A Machine Learning-Based Price Prediction for Istanbul Airbnb Data

08/06/2023

Aslı Gültekin, Oğuzhan Gündüz

T.C.
Dokuz Eylül University
Faculty of Science
Department of Mathematics

Aslı Gültekin
asligltkn1@gmail.com

Oğuzhan Gündüz
oguzhan.gunduz111@gmail.com

Acknowledgments

We would like to thank our consultant Prof. Dr. Özlem Ege Oruç for the guidance, motivation and support she gave us in this project. We express our deep gratitude for her suggestions and patience.

Oğuzhan Gündüz
Aslı Gültekin

Özet

Airbnb, ev sahipleri ile konuklar arasında bağlantı sağlayan büyük bir platform şirketidir. Bu yaygın olarak kullanılan platform, ev sahiplerinin konuklar için en uygun ev fiyatını belirlemeleri açısından önemlidir. Makine öğrenmesinde sıkça kullanılan bir tahmin modelleme tekniği olan regresyon modeli, genellikle değişkenler arasındaki nedensel ilişkileri keşfetmek ve tahmin analizi yapmak için kullanılır. Bu projede amacımız, İstanbul'daki Airbnb konut ilan fiyatlarını tahmin etmek için çeşitli makine öğrenmesi regresyon yöntemlerini kullanmaktır. Çalışmada rastgele orman regresyon modeli, doğrusal regresyon modeli, K-en yakın komşu regresyon modeli ve Gradient Boosting Regresyon modeli gibi dört regresyon modeli seçilmiştir. Elde edilen sonuçlara göre, Rastgele Orman Regresyon modeli bu veri seti için en iyi tahmini sağlayan model olarak belirlenmiştir.

Anahtar Kelimeler: Makine öğrenmesi, Regresyon modeli, R-squared.

Abstract

Airbnb is a major platform company that connects hosts and guests. In this widely used platform, determining the most suitable house prices for guests is of utmost importance for hosts. Regression modeling is a popular technique in machine learning used for prediction modeling. It is commonly employed to explore causal relationships between variables and conduct prediction analysis. The aim of this project is to predict housing listing prices for Istanbul Airbnb using various machine learning regression methods. Four regression models have been selected for this study: Random Forest Regression, Linear Regression, K-Nearest Neighbor Regression, and Gradient Boosting Regression. Based on the results obtained, the Random Forest Regression model has been identified as the most accurate model for predicting prices in this dataset

Keywords: Machine learning, Regression model, R-squared.

Contents

1 Introduction

Airbnb is a prominent online platform that offers short-term rentals of homes and rooms, providing travelers with unique accommodations worldwide. Unlike traditional hotels, Airbnb allows hosts to set their own prices based on their experience. However, determining the optimal price that balances profitability and popularity can be challenging for new hosts.To address these challenges and assist both hosts and guests, it is crucial to evaluate the reasonability of current prices and identify favorable booking times.

Machine learning, a rapidly advancing field in computer technology, offers valuable tools for tackling these tasks. Machine learning has permeated various aspects of life, and its application in predicting Airbnb prices can benefit both consumers and hosts. Several studies have been conducted to develop price prediction models for Airbnb. In 2019, P. Rezazadeh et al. conducted a study titled "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis" . Their aim was to develop a reliable price prediction model using machine learning, deep learning, and natural language processing techniques. This model aimed to assist property owners and customers in evaluating prices based on limited information about the property. Another study by M. Mahyoub and colleagues, titled " Airbnb Price Prediction Using Machine Learning," compared the performance of various machine learning algorithms and methodologies in predicting Airbnb prices. [2][3][4]

The aim of this study is to predict Airbnb housing prices in Istanbul using various machine learning models and determine the best-performing model by comparing their performances. The study involves a data preprocessing process to prepare the data obtained from the Airbnb website for analysis specifically for Istanbul city. Subsequently, a data visualization process is conducted to examine the relationship between rental prices and different factors. Finally, different machine learning models are applied for price prediction, and evaluation criteria such as RMSE (Root Mean Squared Error) and R-squared are used to determine the most accurate model for this dataset. The study consists of four sections: Introduction, methodology where the machine learning methods used in the study are introduced, application and conclusion.

2 Methodology

Machine learning is a subfield of artificial intelligence that focuses on the development of algorithms and models enabling computers to learn and make predictions or decisions without explicit programming. It involves creating mathematical models and algorithms that enable machines to analyze and interpret complex data, identify patterns, and make predictions or take actions based on those patterns. Machine learning algorithms learn from historical data and iteratively improve their performance over time by adjusting parameters or updating models. They can be categorized into supervised learning, unsupervised learning, and reinforcement learning, depending on the type of input data and desired output. Machine learning finds wide applications in domains such as image and speech recognition, natural language processing, recommendation systems, autonomous vehicles, and fraud detection, among others.

The primary goal of machine learning is to enable computers to learn from data and enhance their performance on specific tasks without explicit programming. By leveraging statistical techniques and algorithms, machines can analyze large datasets, identify patterns, and make accurate predictions or decisions. This technology has revolutionized numerous industries by automating tasks, extracting valuable insights from data, and enhancing decision-making processes. Machine learning has diverse applications in healthcare, finance, manufacturing, marketing, and various other fields where the ability to analyze and interpret data can lead to improved outcomes. It holds the potential to uncover hidden patterns, optimize processes, and drive innovation across sectors.

2.1 Machine Learning Models

2.1.1 K-Nearest Neighbor Regression

K-Nearest Neighbors (KNN) is a popular and versatile algorithm used for classification tasks in machine learning [5]. It works based on the assumption that similar instances tend to belong to the same class. The algorithm determines the class of a new instance by considering the classes of its K nearest neighbors in the training data. Several techniques can enhance the performance of KNN, such as choosing appropriate distance metrics, determining the optimal value of K, handling imbalanced data, feature selection and dimensionality reduction, and scaling and normalization. These techniques help improve the accuracy and reliability of KNN in real-world applications.

2.1.2 Linear Regression

Linear regression is a statistical modeling technique used to analyze the relationship between a dependent variable and one or more independent variables [6]. The objective is to find the best-fitting linear equation that represents this relationship.

The general form of a simple linear regression equation is given as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where y represents the dependent variable, x represents the independent variable, β_0 is the y-intercept, β_1 is the coefficient or slope, and ϵ represents the error term.

The linear regression model estimates the values of β_0 and β_1 by minimizing the sum of squared residuals. This is achieved through the least squares method, which aims to find the line that best fits the data.

To assess the quality of the linear regression model, the coefficient of determination (R^2) is commonly used. It measures the proportion of the variance in the dependent variable that can be explained by the independent variable(s).

The parameters β_0 and β_1 can be estimated using techniques like ordinary least squares (OLS), gradient descent, or matrix methods. These methods ensure unbiased estimates and provide the best fit to the data.

In multiple linear regression, where there are more than one independent variables, the equation can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Here, $x_1, x_2, \ldots, x_n$ represent the independent variables, and $\beta_1, \beta_2, \ldots, \beta_n$ represent their respective coefficients.

