room in apartment', 'Private room in bed and breakfast','Private room in condominium', 'Private room in house', 'Private room in serviced apartment', 'Private room in townhouse', 'Room in aparthotel', 'Room in boutique hotel', 'Room in hotel', 'Shared room in apartment', 'Hotel room','Entire home/apt', 'Hotel room', 'Private room', 'Shared room']

- In the first part of our model test, We had 65 features in the dataset and our mean squared error which comes from a comparison of test data and data from the prediction in the model was equal to 580.419. We dropped some unnecessary features which were in the range of coefficient of 0.0..- 0.00... In this range, There were 18 unnecessary feature, these are

Name of Feature	**Coefficient**
Atasehir	*0.080843*
Avcilar	*0.059768*
Bagcilar	*0.040850*
Bahcelievler	*0.048892*
Bakirkoy	*0.079385*
Bayrampasa	*0.010626*
Beylikduzu	*0.067246*
Catalca	*0.031474*
Cekmekoy	*0.031351*
Esenler	*0.007115*
Gaziosmanpasa	*0.088927*
Gungoren	*0.004676*
Kartal	*0.050583*

Pendik	*0.031014*
Sancaktepe	*0.017832*
Sultanbeyli	*0.003509*
Sultangazi	*0.004400*
Tuzla	*0.068309*

After dropped these features, our mean squared error was equal to 573.014 and our new number of feature was equal to 47.

- In the second part of our model test, We had 47 features in the dataset and our mean squared error which comes from a comparison of test data and data from the prediction in the model was equal to 573.014. We dropped some unnecessary features which were in the range of coefficient of 0.1..- 0.2... In this range, There were 7 unnecessary feature, these are

Name of Feature	**Coefficient**
Entire loft	*0.085276*
Entire condominium	*0.185649*
Entire townhouse	*0.110871*
Entire serviced apartment	*0.243123*
Private room in condominium	*0.212524*
Private room in townhouse	*0.287370*
Room in aparthotel	*0.211855*

After dropped these features, our mean squared error was equal to 566.866 and our new number of feature was equal to 40.

- In the third part of our model test, We had 40 features in the dataset and our mean squared error which comes from a comparison of test data and data from the prediction in the model was equal to 566.866. We dropped some unnecessary features which were in the range of coefficient of 0.2..- 0.3..-0.5.. In this range, There were 6 unnecessary feature, these are

Name of Feature	**Coefficient**
Private room in house	*0.316958*
Private room in serviced apartment	*0.269792*
Shared room in apartment	*0.297676*
Hotel room	*0.389874*

After dropped these features, our mean squared error was equal to 566.566 and our new number of feature was equal to 34.

- In the final part of our model test, We had 34 features in the dataset and our mean squared error which comes from a comparison of test data and data from the prediction in the model was equal to 566.566. We dropped rest of the unnecessary features.

Name of Feature	**Coefficient**
Private room in bed and breakfast	*0.704659*
Room in boutique hotel	*0.458505*
Room in hotel	*0.644327*

After dropped these features, our mean squared error was equal to 566.485 and our new number of feature was equal to 32.

Our 32 Features after feature selection:

['bathrooms_text', 'bedrooms','beds', 'price', 'Adalar', 'Arnavutkoy', 'Basaksehir', 'Besiktas', 'Beykoz', 'Beyoglu', 'Buyukcekmece', 'Esenyurt', 'Eyup', 'Fatih', ’'Kadikoy', 'Kagithane', 'Kucukcekmece', 'Maltepe', 'Sariyer', 'Sile', 'Silivri', 'Sisli'', 'Umraniye', 'Uskudar', 'Zeytinburnu', 'Entire apartment', 'Entire house', 'Entire villa', 'Private room in apartment', 'Entire home/apt', 'Private room', 'Shared room']

So, we can show our all test parts in this table.

Table 20: Table of Model Test

Number of Test Process	Number of Features	Mean Squared Error value	Range of coefficient of features which dropped	Number of features after dropping	New Mean Squared Error value
1	65	580.419	0.0...-0.00...	47	573.014
2	47	573.014	0.1... -0.2...	40	566.866
3	40	566.866	0.2..-0.3..-0.5..	34	566.566
4	34	566.566	-	32	566.485

Thanks to this feature selection, unnecessary features were eliminated. Important 32 features for prediction were used in model. After feature selection with Random Forest, Mean Squared Error was reduced to 566.485 in our model.

8. Discussion

First of all, exploratory data analysis was performed for the design of the backend part of the project. For this analysis, the data set was first read from the jupyter notebook. The records in the data set belong to Istanbul Airbnb data in Turkey. In this data set, there are 74 different columns, namely features, and 24519 rows of records. Due to the high number of records included, the data set has been eliminated according to the range in which the price distribution is dense. Therefore, the price feature to be used for estimation in the model was recovered from outlier data and placed in a smooth range. Throughout this project, our approach has been to focus on the region where the price is between 0-10000, but other approaches can be tried and more varied results can be obtained.

