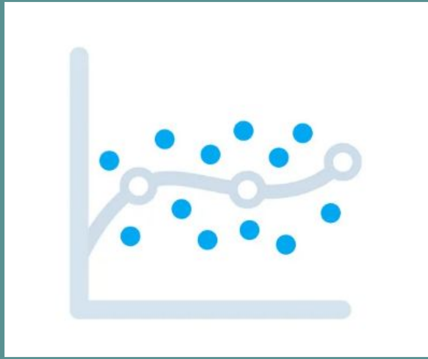
A Machine Learning-Based Price Prediction for Istanbul Airbnb Data

Oğuzhan Gündüz
Aslı Gültekin

CONTENTS

The aim of this study is to create a machine learning regression model to predict accommodation prices using Airbnb data in Istanbul. Within the scope of the project, the effects of various factors, such as location, date range, and number of bedrooms, on prices were examined using the Airbnb data source. After data collection and cleaning processes, the dataset was divided into training and test sets by selecting appropriate features. Various regression algorithms were tested, and the model that yielded the best results was selected. The model was validated on the test data, and the results were evaluated. As a result of the study, a model that can be used to predict Airbnb accommodation prices in Istanbul is proposed, and it is deemed to be a valuable resource for homeowners and accommodation seekers in Istanbul, as well as a guiding tool for future research endeavors.

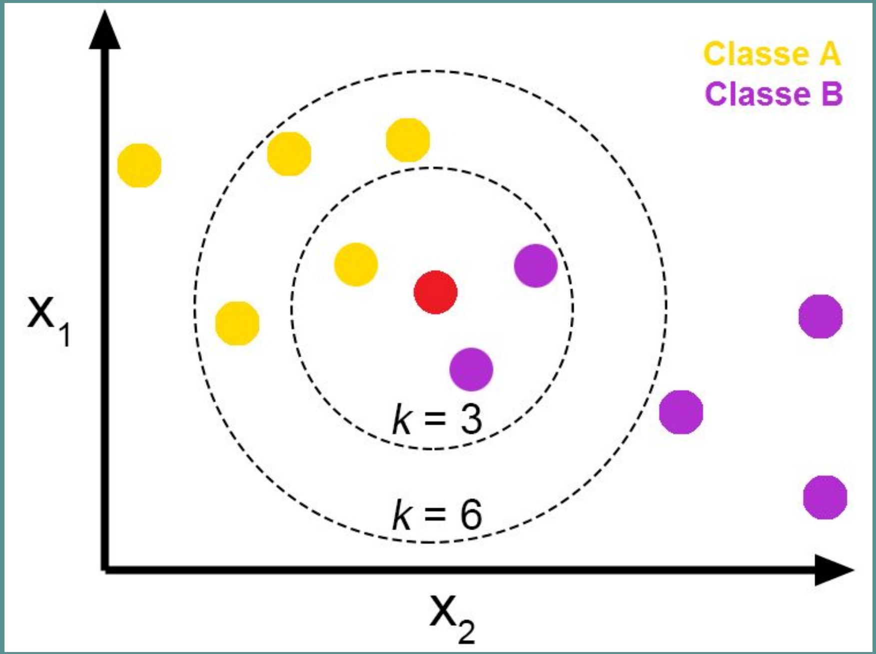
Keywords: Machine learning, Regression model, R-squared.