Linear regression finds wide applications in economics, finance, social sciences, and machine learning. It serves as a valuable tool for understanding variable relationships and making predictions based on learned patterns.

2.1.3 Random Forest Regression

Random Forest Regression is a powerful and versatile machine learning algorithm used for regression tasks [7]. It combines decision trees and ensemble learning to make accurate predictions. By creating multiple decision trees trained on different subsets of the data, it reduces overfitting and improves generalization performance. It can handle large and complex datasets, capture non-linear relationships, and work with both categorical and numerical features. Random Forest Regression is robust to outliers and provides valuable insights into feature importance. Its applications span across industries such as finance, healthcare, and marketing, making it a popular choice for regression problems.

2.1.4 Gradient Boosting Regression

Gradient Boosting Regression, a popular machine learning algorithm, is used for regression tasks [8]. It combines weak prediction models, typically decision trees, to create a strong predictive model. The algorithm iteratively builds models, correcting the errors of previous models. It uses gradient descent optimization to update model parameters and minimize the prediction error. The final prediction is the weighted sum of individual model predictions. Gradient Boosting regression captures complex relationships and nonlinearity in the data, making it effective for price prediction. It includes regularization techniques to prevent overfitting. Despite the need for careful tuning and computational requirements, Gradient Boosting Regression is a powerful tool for accurate regression tasks.

3 Application

3.1 Data Presentation

3.1.1 Source of Data

Airbnb is a highly popular platform for sharing accommodations, providing users with a wide range of lodging options. The Istanbul dataset used in this study was obtained from the Inside Airbnb website. This dataset encompasses diverse features and characteristics of Airbnb accommodations in Istanbul. [1]

3.1.2 Properties of Data

The dataset contains more than 40,000+ posts and consists of 75 different columns. Out of 75 features, It selects only the pricing-related ones from 75 features. Excludes properties like host_id, picture_url, etc..

Features	Description
Bedrooms	This column represents the number of bedrooms in each listing. It states how many beds there are in the advertisements.
Neighbourhood	This column denotes the geographical location of the listing, indicating the specific neighborhood where it is situated.
property type	This column describes the property type of the listing, such as house, aparthotel, boutique hotel, guesthouse, apartment, dome, villa, dorm, and more.
room type	This column signifies the type of room offered in the listing, categorizing it as an entire home/apartment, private room, shared room, or hotel room.
amenities	This column presents a list of amenities available within the room, detailing the provided facilities.
price	This column represents the value of the listing, indicating the daily rental price.
minimum nights	This column denotes the maximum number of nights allowed for a stay.
review scores rating	This column reflects the rating score derived from user reviews, representing an overall evaluation.
review scores accuracy	This column signifies the accuracy rating provided in user reviews.
review scores cleanliness	This column indicates the cleanliness rating mentioned in user reviews.
review scores checkin	This column represents the check-in experience rating given by users.
review scores communication	This column signifies the communication rating provided in user reviews.
review scores location	This column denotes the location rating mentioned in user reviews.
review scores value	This column reflects the value rating provided in user reviews.
reviews per month	This column represents the number of monthly reviews received by the listing.
bathrooms	Number of bathrooms
accommmodates	Number of amenities assigned/defined by Airbnb.

3.1.3 EDA and Pre Processing

The "review Avg" feature is created by averaging the features related to the review score. (review scores rating, review scores accuracy, review score cleanliness, review scores checkin, review scores communication, review scores location, review score value).
In order to find out how many people can stay in the ad, the "max person to stay" feature is created as the maximum number of people. Since some beds have a capacity of 1 and some beds

have a capacity of 2, the number of people can be found according to the number of beds.

By using the neighborhood cleaning feature, advertisements on the European and Anatolian sides were found and is Anadolu was created.

property_type	room_type	neighbourhood	max_person_to_stay	bedrooms	olanaklar	review_Avg	reviews_per_month	is_anadolu	price	minimum_nights	accommodates	bathrooms
loft	Private room	Sisli	2	1	1.75	1.594933	0.27	0	384	2	2	1
rental unit	Entire home/apt	Sariyer	4	2	1.00	0.000000	0.00	0	2396	3	2	1
rental unit	Private room	Beyoglu	2	1	0.25	0.000000	0.00	0	958	3	3	1
rental unit	Private room	Beyoglu	1	1	0.75	0.000000	0.01	0	2087	1	2	1
rental unit	Private room	Sisli	1	1	1.00	1.609438	0.01	0	1102	2	2	1

Head Of Dataset

The first four samples from the dataset were analyzed to provide a glimpse into the range of listings available on Airbnb in Istanbul. A comprehensive review of the descriptive features, including minimum, maximum, mean, and standard deviation, was performed to gain insight into the characteristics of the dataset.

	max_person_to_stay	bedrooms	olanaklar	review_Avg	reviews_per_month	is_anadolu	price	minimum_nights	accommodates	bathrooms
count	34818.0	34818.0	34818.0	34818.0	34818.0	34818.0	34818.0	34818.0	34818.0	34818.0
mean	3.0	1.4	1.4	0.9	0.7	0.1	1830.9	1.9	3.3	1.2
std	3.2	1.2	0.7	0.8	1.1	0.2	15733.3	1.2	2.0	0.9
min	1.0	0.0	0.0	0.0	0.0	0.0	60.0	1.0	1.0	0.0
25%	1.0	1.0	0.8	0.0	0.0	0.0	626.0	1.0	2.0	1.0
50%	2.0	1.0	1.5	1.5	0.2	0.0	1000.0	1.0	3.0	1.0
75%	4.0	2.0	2.0	1.6	1.0	0.0	1605.8	3.0	4.0	1.0
max	163.0	50.0	5.0	1.6	18.8	1.0	1878296.0	5.0	16.0	50.0

The price variable exhibits a considerably large standard deviation, suggesting potential presence of outliers. Upon assessing the interquartile range (IQR), it appears that there may be outlier values. Notably, there is a substantial disparity between the minimum and maximum values. Therefore, it is advisable to consider a transformation of the data to address this issue.

Similarly, when examining the "minimum_night" variable, the IQR reveals the possibility of extremely high maximum values. It is crucial to perform outlier detection and cleansing for this variable as well.

On a positive note, the variables "max_person_stay", "bedrooms", "review_Avg", and "review_per_month" demonstrate a relatively homogeneous distribution, as evidenced by the proximity of their mean and standard deviation values.

	property_type	room_type	neighbourhood	is_anadolu
count	34818	34818	34818	34818
unique	17	4	39	2
top	rental unit	Entire home/apt	Beyoglu	0
freq	21897	22656	8321	32721

The property_type variable reveals that the majority of rental units fall under 17 distinct types. Meanwhile, the neighborhood variable exhibits 39 unique types, with "Beyoğlu" emerging as the most prevalent neighborhood. Regarding the is_anadolu variable, it signifies whether a property is located in the Anatolian or European region. Upon examining the data, it becomes apparent that the most prevalent value is 0, indicating that the majority of properties are not situated in the Anatolian region.