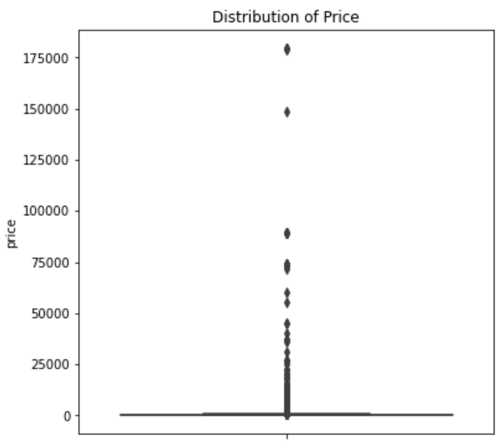

Figure 15: Box Plot For Distribution of Price

Then, operations such as dealing with missing data, dropping unnecessary features, dropping completely empty records and dealing with categorical data required for exploratory data analysis were carried out. As a result of these processes, the data set has become suitable for model training.

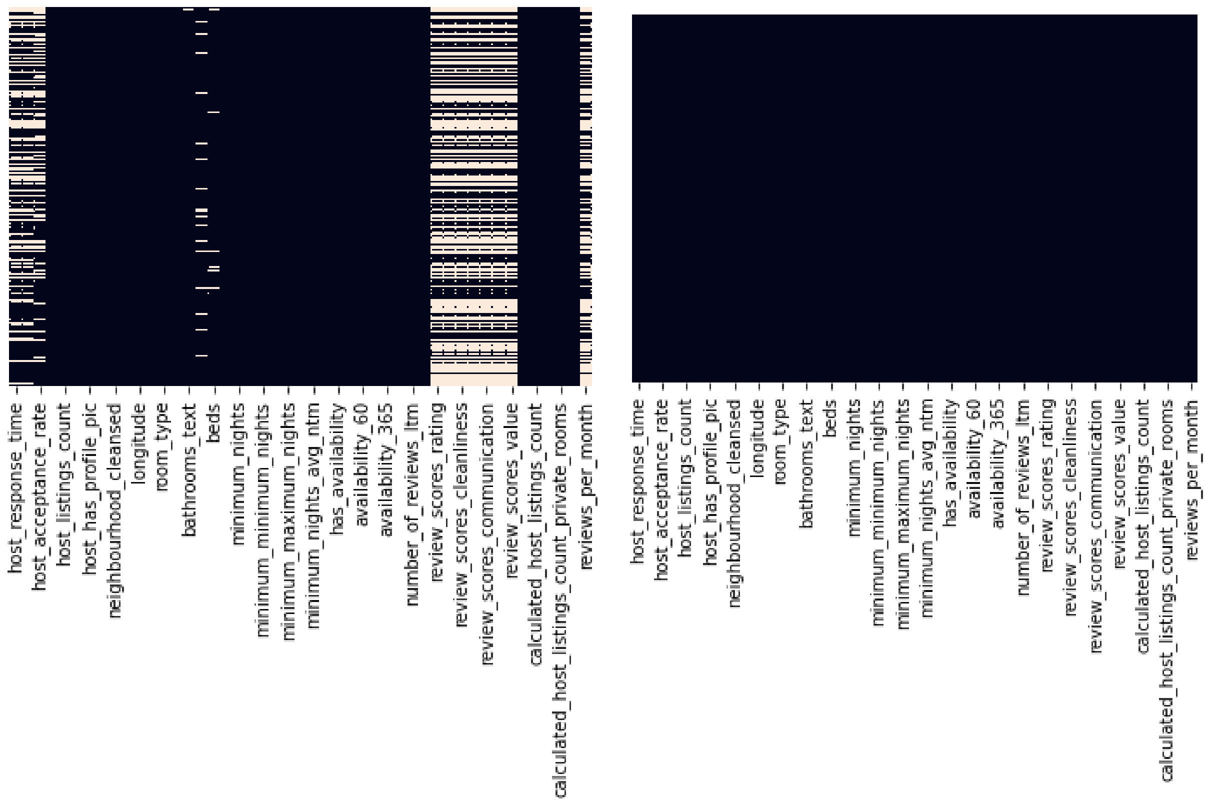

Figure 16: Heat Maps of Exploratory Data Analysis

Random forest regressor was chosen as the model to be trained, because the price, which is the feature to be used for prediction, includes continuous numerical values, so regressor models were chosen instead of classification models. In addition, a suitable model has been operated by considering hardware constraints such as GPU and CPU. For model training, 25% of the data set is reserved for testing and the rest for training. The feature selection process was applied to examine the effect of the features in the data set on the price. During this process, some of the features that had a low feature selection coefficient and were not suitable for the main purpose of the project were eliminated. In addition, as a result of this section, the main features presented in the frontend and their sub-features have emerged. In short, the features that emerge as a result of the feature selection process are 32 in total.