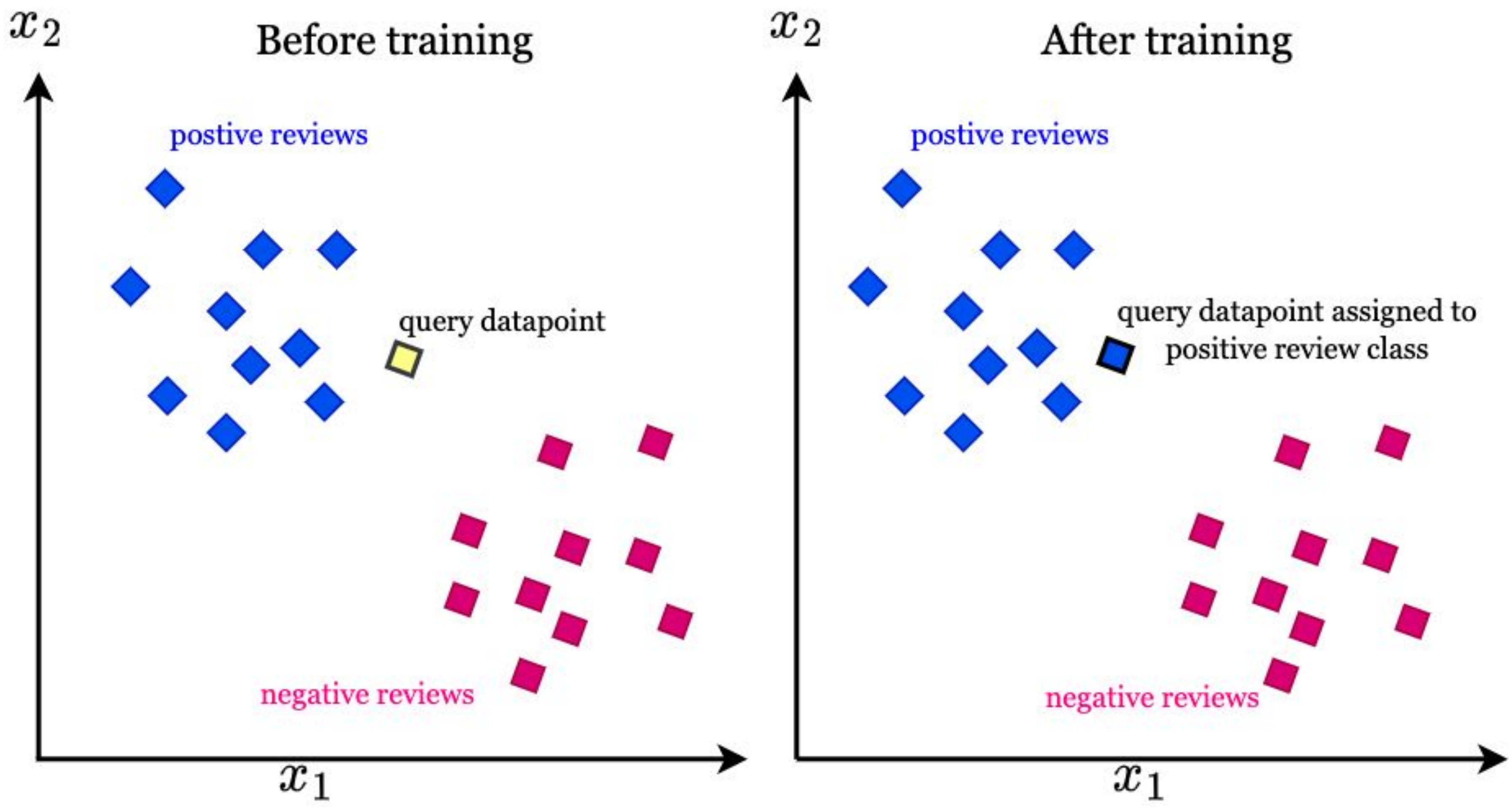


K-Nearest Neighbor Regression

K-Nearest Neighbors (KNN) is a popular and versatile algorithm used for classification tasks in machine learning. 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.





x_2
Before training
postive reviews
query datapoint
negative reviews
x_1
x_2
After training
postive reviews
query datapoint assigned to
positive review class
negative reviews
x_1

