# Analyzing Spatial Trends in Airbnb Prices Across Europe

1st Zeynep Acar
*Computer Engineering*
*Yildiz Technical University*
Istanbul, Turkey
zeynep.acar1@std.yildiz.edu.tr

2nd Betül Çelik
*Computer Engineering*
*Yildiz Technical University*
Istanbul, Turkey
betul.celik2@std.yildiz.edu.tr

***Abstract*—This essay explores the spatial trends in Airbnb prices across Europe and identifies the most favorable cities for hosting. Leveraging a dataset obtained from Kaggle, which includes details of Airbnb listings in different cities, machine learning techniques are employed to predict accurate prices and provide insights for hosts. Various data analysis techniques, including visualizations and correlation matrices, are utilized to understand the dataset and uncover significant factors prices. The data preprocessing stage involves cleaning the dataset, handling outliers, and transforming textual and categorical data into numeric representations. Feature selection is performed using the Random Forest Regressor algorithm, which reveals the most influential features for price predictions. Six machine learning algorithms, including Random Forest, Gradient Boosting, XGBoosting, Decision Tree, Support Vector Regression (SVR), and KNeighbors Regressor, are applied to the dataset.Several evaluation metrics employed to assess the performance of the models. The results indicate that the Random Forest algorithm outperforms the others. The study provides valuable insights for hosts to set competitive listing prices and make informed investment decisions. Ultimately, this research contributes to the understanding of Airbnb market dynamics and aids hosts in optimizing their rental income.**

***Index Terms*—Airbnb, machine learning, feature selection, dataset, preprocessing**

## I. Introduction

Airbnb is a company that provides accommodation services to travelers. This company offers people the opportunity to rent, lease, or host a place according to their needs. At this stage, house prices can vary significantly depending on the features of the house, its location, and the city it is located in. The platform has become increasingly popular in recent years. [1]

By leveraging the power of machine learning, hosts can gain a competitive edge in the Airbnb market. Accurate price predictions enable hosts to set their listing prices competitively, ensuring a balance between attracting guests and maximizing their rental income. Furthermore, understanding which cities command higher prices can inform hosts' investment decisions and help them identify lucrative markets to expand their Airbnb portfolio.

In this essay, we will be analyzing the spatial trends in Airbnb prices across Europe and identifying the most favorable cities for hosting. We will begin by discussing the methodology used to collect and analyze the data. Finally, we will present our findings and discuss the implications of our research.

## II. Data Set

The dataset that used for prediction analysis has been obtained from the Kaggle website. The dataset includes details of Airbnb listings in different cities. Each data point represents the details of an Airbnb listing in a city.

The dataset consist of 41.714 rows and 19 columns. The variables in the dataset include city, price, day, room type, whether it is a shared room or private room, person capacity, superhost status, multiple rooms availability, business reservation, cleanliness rating, guest satisfaction, number of bedrooms, distance to city center, distance to metro, attraction index, normalized attraction index, restaurant index, and normalized restaurant index.

This dataset will be used to understand various factors that influence the prices of Airbnb listings and make price predictions. By analyzing this information in the dataset, accurate price predictions can be made using machine learning algorithms, and better pricing strategies can be developed for Airbnb hosts.

### A. Analyzing the Dataset

In order to gain a better understanding of the dataset and visualize the distribution of its components, various graphs were used to gather information.

When we look at Fig 1. we can see that the distribution of airbnb prices in Europe by cities. While the prices in Amsterdam and Paris are more expensive than other cities, it is seen that the prices in Athens are cheaper.

Fig. 2 illustrates the distribution of prices based on the city and day. Although the day of the week does not have much effect, it is seen that the weekend is a little more expensive in general.

## III. Data Preprocessing

Data preprocessing is a critical step involving the preparation of the dataset in order to obtain accurate results in the analysis on Airbnb prices. This section includes processes such as cleaning the dataset, processing missing data, handling outliers. After examining the dataset, we found that the dataset