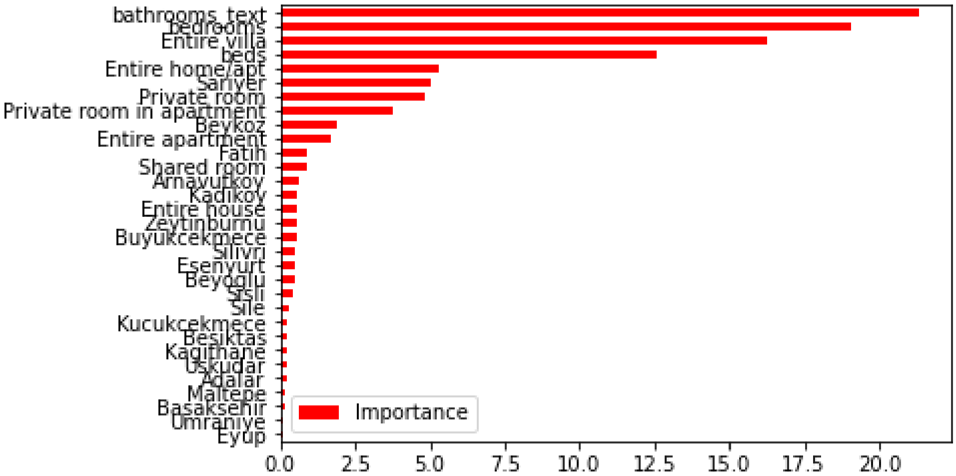

Figure 17: Barh Plot of Feature Selection Process

In the elimination process, attention was paid to the continuous decrease of the mean squared error value in order to improve the estimation of the model, because regression loss function, meanly mean squared error (MSE), value is the average of the difference between the actual value and the predicted value for all predictions made by the algorithm. In a word, this is the mean of the squares of all errors. As a result of this process, the initial MSE value of 580.4197063127456 was reduced to 566.485489050685. This means that approximately 30 out of 100 data are predicted correctly. A better results can be obtained by working on a different data set with more consistent records, focusing on different parts of the data set, taking test and train data at different rates, or trying other models and methods.

As a result of all the studies mentioned in the report, the project realizes the price estimation in an optimized way. In addition, we expect this project we have developed to shed light on it as an efficient and dynamic resource for similar studies in the future.

9. Conclusion

As a result, thanks to the Innovative System Design and Development I-II courses, we learned some machine learning algorithms and the models in which these algorithms are applied while examining the literature studies of similar projects. Additionally, we had the opportunity to examine Airbnb data sets, make sense of the data and use them in models. After choosing the appropriate data set for Airbnb Istanbul, we tried the algorithms and models we will use for price prediction and developed the model that could best predict the data set. While doing all this, we learned to determine the requirements of a real system, design and test it. In this process, before the implementation phase of the project, we had some difficulties in the models we will choose for the system, determining the features for price estimation and choosing the appropriate data set, however, overcoming all these difficulties, we completed the project on time by designing it in accordance with all its requirements. In our system, features such as house type, room type, number of bedrooms, number of beds, number of bathrooms and neighborhood are used for price determination. We developed the random forest regressor machine learning model in the back-end part of the project, and completed the web application with flask in the front-end part.

Acknowledgement

First of all, we would like to thank our valuable project consultant Roya Choupani, who answered our questions tirelessly during our weekly meetings throughout the project and always made us feel her support, and then to all our professors and authors whose documents and articles we reviewed. We are truly grateful for everything you do, Ms. CHOUPANI.

References

[1] V. Raul Perez-Sanchez, Leticia Serrano-Estrada , Pablo Marti , Raul-Tomas Mora-Garcia. The What, Where, and Why of Airbnb Price Determinants. December 2018.

[2] Yuan Caı,Yongbo Zhou, Jıanyu (Jenny) Ma, Noel Scott. Price Determinants of Airbnb Listings: Evidence From Hong Kong. May 2019.

[3] Insu Hong, Changsok Yoo. Analyzing Spatial Variance of Airbnb Pricing Determinants Using Multiscale GWR Approach. June 2020.

[4] Yuanhang Luo (royluo), Xuanyu Zhou (xuanyu98), Yulian Zhou (zhouyl). Predicting Airbnb Listing Price Across Different Cities. December 14, 2019

[5] http://insideairbnb.com/get-the-data.html