Multiple Linear Regression

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

Y-intercept Population slopes Random Error

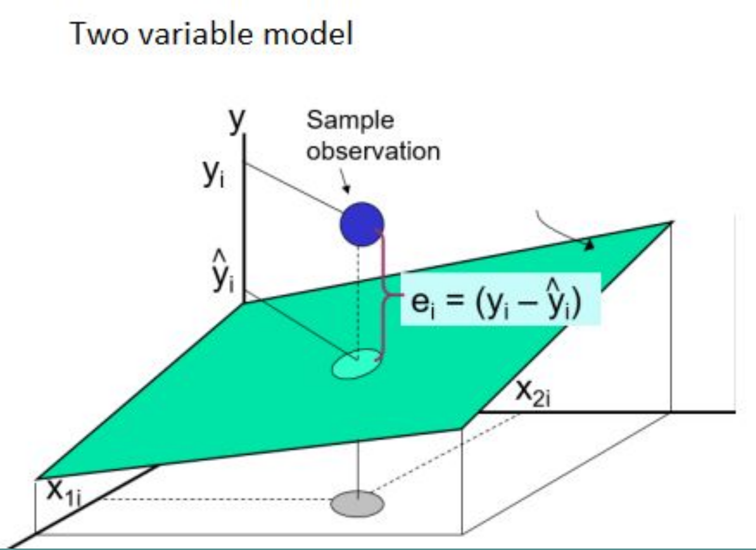
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

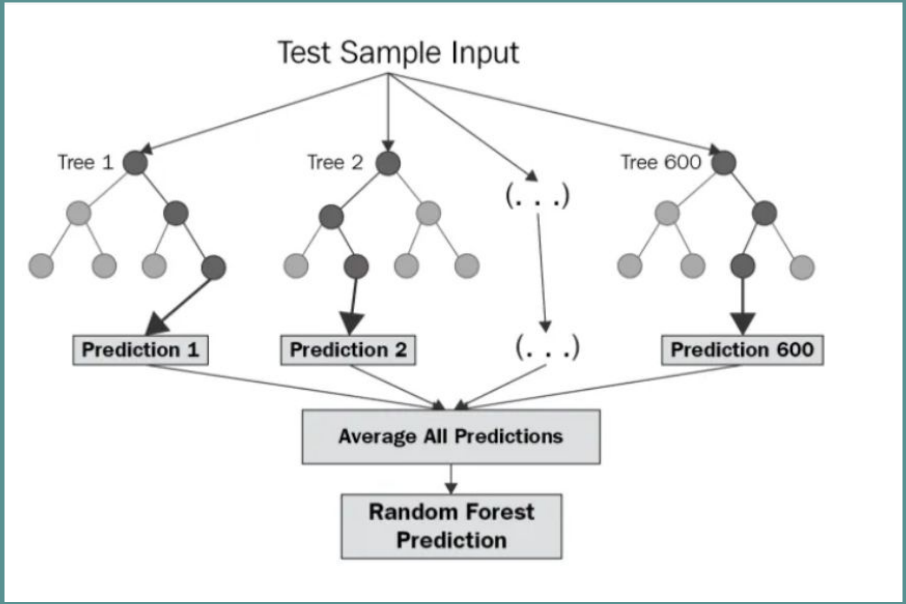
The parameters β0,β1,β3... are estimated using the least squares (OLS) method.



$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

To assess the quality of the linear regression model, the coefficient of determination (R2) is commonly used. It measures the proportion of the variance in the dependent variable that can be explained by the independent variable(s).

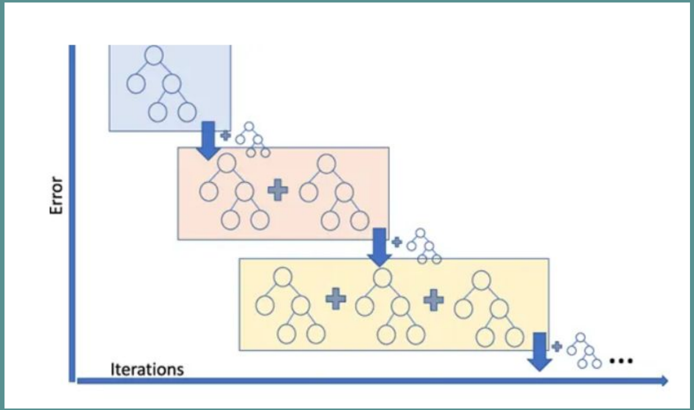
Random Forest Regression



Random Forest Regression is a supervised learning algorithm method. 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.

Gradient Boosting Regression

Gradient Boosting Regression, a popular machine learning algorithm, is used for regression tasks. 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. Gradient Boosting regression captures complex relationships and nonlinearity in the data, making it effective for price prediction.