3.1.4 Visualizing Data

A transformation is deemed necessary for the price variable due to the substantial difference between its minimum and maximum values, making it challenging to interpret the data on the chart without applying a transformation. To address this issue, a logarithm transformation has been applied to the price variable.

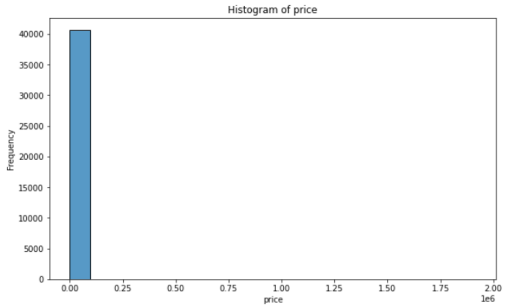


Figure 2: Price Distribution

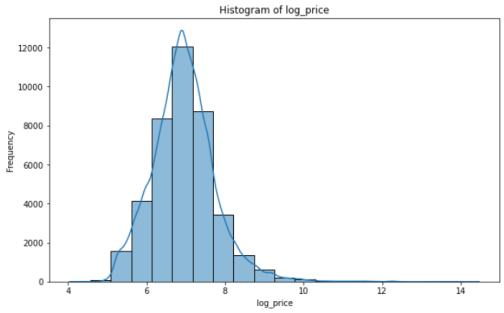


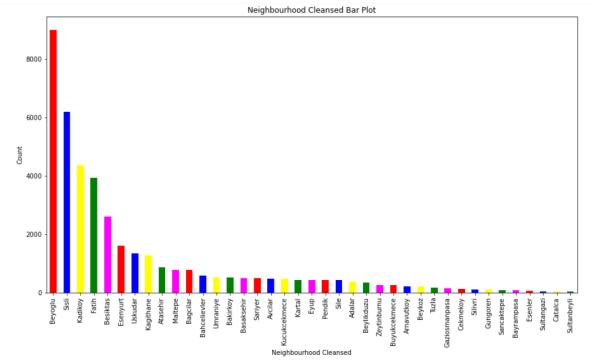
Figure 3: Histogram of Log Price

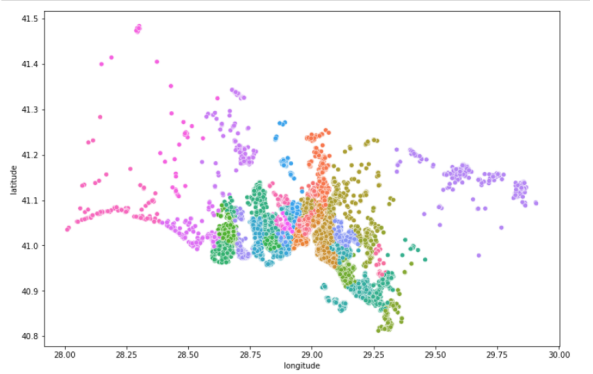
```
Test statistic: 0.9655727744102478
p-value: 0.0
The log_price variable does not have normal distribution.
```

When the Shapiro-Wilk test is applied, it is observed that the variable is not normally distributed.

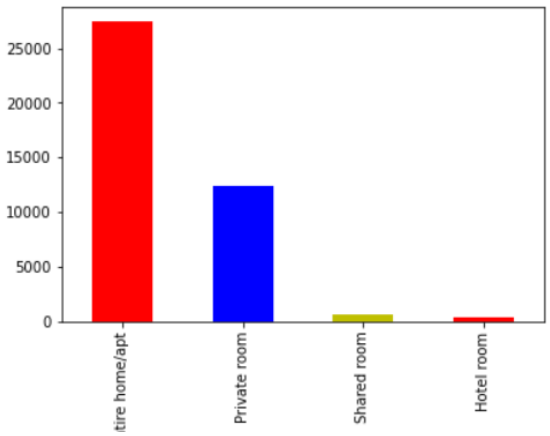
By examining the neighborhood variable, the distribution of neighborhoods in Istanbul can be understood. In particular, 'Beyoğlu' is the most common neighborhood, leaving the others behind in frequency. This is followed by the remaining neighborhoods, which appear in a sequential order.

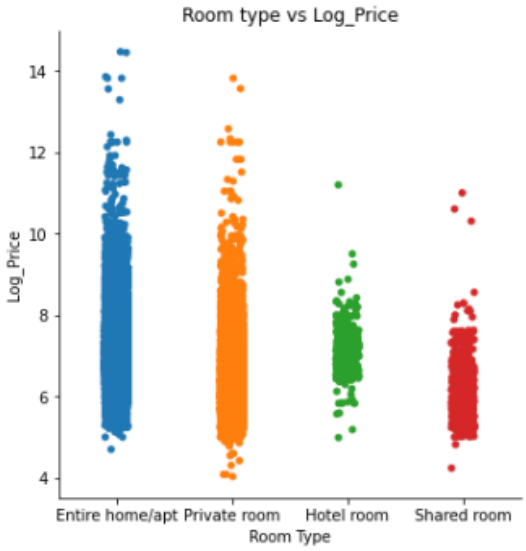
Latitude and longitude variables can be used to get more idea about the geographical location of Istanbul. By making use of these variables, it becomes possible to derive the spatial representation of Istanbul on a map and thus visualize the geographical shape of the city.





Upon visualizing the distribution of the room_type variable, it becomes evident that the Entire home/apt type showcases a broader price range compared to the shared room and hotel room types, which exhibit narrower price ranges.





A pie chart is created to show the distribution of categorical variables. A catplot is created to examine the relationship between the transformed "log_price" variable and other categorical variables. This chart will provide insight into how categorical variables affect pricing patterns.