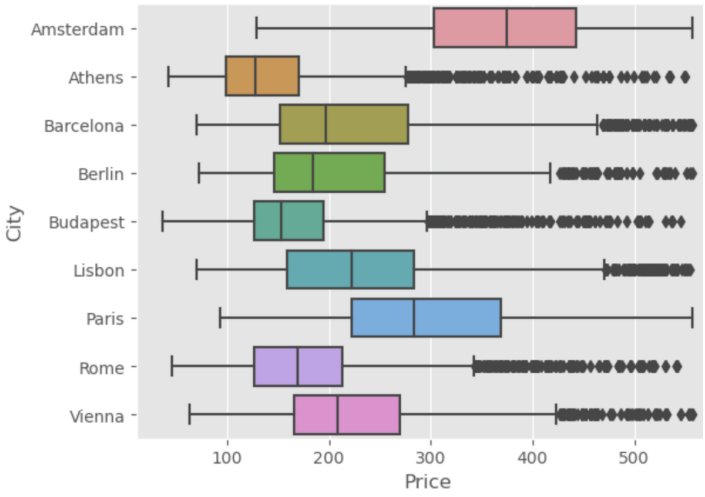


Fig. 1. Prices by Cities

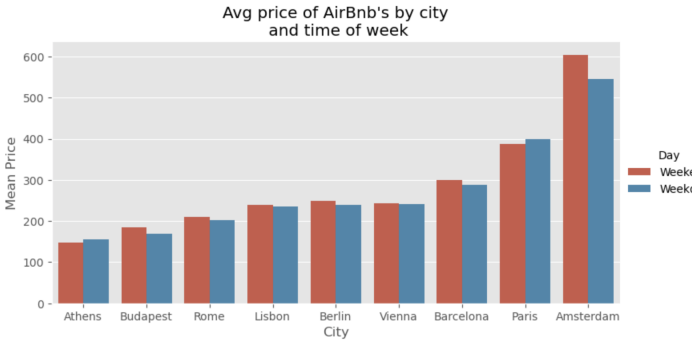


Fig. 2. Prices by City and Time of the Week

has no missing values. So that we did not have to apply any algorithm for filling missing values. We identified outliers in the dataset to improve the accuracy of our prediction results and we have removed them from the dataset. Price outliers shown in the Fig. 1.

When analyzing the dataset, we observed that alongside numeric values, there are also textual and categorical values present. In order to apply machine learning models to our training set, we transformed the textual and categorical fields into numeric values. By doing so, we ensured that all features in our dataset were represented as numeric values to facilitate the application of machine learning algorithms. To achive the conversion of textual and categorical data into numeric representations, we used the "LabelEncoder" from the sklearn library. This conversion process facilitated better analysis and modeling, allowing us to uncover patterns, relationships and ultimately achieve more accurate and effective models.

### A. Correlation Matrix

The correlation matrix is a useful tool for analyzing the relationships between variables in a dataset. It provides a visual representation of the correlation coefficients between

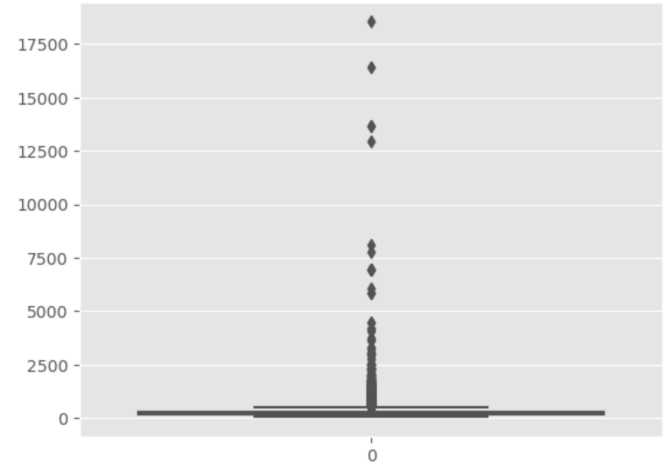


Fig. 3. Price Distribution

pairs of variables, indicating the strength and direction of their linear relationships.


Fig. 4. Correlation Matrix

The correlation matrix in Figure 4 shows that there is a high correlation between the City Center, the Attraction Index, and the Constraint Index.

## IV. Applying ML Algorithms to Data Set

In this section, we will first describe how feature selection is performed, followed by an overview of the models that will be applied to the dataset.

### A. Feature Selection

The feature selection process plays a crucial role in determining the relevant features that contribute significantly to the predictive model's performance. By identifying and selecting the most important features, we gain insights into the factors that drive the rent and understand which features are deemed

crucial by the model. [2] There are several approaches to feature selection. In this study we applied Random Forest Regressor algorithm.