APPLICATION

- ## 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.

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.

>>http://insideairbnb.com/get-the-data/

• Properties of Data

Some of the features used for predicting the price.

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.

• EDA(Exploratory Data Analysis) 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).

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

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.

amenities

['Essentials', 'Wifi', 'Bathtub', 'Dryer', 'Coffee maker: espresso machine, pour-over coffee', 'Game console: PS4', 'Backyard', 'Luggage dropoff allowed', 'Mosquito net', 'Drying rack for clothing', 'Barbecue utensils', 'Refrigerator', 'Public or shared beach access', 'Room-darkening shades', 'Kitchen', 'Hot tub', 'Microwave', 'Bed linens', 'Bread maker', 'Private entrance', 'Dedicated workspace', 'Hot water', 'Wine glasses', 'Baking sheet', 'Host greets you', 'Sound system', 'Iron', 'Cleaning products', 'Blender', 'Pocket wifi', 'Hangers', 'Childrenâ€™s books and toys for ages 5-10 years old', 'Exercise equipment: free weights', 'Long term stays allowed', 'Coffee', 'Clothing storage: closet', 'Cleaning available during stay', 'Hair dryer', 'Safe', 'Dishes and silverware', 'Books and reading material', 'Oven', 'Air conditioning', 'Dining table', 'Hot water kettle', 'Extra pillows and blankets', 'Heating', 'Shampoo', 'Cooking basics']

['Wifi', 'Dryer', 'Heating', 'Kitchen']