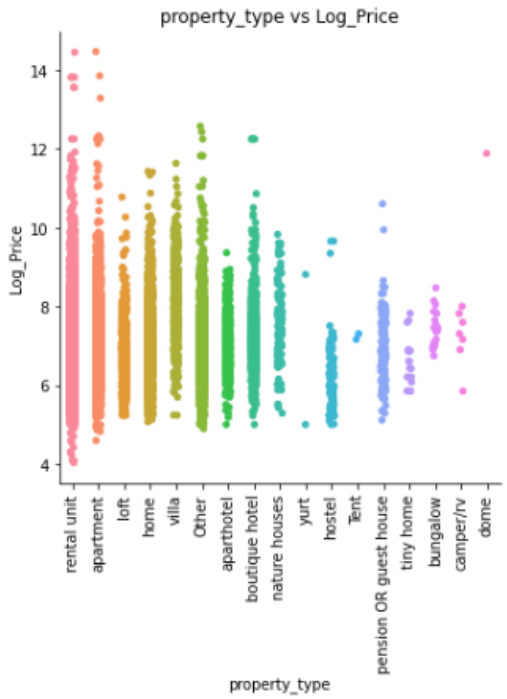


Figure 4: Property Type vs. Price

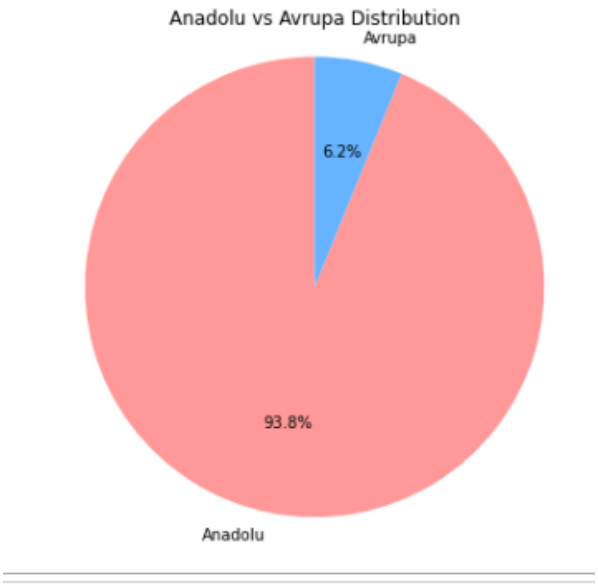
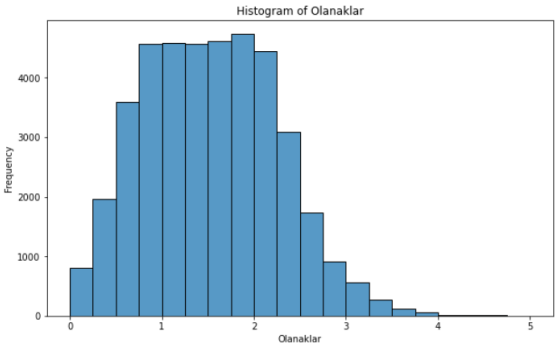
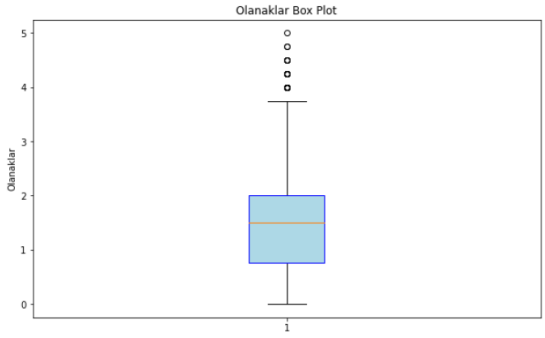


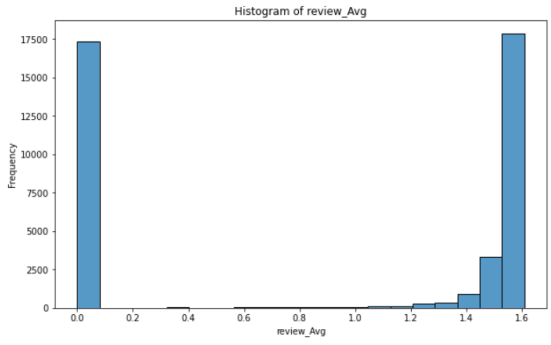
Figure 5: Anadolu Distribution

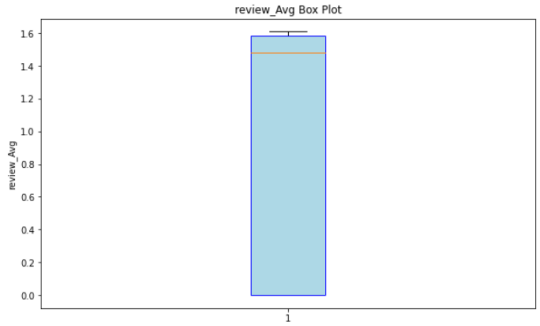
When looking at the distribution of the "olanaklar" (in English is "amenities") variable, which is a numeric value, it is observed to be right-skewed. The box plot also indicates the presence of values that can be considered outliers.



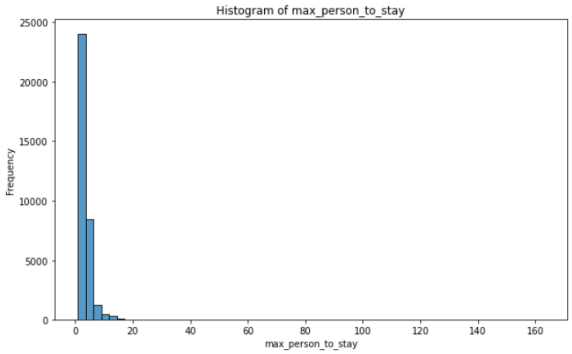


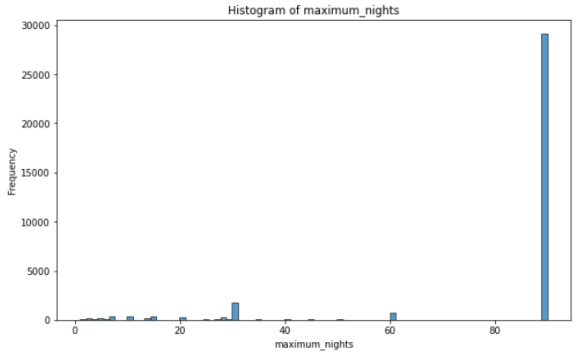
Examining the histogram of the variable review_Avg, it is clearly seen that the distribution is skewed to the left. This indicates that a significant number of rooms/houses did not receive any reviews, resulting in an average of 0 reviews. Conversely, where reviews are available, the values represent the average rating for the facility. Typically, the review scores range from 0 to 5, with 5 being the highest score.



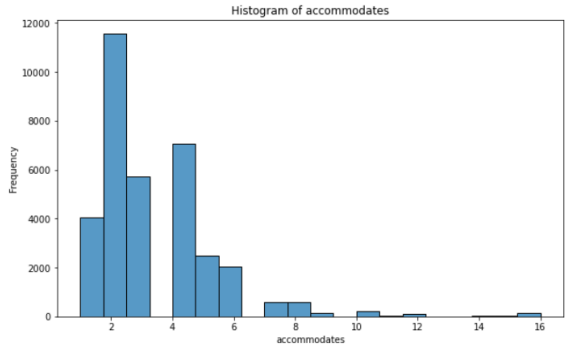


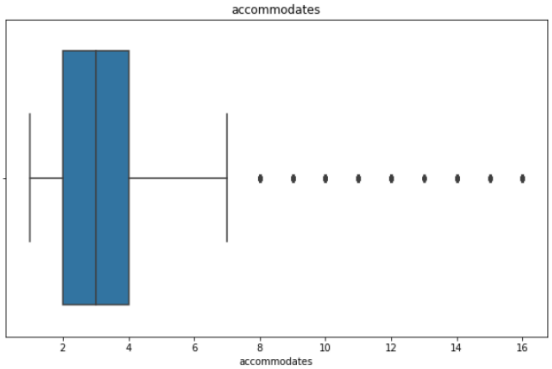
The variable max person stay shows a rigth skewed distribution. Outliers accumulated on the right.





The accommodates variable exhibits a right skew, with outliers accumulating on the right side.





Having gained an understanding of the variables and their distribution, the next step involves creating the model using these features and proceeding to estimate the "price" variable.