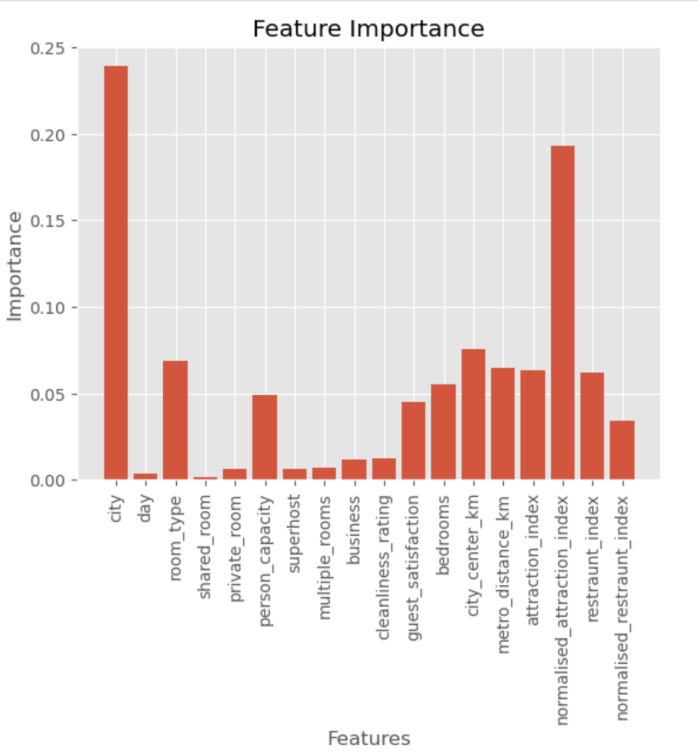


Fig. 5. Feature Importance

After processing the dataset using the Random Forest Regressor, we obtained a graphical representation as shown in Figure 2. This graph illustrates the results or insights derived from the model's predictions and their relationship with the dataset variables. By analyzing the graph we can gain insights into the importance of different features in the dataset and understand how they contribute to the model's predictions. As shown in the figure its clearly seen that the city and the normalised attraction index features has more affect on the dataset. And we chosed nine features based on the graph.

### *B. Selecting and Training ML Models*

Selecting and training the right machine learning (ML) model is a critical step in building effective predictive or analytical systems. ML models are designed to learn patterns and make accurate predictions from data, but the performance and suitability of different models can vary based on the nature of the problem, the characteristics of the dataset, and the desired outcomes.

For the evaluations six different algorithms were used. These algorithms are Random Forest, Gradient Boosting, XGBoosting, Decision Tree, Support Vector Regression (SVR), KNeighbours Regressor were assessed.

After evaluating different options within the hyperparameter space for the models, the following parameters were found optimal under the specified conditions. For the Random Forest algorithm the model parameter with the 100 tree estimators performed well. For the Graident Boosting 100 estimators, 0.1 learning rate and 5 max depth was chosen, similarly 200 estimators , 0.01 learning rate and 5 max depth were the best options for XGBoosting algorithm. For SVR algorithm default parameters were used which is epsilon is 0.1, C is 1.0 and the kernel type is rbf.

When applying the decision tree algorithm to the dataset max depth was chosen 7 to perform better and finally for the KNeighbours algorithm 5 neighbours and p value is 1 selected.

In this study Python programming language, Sci-kit Learn machine learning libraries were used for the model development.

### *C. Model Evaluation Criteria*

The performance of the trained ML models was evaluated using several metrics, including R2 (R-squared), MAE (Mean Absolute Error), MSE (Mean Squared Error) , MAPE (Mean Absolute Percentage Error), and RMSE (Root Mean Squared Error). These metrics provide valuable insights into the accuracy, precision, and overall goodness-of-fit of the models, allowing us to assess their performance and compare them against each other. By examining these evaluation metrics, we can make informed decisions about the effectiveness and reliability of the ML models in making predictions or solving the given problem.

R-squared (R2) measures the ability of a machine learning model to explain the variance in the dependent variable using the independent variables.

Mean Absolute Error (MAE) is a metric commonly used to measure the average magnitude of errors between predicted and actual values in a regression model. It provides a straightforward way to understand how far off, on average, the model's predictions are from the true values.

Mean Squared Error (MSE) measures the average squared difference between the predicted and actual values. It is calculated by taking the mean of the squared differences between each predicted value and its corresponding true value.

Root Mean Squared Error (RMSE) is the square root of MSE and provides a more interpretable measure since it is in the same unit as the dependent variable. It represents the average magnitude of the prediction errors.

Mean Absolute Percentage Error (MAPE) measures the average percentage difference between the predicted and actual values. It is calculated by taking the mean of the absolute percentage differences.

| Algorithms | $R^2$ | MSE | RMSE | MAE | MAPE |
|---|---|---|---|---|---|
| Random Forest | 0.78 | 2296.11 | 47.91 | 31.58 | 15.64 |
| Decision Tree | 0.55 | 4705.60 | 68.59 | 49.64 | 24.67 |
| XGBoost | 0.46 | 5616.83 | 74.94 | 50.66 | 21.92 |
| Gradient Boosting | 0.64 | 3775.33 | 61.44 | 44.61 | 22.17 |
| SVR | 0.01 | 10416.09 | 102.05 | 74.73 | 36.62 |
| KNeigbours | 0.70 | 3106.07 | 55.73 | 30.68 | 14.84 |