amenities
['Essentials', 'Wifi', 'Bathtub', 'Dryer', 'Coffee maker: espresso machine, pour-over coffee', 'Game conso
['Essentials', 'Wifi', 'Drying rack for clothing', 'Refrigerator', 'Ceiling fan', 'Coffee maker: Nespresso', 'Kit
['Microwave', 'Hangers', 'Fire extinguisher', 'Bed linens', 'Essentials', 'Wifi', 'Hot water', 'Bathtub', 'Patio
['Essentials', 'Wifi', 'Bathtub', 'Luggage dropoff allowed', 'Drying rack for clothing', 'Refrigerator', 'Room-
['Fire extinguisher', 'Essentials', 'First aid kit', 'Wifi', 'Dryer', 'Heating', 'Shampoo', 'Kitchen']
['Essentials', 'Wifi', 'Breakfast', 'Shampoo', 'Kitchen']
['Wifi', 'Dryer', 'Air conditioning', 'Hair dryer', 'Heating', 'Elevator', 'Kitchen']
['Essentials', 'Wifi', 'Patio or balcony', 'Drying rack for clothing', 'Refrigerator', 'Kitchen', 'Ocean view', 'B
['Essentials', 'Wifi', 'Air conditioning', 'Heating', 'Elevator', 'Kitchen', 'Indoor fireplace']
['Essentials', 'Wifi', 'Luggage dropoff allowed', 'Breakfast', 'Refrigerator', 'High chair', 'Room-darkening s
['Essentials', 'Window AC unit', 'Wifi', 'Dryer', 'Private patio or balcony', 'Luggage dropoff allowed', 'BBQ
['Free dryer â€“ In unit', 'Essentials', 'Wifi', 'Luggage dropoff allowed', 'City skyline view', 'Mosquito net
['Essentials', 'Wifi', 'Luggage dropoff allowed', 'Refrigerator', 'Kitchen', 'Microwave', 'Bed linens', 'Host g
['Wifi', 'Dryer', 'Heating', 'Kitchen']
['Essentials', 'Wifi', 'Refrigerator', 'Kitchen', 'Microwave', 'Single level home', 'Bed linens', 'Hot water', 'I
['Wifi', 'Air conditioning', 'Heating', 'Breakfast', 'Elevator', 'Kitchen']
['Kitchen', 'Wifi']
['Hangers', 'Host greets you', 'Essentials', 'Wifi', 'Hot water', 'Long term stays allowed', 'Luggage dropoff
['Essentials', 'Wifi', 'Luggage dropoff allowed', 'Refrigerator', 'Kitchen', 'Microwave', 'Host greets you', 'H
['Hangers', 'Fire extinguisher', 'Private entrance', 'Essentials', 'First aid kit', 'Wifi', 'Air conditioning', 'Hai
['Essentials', 'Wifi', 'Patio or balcony', 'Luggage dropoff allowed', 'Refrigerator', 'Kitchen', 'Microwave', 'I
['Hangers', 'Fire extinguisher', 'Bed linens', 'Essentials', 'First aid kit', 'Wifi', 'Dedicated workspace', 'Clot
['Essentials', 'Patio or balcony', 'Dryer', 'Luggage dropoff allowed', 'Drying rack for clothing', 'Refrigerato
['Essentials', 'Wifi', 'Patio or balcony', 'Drying rack for clothing', 'Refrigerator', 'Elevator', 'Kitchen', 'Bed l
['Essentials', 'Wifi', 'Window AC unit', 'Refrigerator', 'Kitchen', 'Microwave', 'Bed linens', 'Private entran
['Wifi', 'Dryer', 'Heating', 'Breakfast', 'Elevator', 'Kitchen']
['Coffee maker: pour-over coffee', 'Rice maker', 'Essentials', 'Wifi', 'Patio or balcony', 'Luggage dropoff a
['Essentials', 'Wifi', 'Luggage dropoff allowed', 'Refrigerator', 'Kitchen', 'Microwave', 'Bed linens', 'Host g
['Wifi', 'Dryer', 'Air conditioning', 'Heating', 'Kitchen']
['Elevator', 'Wifi', 'Dryer', 'Air conditioning', 'Heating', 'Gym', 'Pool', 'Kitchen', 'Indoor fireplace']
['Essentials', 'Wifi', 'Dryer', 'Luggage dropoff allowed', 'City skyline view', 'Refrigerator', 'Kitchen', 'Hot w
['Private patio or balcony', 'Wifi', 'Backyard', 'Dryer', 'Refrigerator', 'Kitchen', 'Dedicated workspace', 'Ho
['Essentials', 'Wifi', 'Window AC unit', 'Luggage dropoff allowed', 'Mosquito net', 'Refrigerator', 'Kitchen
['Hangers', 'Essentials', 'Wifi', 'Hair dryer', 'Heating', 'Shampoo', 'Iron', 'Kitchen', 'Elevator']
['Essentials', 'Wifi', 'Refrigerator', 'Kitchen', 'Microwave', 'Bed linens', 'Carbon monoxide alarm', 'Iron', '(
['Essentials', 'Wifi', 'Private patio or balcony', 'Luggage dropoff allowed', 'Drying rack for clothing', 'Roor

```python
class FeatureCounter:
    def __init__(self):
        self.data_selected = data_selected

    def count_features(self, feature_name, feature_list):
        self.data_selected[feature_name] = 0
        for index, row in self.data_selected.iterrows():
            amenities = row["amenities"]
            count = 0
            for feature in feature_list:
                if feature.lower() in amenities.lower():
                    count += 1
                amenities_list = eval(amenities)
                amenities_list = [item for item in amenities_list if feature.lower() not in item.lower()]
            self.data_selected.at[index, feature_name] = count
            self.data_selected.at[index, "amenities"] = str(amenities_list)

    def run(self, feature_lists):
        for feature_list in feature_lists:
            feature_name = feature_list[0]
            features = feature_list[1]
            self.count_features(feature_name, features)
        self.data_selected.to_csv("AirBnb_Data_Amenities.csv", index=False)



fc = FeatureCounter()

features = [['furniture_and_amenities', ['Chair', 'Clothing storage', 'Dedicated workspace', 'Elevator', 'Essentials', 'Iron', 'S
    ['entertainment_and_electronics', ['Bluetooth', 'Netflix', 'Sound system', 'TV']],
    ['kitchen_and_dining', ['Blender', 'Coffee maker', 'Cooking basics', 'Dishes', 'Fridge', 'Kettle', 'Kitchen', 'Microwave', 'C
    ['bathroom_and_toiletries', ['Shampoo', 'Bathtub', 'Bath', 'Body soap', 'Crib', 'Conditioner', 'Hair dryer', 'Hot tub', 'Hot
    ['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', ['Wifi', 'Ethernet connection', 'Long term stays allowed', 'View', 'Dryer', 'Pets allowed', 'Freezer']],
    ['pets_allowed', ['Pets allowed']],
    ['internet_options', ['Wifi', 'Ethernet connection']]
]



fc.run(features)
```

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.