3.2 Methods

3.2.1 The Dependent Variable

The listing price is the dependent variable in this study. To minimize the impact of outliers on subsequent prediction results, a specific data range has been chosen to focus on.

Furthermore, to further refine the dataset, a z-score filter has been applied. Only data points with z-score values between -3 and 3, representing prices within a standardized range, have been considered. This filtering process allows for a concentration on the majority of data points while reducing the influence of extreme values.

By implementing these selection criteria, the aim is to ensure a more representative and reliable dataset for analysis, thereby enhancing the accuracy of price prediction models. The resulting dataset comprises over 34,000 data points, providing a robust foundation to capture underlying patterns and relationships effectively.

3.2.2 Independent Variables

The listing price of a property is influenced by numerous factors, including the property type, room type, bed type, accommodation capacity, amenities, and more. In this study, we have employed a selective approach by narrowing down the original pool of 75 variables to 10 independent variables that exhibit a direct correlation with the price. During this process, columns such as 'name', 'host_id', 'host_url', and other unrelated variables were excluded. By focusing on these carefully chosen independent variables, our aim is to capture the key determinants that significantly contribute to the variability in listing prices.

3.2.3 Data Processing

In order to simplify the programming process, all variables were transformed into numeric variables during the data processing phase. Missing values in the 'bedrooms' and 'review scores rating' variables were supplemented with zeros. The 'amenities' data, which was in the form of a list or array, was separated and assigned numeric values. For instance, a counter was incremented whenever certain keywords such as "wifi," "free park," "park," "coffee," and "washer" were found in the list. The incremented number was then assigned to a newly created column named "olanaklar" (amenities). This allowed us to convert the array-format data into a numeric format.

Within the dataset, there exists a column labeled "amenities" that comprises a list or array of values. In order to properly organize this data, it is necessary to categorize them into distinct groups. The designated categories for classification are as follows:

Category	Element Lists
furniture_and_amenities	['Chair', 'Clothing storage', 'Dedicated workspace', 'Elevator', 'Essentials', 'Iron', 'Smoking allowed']
entertainment_and_electronics	['Bluetooth', 'Netflix', 'Sound system', 'TV']
kitchen_and_dining	['Blender', 'Coffee maker', 'Cooking basics', 'Dishes', 'Fridge', 'Kettle', 'Kitchen', 'Microwave', 'Oven', 'Refrigerator', 'Toaster']
bathroom_and_toiletries	['Shampoo', 'Bathtub', 'Bath', 'Body soap', 'Crib', 'Conditioner', 'Hair dryer', 'Hot tub', 'Hot water', 'Shower', 'Washer']
outdoor_and_recreation	['BBQ', 'Balcony', 'Console', 'Exercise equipment', 'Fireplace', 'Hammock', 'Pool', 'Sauna']
climate_control_and_utilities	['Air conditioning', 'Ceiling fan', 'Heating', 'Stove']
safety_and_security	['Alarm', 'Fire extinguisher', 'First aid kit', 'Lock', 'Security', 'Smoke alarm']
parking	['Carport', 'Garage', 'Parking']
miscellaneous	['Long term stays allowed', 'View', 'Dryer', 'Pets allowed', 'Freezer']
pets_allowed	['Pets allowed']
internet_options	['Wifi', 'Ethernet connection']

Then, new features are generated by dividing the data into different categories. Implemented a code snippet in Python to search for specific "words" within the "amenities" attribute of each category. For example, in the "furniture_and_amenities" category, words such as "Chair" or "Clothes" were searched. Upon matching these words, the corresponding category attribute is assigned a value of +1. Using this methodology, a total of 11 new features (columns) were derived by effectively parsing the "amenities" attribute.

Subsequently, the correlations between the category features and the "price" value were examined. The top 4 categories displaying the strongest correlation with "price" were selected, and their average values were calculated. This new column, named "olanaklar" (meaning "amenities" in Turkish), was assigned as a feature to predict the "price" value.

By assigning incremented numbers to the "olanaklar" column, the data was transformed from an array format to a numeric format. This transformation enabled the utilization of this feature in predicting the "price" value.

3.3 The Outputs

3.3.1 Assumptions

In this section, the assumptions of the model will be discussed. Upon examining the correlation graph, a correlation is observed between the 'bedrooms' and 'max_person_to_stay' variables. This correlation can be attributed to the fact that the number of people who can stay in a room or property is typically associated with the number of bedrooms. Sellers often present the capacity of their homes based on the number of bedrooms available.

Similarly, a similar observation can be made regarding the relationship between the number

of bedrooms and the number of bathrooms. It is expected that the number of bathrooms in a property is proportional to the number of bedrooms.

Furthermore, a correlation is noticed between 'reviews_per_month' and 'review_Avg'. This correlation is reasonable as the number of reviews received for a property determines the average number of reviews per month. Hence, it is normal to observe this correlation between the two variables.

