Price Prediction Model for the AirBnB Offerings based on Location

**Aim and Objectives**

1. **AIM:** Airbnb is a home-sharing platform that allows home-owners to put their properties online, so that guests can pay to stay in them. Pricing a property and evaluating the price for a property are challenges that owners and customers of Airbnb rentals face on a daily basis. The aim of this report is to create a model for predicting the price of an Airbnb listing using property specifications, owner information. Owners and customers can use the resulting model to estimate the expected value of an Airbnb listing. Linear regression, tree-based models,Boasting based models and Bayesian Ridge are trained and tuned on a dataset of Airbnb listings from New York city 2013, and the resulting models are compared in terms of Accuracy score, Root mean square error and R-square.
2. **Objective :** 
   1. To gather and preprocess the dataset listing activity and metrics in New York City (NYC) for the year 2019 .
   2. Implement feature engineering to identify and select location based attributes and other factors that influence Airbnb rental prices in New York City.
   3. To develop a reliable price prediction model using machine learning techniques to help both the property owners and the customers with price evaluation given minimal available information about the property.
   4. To train ,evaluate and compare the performance of each algorithm using established metrics such as Accuracy Score, Root Mean Square Error and R-squared value to determine the best and most reliable model for price prediction.

**Literature Review**

Previously, Yu and Wu [3] tried to implement a location based price prediction using feature analysis along with linear regression, SVR and Random Forest regression. They also attempted to classify the prices into 7 classes using Naive Bayes, Logistic Regression, SVC and Random Forest. They declared a best RMSE of 0.53 for their SVR model and a classification accuracy of 69% for their SVC model with PCA. In another paper, Ma et al.[8] have applied Linear Regression, Regression Tree, Random Forest Regression and Gradient Boosting Regression Trees to analyzing warehouse rental prices in Beijing. They concluded that the tree regression model was the best-performing model with an RMSE of 1.05 CNY/m2-day.

Zervas et al conducted research showing the role of geographical and environmental features in influencing travelers decisions and property owners pricing strategies. The findings show that factors such as proximity to the city, safety and accessibility influence the perceived value of Airbnb properties, which in turn affects the specific location of their price dynamics. Li et al. [6] studied a clustering method called Multi-Scale Affinity Propagation and applied Linear Regression to the obtained clusters in an effort to create a price prediction model for Airbnb in different cities. He took the distance of the property to the city landmarks as the clustering feature. Paper by Chiny et al. [10], Zhou and Tong [4], Trang et al. [5], Kokasih and Paramita [14], and Ma et al. [13] are recent publications around rental prices in the lodging industry. This research has tried to improve and add to the experimented methods from the literature by focusing on a variety of feature selection techniques, implementing Neural Networks.

Guttentag's research focuses on the impact of neighborhood characteristics and geographical positioning on the demand and pricing of Airbnb accommodations. This study implements decision tree algorithms for analyzing dataset and identifying significant decision points, enabling the identification of key neighborhood features affecting pricing dynamics. The existing literature collected demonstrates the critical role of location-specific attributes in establishing Airbnb's product pricing strategy and market dynamics. While some studies have highlighted the importance of proximity to materials and sites, the development of powerful predictive models that combine these factors will help participants make informed decisions and optimize pricing strategies in a short-term competitive environment. important.

**Methodology**

## New York City Airbnb Open Data, the dataset was collected from kaggle. This dataset includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.To preprocess the data, we carefully examined each feature in the dataset. Our pre-processing steps involved the following:

## Remove features with frequent and irreparable missing fields.

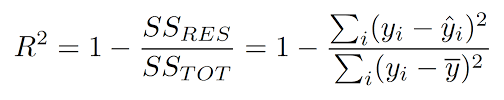
## We also eliminated constant-valued fields and duplicate features.

* We converted the categorical values of the features into integer values to easily train the dataset using different regression approaches.