TABLE I
EVALUATION RESULTS

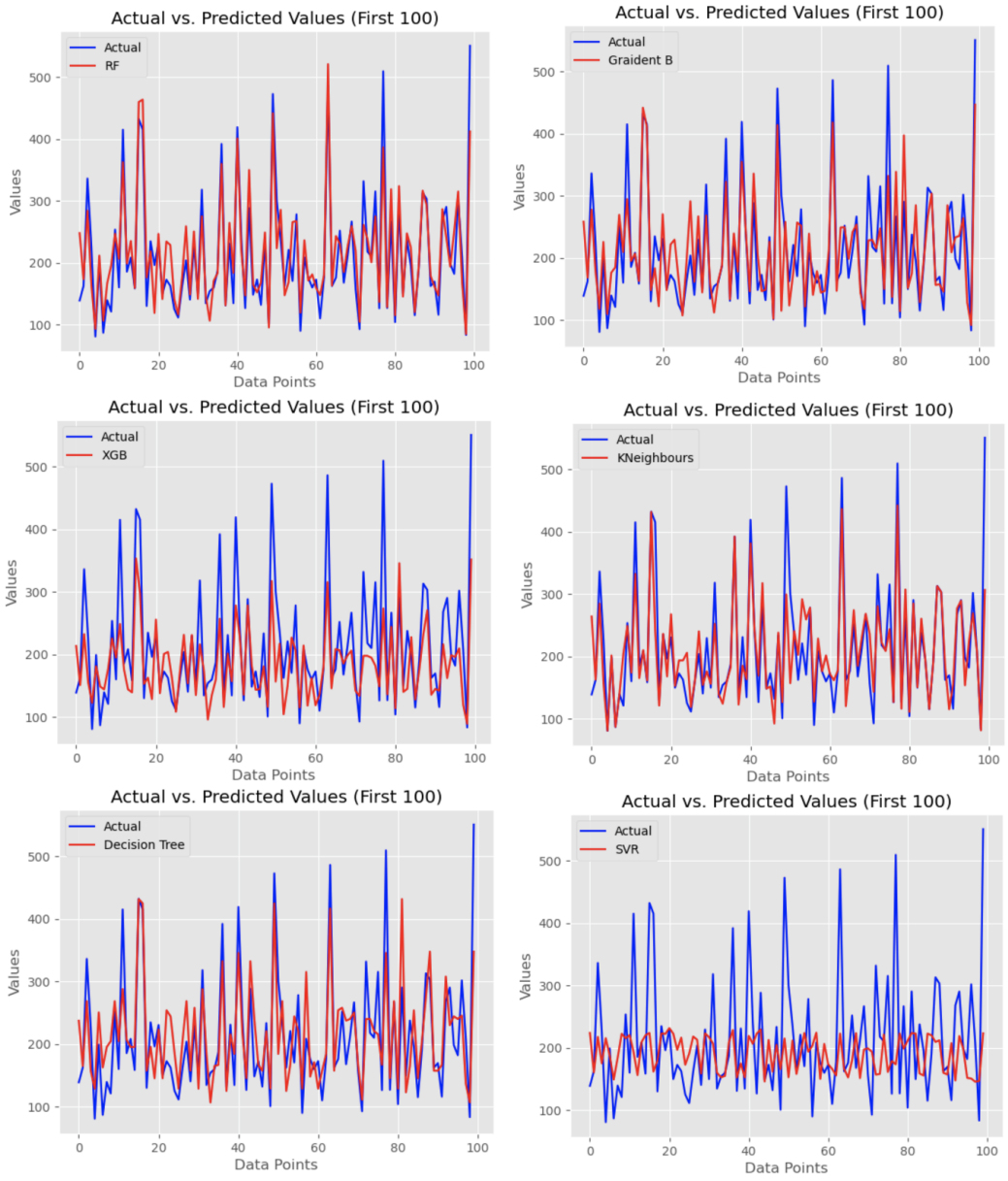


Fig. 6. Model Results

Figure 6 shows the graphs showing the comparison of the estimation results with the actual values obtained by applying the models. We can see that random forest and kneighbours algorithm performed better than the other algorithms from the Fig. 6 and Table 1.

## V. Conclusion

In this study, our objective was to develop a machine learning algorithm that could accurately predict house prices on Airbnb by considering various features. Additionally, we aimed to determine the optimal city for renting based on these predicted prices. By analyzing the characteristics of houses and estimating prices in different cities, we were able to identify the city that offers the highest rental prices. This information can be valuable for owners looking to maximize their earnings on the Airbnb platform. Our algorithm enables us to provide informed decisions and recommendations to owners, helping them find the most lucrative options available. Finally, according to the model results, we think that the random forest algorithm is better.

## References

[1] Brahmaih, Kala. (2020). Predicting Airbnb Listing Price Across NewYork. 10.13140/RG.2.2.28089.70246.

[2] Ma, Yixuan , Zhang, Zhenjiang , Ihler, Alexander , Pan, Baoxiang. (2018). Estimating Warehouse Rental Price using Machine Learning Techniques. International Journal of Computers Communications and Control. 13. 235-250. 10.15837/ijccc.2018.2.3034.