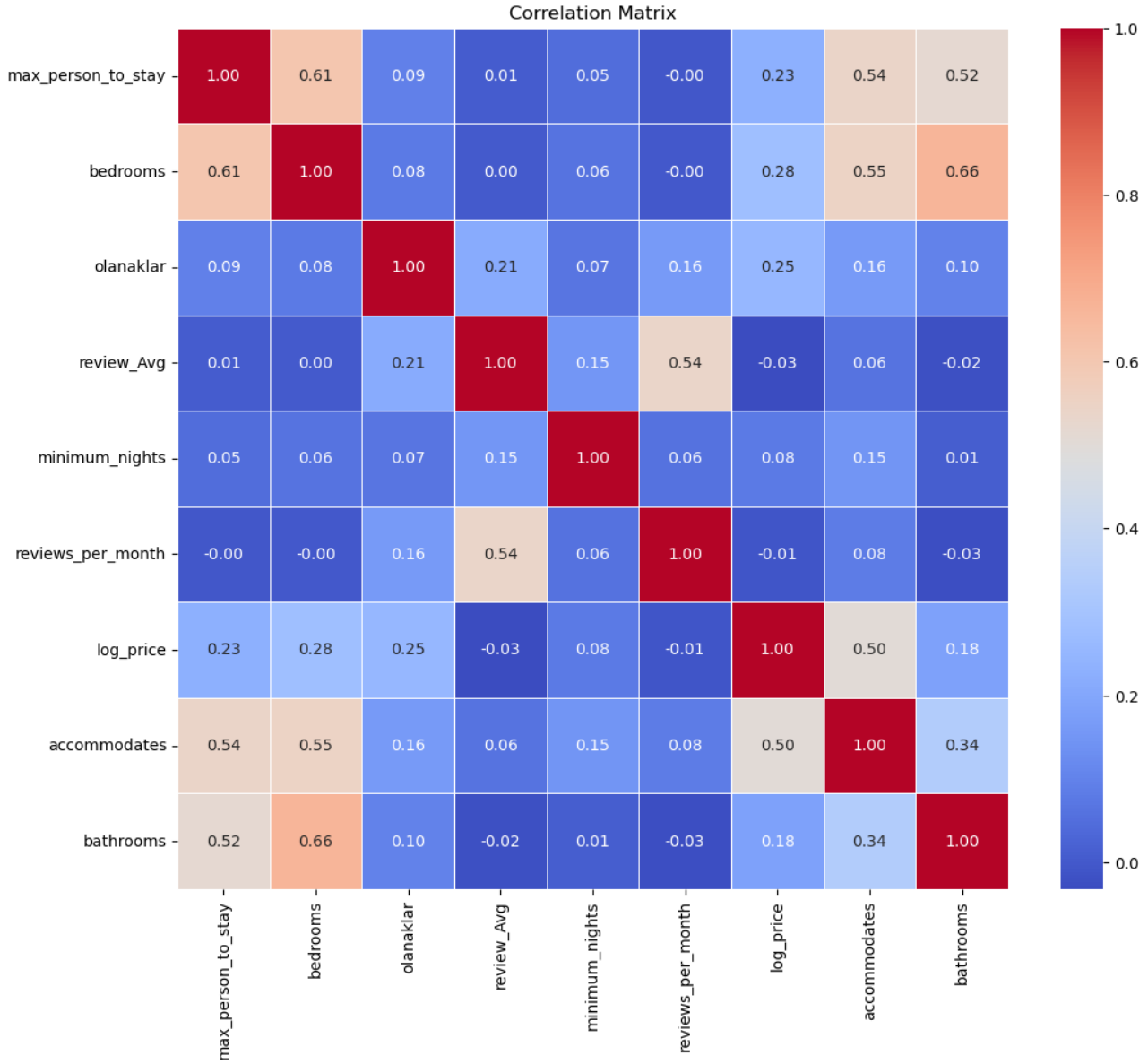
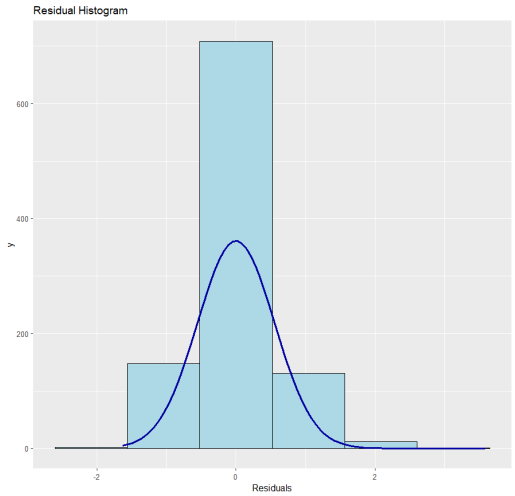


Figure 6: Correlation



The Residual Histogram provides information about the distribution of residuals. From the graph, it can be observed that the residuals follow a normal distribution, as they are centered around the mean and exhibit a relatively homogeneous distribution.

Figure 7: Residuals Histogram

In addition, upon examining the averages of the residuals, it is evident that they are close to zero. Specifically, the average of the residues is calculated to be 6.413251930472326e-16.

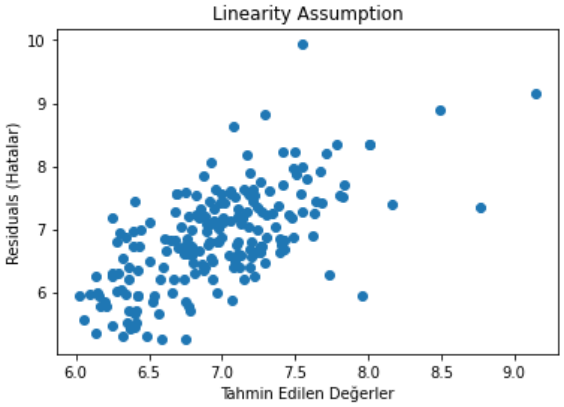
```
Rainbow Test Statistics: 0.9865744958165295
Rainbow Test p-value: 0.7873818799554115
```

Figure 8: Residuals Histogram

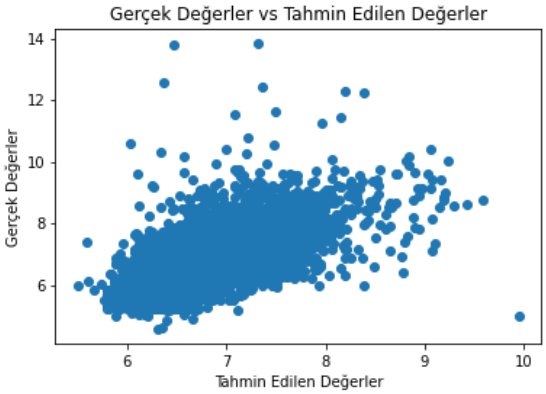
The rainbow test is used as a measure of model fit to evaluate the linearity assumption. A test statistic close to 1 indicates that the linearity assumption is well met. In addition, a p value greater than 0.05 indicates that the null hypothesis cannot be rejected and one can conclude that the linearity assumption is valid.

When evaluating the graphs, the "Residuals vs Fitted Value" plot should not display any noticeable pattern or trend. This indicates that the regression model accurately captures the linear relationship, and the residuals are randomly distributed. In the left visualization, the plot is generated by taking a sample of 1000 observations from the model, while the right visualization shows the plot for the entire dataset. Both plots exhibit no apparent pattern or trend, confirming the fulfillment of the linearity assumption.

Therefore, based on the results of the Rainbow test and the assessment of the graphs, a confident conclusion can be drawn that the linearity assumption is satisfied.



(a) Linearity Assumption 1000



(b) Linearity Assumption

To assess whether the assumption of constant variance is violated, an examination of the residual vs. fitted values plot is conducted. The graph displays an outward expansion, resembling a megaphone shape, indicating that the variance is not constant but rather heteroskedastic. To obtain a more conclusive determination, the Breusch-Pagan test is performed.

The Breusch-Pagan test evaluates the hypothesis that the variance is constant (null hypothesis, H0) against the alternative hypothesis (H1) that the variance is not constant. The test yields a small p-value of 0.0005313317, leading to the rejection of the null hypothesis and the conclusion that the variance is not constant.

Therefore, based on the residual vs. fitted values plot and the results of the Breusch-Pagan test, it can be concluded that the variance of the errors is not constant. The variability of the errors is dependent on the independent variables.