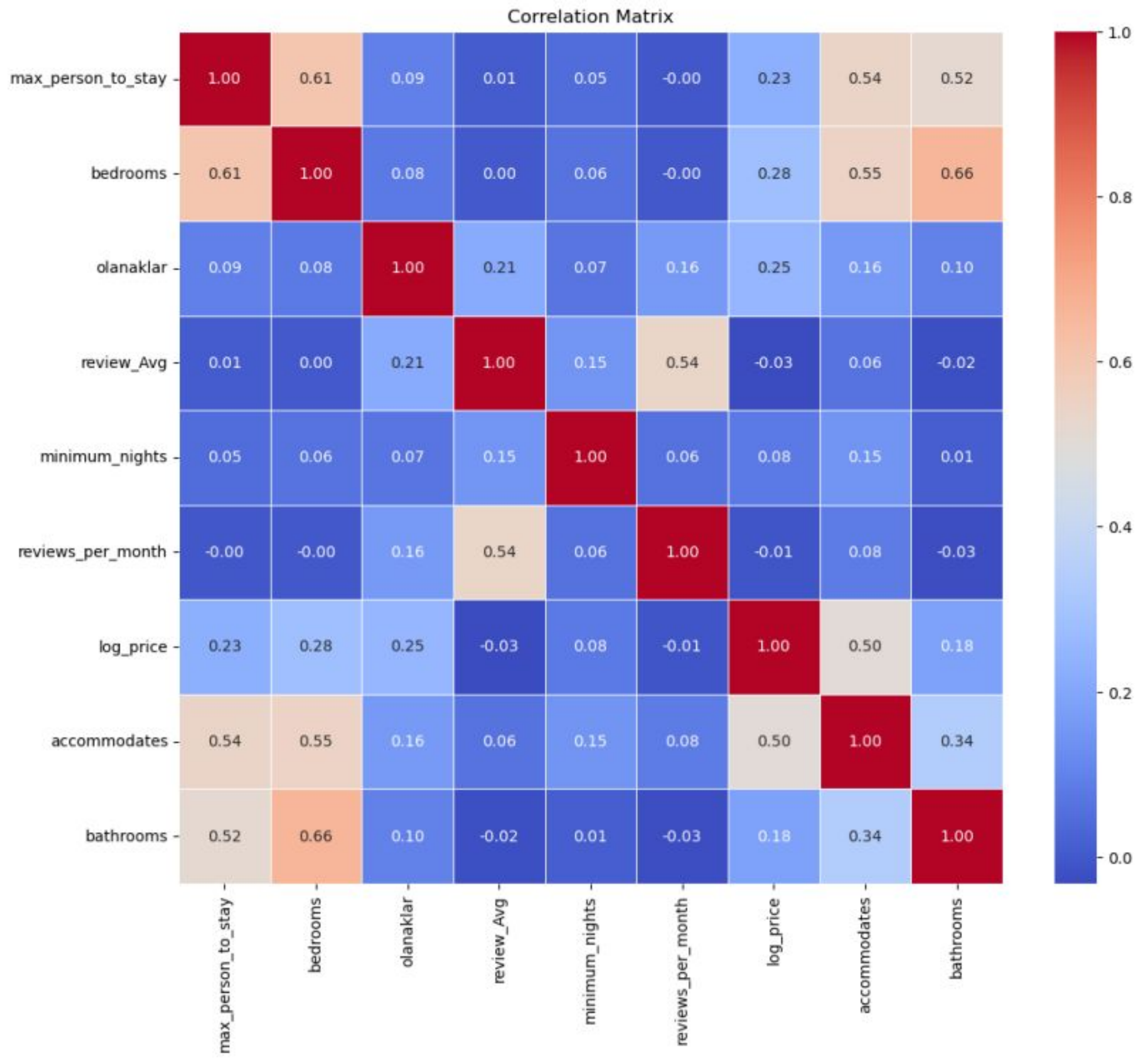
Upon matching these words, the corresponding category attribute is assigned a value of +1. Using this methodology, a total of 11 new features (columns) were derived by effectively parsing the "amenities" attribute.