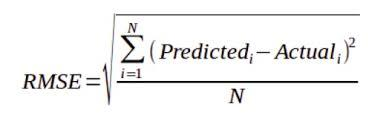
Regression is a supervised machine learning technique which is used to predict continuous values. The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data. The three main metrics that are used for evaluating the trained regression model are variance, bias and error. The 6 algorithms used to determine the best and most reliable model for price prediction are:

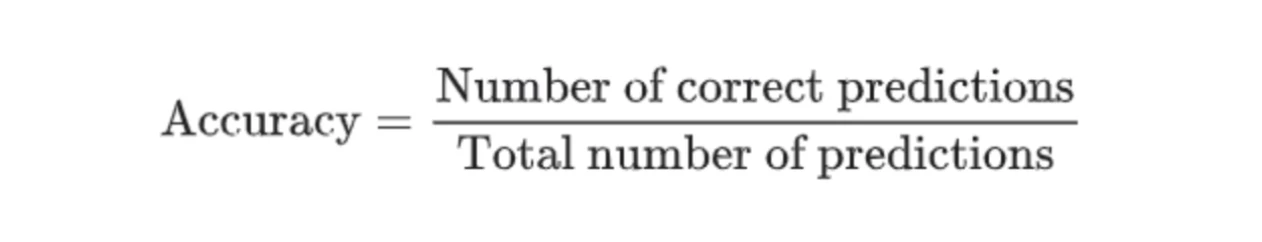
1. Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. Linear regression is used to understand the impact of various location-specific attributes on Airbnb rental prices in New York City (NYC).
2. Decision Tree is a supervised machine learning algorithm that can be used for both classification and regression tasks. Decision Trees can be prone to overfitting.
3. Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mean prediction of the individual trees for regression tasks.
4. Gradient Descent is an iterative optimization algorithm used to minimize the cost function or error in machine learning models.Gradient Descent used in consideration with various regression algorithms to optimize model parameters and determine the optimal weights for the different features affecting Airbnb rental prices in New York City (NYC).
5. A binary classifier was developed, using a traditional machine learning approach, with Python’s implementation of XGBoost (Extreme Gradient Boosting). It is a boosting algorithm based on an ensemble of decision trees.
6. Bayesian Ridge regression is a probabilistic model that applies the principles of Bayesian inference to estimate a regression model with a linear relationship between the features and the target variable. Bayesian Ridge regression is used to understand the uncertainties associated with the relationship between location-specific attributes and Airbnb rental prices in New York City (NYC).

To assess the performance of the regression techniques, we employed evaluation metrics for an evaluation of the model’s performance. We calculated metrics Accuracy score,Root Mean Square Error. and R-Square. R-squared is a statistical measure that represents the goodness of fit of a regression model. The value of R-square lies between 0 to 1. Where we get R-square equals 1 when the model perfectly fits the data and there is no difference between the predicted value and actual value.

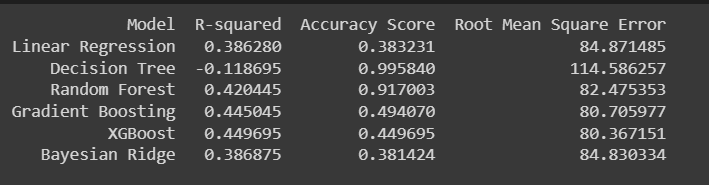


The Root Mean Squared Error (RMSE) is one of the two main performance indicators for a regression model. It measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value (accuracy).

The Accuracy score is calculated by dividing the number of correct predictions by the total prediction number.



All six algorithms are compared and evaluated using these metrics, and the results



The model faces challenges related to overfitting or underfitting, leading to a decreased ability to generalize to new data or an oversimplified representation of the complex pricing dynamics within the Airbnb market. Poor-quality data, including missing values or outliers, brings a huge impact on the model's performance and accuracy which leads to biased predictions or unreliable insights.

**Conclusion**

Based on the findings it is concluded that Decision Tree has an excellent Accuracy score with 99.58%, but considering overall performance in negative R -square shows a high biased or high variance data model. Whereas other algorithms Linear Regression, Gradient Boosting, XGBoost and Bayesian Ridge accuracy lie between 38% to 50%. The best fit model which performs well with all three evaluation metrics is Random-Forest Regressor with an accuracy of 91% and RMSE score of 82.47%.