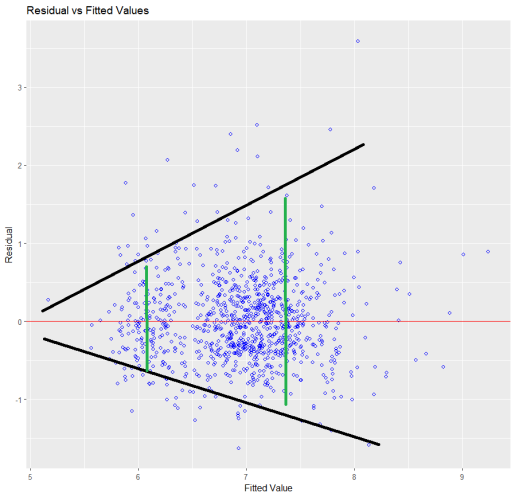


Figure 10: Residuals Histogram

```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

              Data
---------------------------------
Response : total
Variables: fitted values of total

        Test Summary
-------------------------------
DF            =    1
Chi2          =    12.00236
Prob > Chi2   =    0.0005313317
```

Figure 11: Breusch Pagan Test

3.4 The Model

The features used to build our models are as follows:

Table of Features

Variable	Category	Description
max_person_to_stay	int	Maximum number of people to stay.
bedrooms	int	Number of bedrooms.
olanaklar	float	A column of amenities created using the "amenities" column.
review_Avg	float	Average of reviews given by users.
minimum_nights	int	Minimum number of days to stay.
reviews_per_month	float	Amount of reviews per month.
is_anadolu	bool (categorical)	If the house or room is located in a district within Anatolia, the statement is true; otherwise, it is false.
accommodates	int	Number of amenities assigned/defined by Airbnb.
bathrooms	int	Number of bathrooms.
encoded_property_type	categorical	Type of building/propery, 'camp', 'hotel', etc..
encoded_room_type	categorical	Room type, example values; 'Private room', 'Entire home/apt', etc..
encoded_neighbourhood	categorical	The county where the house or room is located.
log_price	float	Logarithm of house or room prices.

The dataset was divided into a training set and a test set, with the test dataset accounting for 20% of the total data and the training dataset comprising the remaining 80%. This partitioning allowed for evaluating the models' performance on unseen data. The training set, representing 80% of the data, was used for model training to capture underlying patterns and relationships. Conversely, the test set served as a benchmark to assess the models' generalization capability and accuracy on new instances. By keeping this subset of data separate during training, an unbiased evaluation of the models' performance was ensured. The division into training and test sets provided valuable insights into the models' ability to generalize and their potential for real-world applications.

During the model construction, all available features were utilized. Twelve independent variables were employed to predict the logarithmically transformed 'price' value. Five different machine learning algorithms, namely KNN, Linear Regression, Gradient Boosting Regression, and Random Forest Regression, were utilized to achieve the best possible prediction. The models' performance was evaluated based on the R-squared (R^2) and root mean square (RMS) values. The results of the models are presented in Table 1.

Table 1: All Models

Model	**R-squared**	**RMS**
K-NN	0.521775	0.554127
Linear Regression	0.381371	0.630243
Gradient Boosting Regression	0.486660	0.574111
Random Forest Regression	0.889190	0.266736

Upon analyzing the outputs of our models, a comprehensive evaluation of the results reveals the following findings.

- **K-NN**:
  - R-squared: 0.521775
  - RMS: 0.554127

  The K-NN model performs reasonably well, with a moderate R-squared value suggesting that it captures a moderate amount of the variation in the target variable. The RMS value indicates a relatively small average deviation between the predicted and actual values, indicating reasonably accurate predictions.

- **Linear Regression**:
  - R-squared: 0.381371
  - RMS: 0.630243

  The Linear Regression model explains a moderate amount of the variation in the target variable, as indicated by the R-squared value. The RMS value suggests a moderate root mean square error, indicating a reasonable average deviation between the predicted and actual values.

- **Gradient Boosting Regression**:
  - R-squared: 0.486660
  - RMS: 0.574111

  The Gradient Boosting Regression model performs reasonably well, capturing a significant portion of the variation in the target variable. The R-squared value suggests a relatively good fit to the data. The RMS value indicates a reasonable average deviation between the predicted and actual values.

- **Random Forest Regression**:
  - R-squared: 0.889190
  - RMS: 0.266736

  The Random Forest Regression model exhibits strong performance. It explains a large portion of the variation in the target variable, as indicated by the high R-squared value. The low RMS value suggests a small average deviation between the predicted and actual values, indicating accurate predictions.

In summary, the **K-NN**, **Linear Regression**, **Gradient Boosting Regression**, and **Random Forest Regression** models demonstrate varying levels of success. Among them, the **Random Forest Regression** model stands out with a high R-squared value and low RMS, indicating its superior ability to capture the underlying patterns and make accurate predictions.

3.5 Different Models

Building and selecting different models hold significant importance in the field of data analysis and machine learning. Exploring a diverse range of models enables the extraction of valuable insights and a deeper understanding of the underlying patterns and relationships within the dataset.

A crucial aspect of this study involves utilizing various features to construct predictive models. Each feature represents a distinct aspect or characteristic of the data, and by incorporating multiple features, a more comprehensive view of the phenomenon under investigation can be captured. This approach allows for the consideration of a wide range of factors that may contribute to the target variable, leading to more accurate predictions and deeper insights.

In this study, the R language was employed to create and evaluate different models. R provides a rich ecosystem of statistical and machine learning packages that facilitate the implementation of complex algorithms and techniques. However, due to the large size of the dataset, consisting of over 40,000 rows and 12 features, the computational performance of R tends to decrease. To address this challenge, a strategic decision was made to work with a smaller sample size of 1000 rows. This sample size enables the effective application of the 'ols_step_best_subset()' method and the evaluation of each model's performance.