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']

● The Outputs

Upon examining the correlation graph, a correlation is observed between

- the ’bedrooms’ and the ’max_person_to_stay’ variables,
- the ‘bedrooms’ and the ‘bathrooms’ variables,
- the ’reviews_per_month’ and the ’review_Avg’ variables,



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%.

The division into training and test sets provided valuable insights into the models' ability to generalize and their potential for real-world applications.

Twelve independent variables were employed to predict the logarithmically transformed 'price' value.

Four different machine learning algorithms, namely **KNN**, **Linear Regression**, **Gradient Boosting Regression**, and **Random Forest Regression**, were utilized to achieve the best possible prediction.

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

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

In summary, the **KNN**, **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.

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.

y, 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

Model	R-Square	Adj. R-Square	Pred 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
```

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.

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 New Models outputs are:

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

Old Models and New Models Comparison

Old 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

New 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

Test The Model with a Different Way

Here, the 10th data point(index 9) is removed from the training dataset then the models are tested on this removed data point. By doing so, we evaluate how well the models perform on unseen data, as the 10th data point was not included in the training process.

Extracted data and predicts:

```
Removed 10th data point: max_person_to_stay        1.000000
bedrooms                  1.000000
olanaklar                 1.250000
is_anadolu                0.000000
review_Avg                1.386294
minimum_nights            2.000000
reviews_per_month         0.060000
encoded_property_type    13.000000
encoded_room_type         2.000000
encoded_neighbourhood     9.000000
accommodates              2.000000
bathrooms                 1.000000
Name: 1962, dtype: float64
Removed 10th Y: 6.794586581
Removed 10th Y Real Value: | 893.0000001102862
--------------------------------------
--------------------------------------
KNN prediction: 6.496312618666667 | 662.693518542382
Linear regression prediction: 6.295509325044419 | 542.1318974400164
Gradient Boosting Regressor prediction: 6.475720462761978 | 649.1867740620306
Random Forest Regressor prediction: 6.942223495181671 | 1035.0691294226121
--------------------------------------
```

```
Removed 10th data point: max_person_to_stay        2.0
bedrooms                    1.0
olanaklar                   2.0
is_anadolu                  0.0
review_Avg                  0.0
minimum_nights              1.0
reviews_per_month           0.0
encoded_property_type      13.0
encoded_room_type           2.0
encoded_neighbourhood      32.0
accommodates                2.0
bathrooms                   1.0
Name: 19529, dtype: float64
Removed 10th Y: 7.170119543
Removed 10th Y Real Value: | 1299.9999994154837
--------------------------------------
--------------------------------------
KNN prediction: 6.835767691333333 | 930.5424459591657
Linear regression prediction: 6.570773649707399 | 713.921955046347
Gradient Boosting Regressor prediction: 6.647642887715317 | 770.964931208755
Random Forest Regressor prediction: 7.14112647422667 | 1262.8501563091265
--------------------------------------
--------------------------------------
KNN - R-squared: 0.5336989826731611
KNN - RMS: 0.5697520253381738
Linear Regression - R-squared: 0.3974059923608111
Linear Regression - RMS: 0.6476868810947463
Gradient Boosting Regressor - R-squared: 0.5804701222053898
Gradient Boosting Regressor - RMS: 0.5404233981752044
Random Forest Regressor - R-squared: 0.912614094946165
Random Forest Regressor - RMS: 0.246645610308442
```

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 88% through the variables ’property_type’, ’room_type’, ’neighbourhood’, ’max_person_to_stay’, ’bedrooms’, ’amenities’, ’review_Avg’, ’reviews_per_month’, ’is_anadolu’, ’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, Ne¸se 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.

Thank You

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.