Furthermore, It can be observed that the price values ranged from approximately 100 to 2000+. The Neighbourhood\_group values, defined by integers, mostly lie between 0 and 6, indicating that the majority of the listings in the dataset belonged to these groups. However, fewer occurrences of Neighbourhood\_group values 9 and 10 were observed in a specific price range of about 200 to 500. The Neighbourhood\_group values of 0–8 exhibited a wider distribution of prices, covering a range of approximately 100 to 2000+.

The following are some of the main recommendations for future work: A more sophisticated, streamlined process to update/expand the new data set - LM-SA-2020.; Improve the model generalization capability by increasing the size of the data set.; there is great potential to improve the location-based Airbnb product price prediction model through the integration of technologies such as deep learning and natural language processing (NLP). Using deep learning,the model can capture non-social and physical patterns in the New York City Airbnb market, thus improving its predicted capabilities. Additionally, the inclusion of NLP methods allows models to analyze data from visitors and write descriptions to better understand the benefits associated with price dynamics. Moreover, the integration of real-time data and dynamic pricing strategies would enable the model to adapt to the evolving market trends and make data driven decisions.

**References**

1. AirBNB public dataset. <http://insideairbnb.com/get-the-data.html>. Accessed 01 Dec 2018
2. [Ningxin Peng](https://ieeexplore.ieee.org/author/37088872789); [Kangcheng Li](https://ieeexplore.ieee.org/author/37089269444); [Yiyuan Qin](https://ieeexplore.ieee.org/author/37088874138), “ [Leveraging Multi-Modality Data to Airbnb Price Prediction | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/9434659) “ , 2021
3. Yu, H., Wu, J.: Real estate price prediction with regression and classification. CS229 (Machine Learning) Final Project Reports (2016)
4. Zhou, X., Tong, W.: Learning with self-attention for rental market spatial dynamics in the Atlanta metropolitan area. Earth Sci. Inf. 14(2), 837–845 (2021)
5. Trang, L.H., Huy, T.D., Le, A.N.: Clustering helps to improve price prediction in online booking systems. Int. J. Web Inf. Syst. (2021)
6. Li, Y., Pan, Q., Yang, T., Guo, L.: Reasonable price recommendation on AirBNB using multi-scale clustering. In: 2016 35th Chinese Control Conference (CCC), pp. 7038–7041. IEEE (2016)
7. Ma, C., Liu, Z., Cao, Z., Song, W., Zhang, J., Zeng, W.: Cost-sensitive deep forest for price prediction. Pattern Recogn. 17, 107499 (2020)
8. Ma, Y., Zhang, Z., Ihler, A., Pan, B.: Estimating warehouse rental price using machine learning techniques. Int. J. Comput. Commun. Control 13(2) (2018)
9. Chiny, M., Bencharef, O., Hadi, M.Y., Chihab, Y.: A client-centric evaluation system to evaluate guest’s satisfaction on AirBNB using machine learning and NLP. Appl. Comput. Intell. Soft Comput. 2021 (2021)
10. [Lu Jiang](https://arxiv.org/search/cs?searchtype=author&query=Jiang,+L), [Yuanhan Li](https://arxiv.org/search/cs?searchtype=author&query=Li,+Y), [Na Luo](https://arxiv.org/search/cs?searchtype=author&query=Luo,+N), [Jianan Wang](https://arxiv.org/search/cs?searchtype=author&query=Wang,+J), [Qiao Ning](https://arxiv.org/search/cs?searchtype=author&query=Ning,+Q),” [[2301.01222] A Multi-Source Information Learning Framework for Airbnb Price Prediction (arxiv.org)](https://arxiv.org/abs/2301.01222) ”, 2023
11. Kokasih, M.F., Paramita, A.S.: Property rental price prediction using the extreme gradient boosting algorithm. IJIIS: Int. J. Informat. Inf. Syst. 3(2), 54–59 (2020)
12. Hall, M.A., Smith, L.A.: Practical feature subset selection for machine learning (1998)