Here is the output of 'ols_step_best_subset()' method:

```
> ols_step_best_subset(model1)
                                                                    Best Subsets Regression
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
Model Index    Predictors
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
     1         accommodates
     2         encoded_property_type accommodates
     3         encoded_property_type encoded_room_type accommodates
     4         review_Avg encoded_property_type encoded_room_type accommodates
     5         review_Avg encoded_neighbourhood encoded_property_type encoded_room_type accommodates
     6         olanaklar review_Avg encoded_neighbourhood encoded_property_type encoded_room_type accommodates
     7         olanaklar review_Avg encoded_neighbourhood encoded_property_type encoded_room_type bathrooms accommodates
     8         max_person_to_stay olanaklar review_Avg encoded_neighbourhood encoded_property_type encoded_room_type bathrooms accommodates
     9         max_person_to_stay bedrooms olanaklar review_Avg encoded_neighbourhood encoded_property_type encoded_room_type bathrooms accommodates
    10         max_person_to_stay bedrooms olanaklar review_Avg reviews_per_month encoded_neighbourhood encoded_property_type encoded_room_type bathrooms accommodates
    11         max_person_to_stay bedrooms olanaklar review_Avg minimum_nights reviews_per_month encoded_neighbourhood encoded_property_type encoded_room_type bathrooms accommodates
    12         max_person_to_stay bedrooms olanaklar review_Avg minimum_nights reviews_per_month is_anadolu encoded_neighbourhood encoded_property_type encoded_room_type bathrooms accommodates
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

                                                   Subsets Regression Summary
------------------------------------------------------------------------------------------------------------------------------------
                       Adj.        Pred
Model    R-Square    R-Square    R-Square      C(p)         AIC          SBIC          SBC          MSEP        FPE       HSP       APC
------------------------------------------------------------------------------------------------------------------------------------
  1        0.2676      0.2669      0.2645    380.4140    2150.4573     -688.6632    2165.1806    500.8655    0.5019    5e-04    0.7353
  2        0.3408      0.3321        -Inf    244.8440    2069.1544     -791.9538    2142.7707    451.2575    0.4577    5e-04    0.6631
  3        0.4110      0.4014        -Inf    114.9079    1962.5462     -902.1222    2050.8858    403.6041    0.4106    4e-04    0.5937
  4        0.4330      0.4232        -Inf     75.5621    1926.4782     -938.0217    2019.7256    388.9193    0.3960    4e-04    0.5727
  5        0.4565      0.4402        -Inf     33.4989    1908.2468     -977.8837    2060.3872    373.2122    0.3847    4e-04    0.5501
  6        0.4751      0.4588        -Inf      0.5178    1875.4018    -1010.2843    2032.4500    360.7949    0.3723    4e-04    0.5323
  7        0.4795      0.4629        -Inf     -5.8317    1868.9022    -1016.6188    2030.8581    358.1023    0.3699    4e-04    0.5289
  8        0.4828      0.4657        -Inf    -10.0601    1864.5144    -1020.8447    2031.3781    356.1815    0.3683    4e-04    0.5266
  9        0.4851      0.4675        -Inf    -12.2534    1862.1907    -1023.0220    2033.9621    355.0034    0.3675    4e-04    0.5253
 10        0.4869      0.4688        -Inf    -13.6236    1860.7022    -1024.3650    2037.3813    354.1252    0.3669    4e-04    0.5246
 11        0.4875      0.4689        -Inf    -12.8975    1861.3803    -1023.5802    2042.9673    354.0157    0.3672    4e-04    0.5249
 12        0.4876      0.4684        -Inf    -11.0000    1863.2739    -1021.6072    2049.7686    354.3370    0.3679    4e-04    0.5259
------------------------------------------------------------------------------------------------------------------------------------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

Figure 12: Sample OLS Results

Based on the provided model evaluation results, the researchers selected the 8th model as the preferred choice. This model demonstrated a relatively high level of performance with an R-Squared value of 0.4828.

The R-Squared value represents the proportion of the variance in the dependent variable (target variable) that can be explained by the independent variables (features) in the model. In this case, the 8th model accounts for approximately 48.28% of the variance in the target variable, indicating a reasonably good fit to the data.

Additionally, the model exhibited favorable values for other evaluation metrics such as C(p), AIC, SBIC, MSEP, FPE, HSP, and APC. These metrics provide insights into the model's complexity, goodness of fit, and prediction accuracy. For instance, the negative value of C(p) (-10.0601)

suggested that the model might be slightly overfit, but overall, the model's performance was considered satisfactory.

The decision to select the 8th model was based on a careful consideration of various factors. Firstly, its relatively high R-Squared value indicated a good level of explanatory power, suggesting that the selected features were capturing meaningful information. Additionally, the favorable values of evaluation metrics indicated that the model struck a balance between complexity and prediction accuracy.

It's worth noting that model selection is not solely based on evaluation metrics but also involves domain knowledge, contextual understanding, and research objectives. In this particular case, the 8th model demonstrated a combination of good performance and interpretability, making it a suitable choice for the analysis.

Overall, by selecting the 8th model, a reasonable level of predictive accuracy can be expected, along with an understanding of the underlying relationships between the selected features and the target variable.

Let's create new models using the 8th model so the features are "max_person_to_stay," "olanaklar," "review_Avg," "encoded_neighbourhood," "encoded_property_type," "encoded_room_type," "bathrooms," and "accommodates." Then the models outputs are:

Table 2: All Models

Model	R-squared	RMS
K-NN	0.510134	0.560831
Linear Regression	0.374351	0.633809
Gradient Boosting Regression	0.465987	0.585557
Random Forest Regression	0.836845	0.323663

- **K-NN:**
  - The R-squared value has slightly decreased from 0.521775 to 0.510134 in the new model, indicating a slightly weaker fit.
  - The RMS has increased from 0.554127 to 0.560831, implying slightly higher prediction errors.
- **Linear Regression:**
  - The R-squared value has slightly decreased from 0.381371 to 0.374351 in the new model, suggesting a slightly weaker relationship between the features and the target variable.
  - The RMS has increased from 0.630243 to 0.633809, indicating slightly higher prediction errors.
- **Gradient Boosting Regression:**
  - The R-squared value has decreased from 0.486660 to 0.465987 in the new model, indicating a weaker fit.
  - The RMS has increased from 0.574111 to 0.585557, implying higher prediction errors.

- **Random Forest Regression:**
  - The R-squared value has decreased from 0.889190 to 0.836845 in the new model, suggesting a weaker fit.
  - The RMS has increased from 0.266736 to 0.323663, indicating higher prediction errors.

Overall, the new model shows a decrease in performance compared to the old model for most of the algorithms. The R-squared values have generally decreased, indicating a weaker ability to explain the variance in the target variable. Additionally, the RMS values have mostly increased, indicating higher prediction errors in the new model. These differences suggest that the new model might not capture the underlying patterns and relationships in the data as effectively as the old model. Further analysis and adjustments may be necessary to improve the performance of the new model.

4 Conclusion

By familiarizing themselves with the Airbnb dataset and conducting a comprehensive analysis, the researchers applied a regression model. They examined the performance of other machine learning models such as K-NN, Gradient Boosting Regression, and Random Forest Regression. The main objective of the study was to enable homeowners in Istanbul with properties, whether they are houses, rooms, or hotels, to price them effectively. Additionally, they aimed to provide potential travelers coming to Istanbul with the ability to estimate prices based on their preferences for renting houses, rooms, hotels, etc.

Based on the analysis results, it was found that the 'price' variable in Istanbul can be explained by approximately 48% through the variables 'property_type', 'room_type', 'neighbourhood', 'max_person_to_stay', 'bedrooms', 'amenities', 'review_Avg', 'reviews_per_month', 'is_anadolu', 'price', 'minimum_nights', 'accommodates', and 'bathrooms'. Due to the relatively low explanatory power, it might be beneficial to explore different models or examine the variables more comprehensively by including more important variables. The Random Forest model was selected as the best-performing model, as it had a higher R-squared value and lower error metrics.

This study highlights that various methods and models can be used to predict Airbnb house prices, and it can serve as a guiding resource for future research endeavors to achieve better results.

References

[1] Airbnb Public Dataset (2023) "http://insideairbnb.com/get-the-data/"

[2] Kalehbasti, P.R., L.N. for M.L. (2021) "Airbnb price prediction using machine learning and sentiment analysis."

[3] Luo, Y., Zhou, X., and Zhou, Y. (2019) "Predicting Airbnb listing price across different cities"

[4] Yang, S. (2021) "Learning-based Airbnb Price Prediction Model. Proceedings - 2nd International Conference on E-Commerce and Internet Technology."

[5] Kadir S., Muhammed F.Ü., (2016) "Different Apple Varieties Classification Using kNN and MLP Algorithms."

[6] Gülden Kaya Uyanık, Neşe Güler (2013) "A Study on Multiple Linear Regression Analysis."

[7] Mark R. Segal, (2003) "Machine Learning Benchmarks and Random Forest Regression."

[8] Jerome H. Friedman, (2001) "Greedy Function Approximation: A Gradient Boosting Machine."