THE COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS MULTIPLE REGRESSION, XG BOOST AND SVM WITH RESPECT TO RESIDENTIAL ASSET PRICE

Mrs Nidhi, Assistant Professor
Bharati Vidyapeeth's Institute of Management and Information Technology,
Navi Mumbai (India)
mca.nidhipoonia@gmail.com

Saurabh Dnyaneshwar Kathe, Student,
Bharati Vidyapeeth's Institute of Management and Information Technology,
Navi Mumbai (India)

Swapnil Sunil Patil, Student,
Bharati Vidyapeeth's Institute of Management and Information Technology,
Navi Mumbai (India)

ABSTRACT
Tourism is a vital cog in the growth of any country as it is increasing day by day in countries like India. Due to this hotel and Residential asset, stays demand is also increasing rapidly and the best affordable price decision-making for owners plays a very important role. Price prediction and analysis of residential assets are important topics to research in the Indian economy. Existing research papers mostly focus on macroeconomics affecting the prices of residential assets. Here we will focus on micro factors and detailed information about the asset. It can be helpful for an organization in two ways, on one side, it enables space owners to list their space and earn rental money and on the other side, it helps tourists for accessing rented private homes. This paper helps with the increasing competition which reduces prices for customers with better services. It promotes tourism in a region. The prices are predicted based on various factors such as location, neighbourhood, etc. We used various algorithms (multiple regression, XG boost, support vector machine) to predict prices and compare the best one according to error rate.
Keywords: Residential asset price prediction, multiple regression, XG boost, support vector machine modal

Introduction

Tourism is the backbone for the health of any economy, it boosts job creation, foreign currency earnings, infra development, culture and regional development. Technological innovations are improving every sphere of human life, tourism is no exception. To make optimum utilization of resources technology has an important role to play. Technological advancements are upgrading the experience for tourists as well as hotel owners. As tourism is increasing in India, many organizations have promoted tourism in such a way that people can easily rent their residential assets online which can be rented by tourists. These are cost-efficient and even a middle-class family can afford such apartments for a night. In the past, people relied on simple data analysis for calculating the budget but nowadays data science has unlocked the potential for studying complex businesses. Machine learning algorithms work to filter out data or live-streamed which helped to increase the result for the decision-making process. Data collection is very important as we rely on the data for more accurate predictions. Various techniques can be applied to the vast chunk of data to churn own the best possibilities for all stakeholders and parties. One could use traditional regression algorithms or even the latest machine learning models which are better at drawing predictions. Various factors affect the housing prices like the neighbourhood, location, ratings, nearby destinations to visit, cost per night, minimum night's availability, and the number of reviews. The other important factors include the characteristics of the house like the room type (private, public), facilities available, etc. For better decisions and predictions and deciding prices for these assets, we should use different machine learning algorithms. Here we will apply different approaches and try to find out which ML algorithm is best suitable for predicting the residential prices of a property for one day or more days for renting purposes. In our research, we will use multiple regression, XG boost and support vector machine modal and test which is giving the best result for predicting their prices by finding the error rate. For our research first will review the literature on residential prices in tourism and machine learning algorithms then will do a visualisation of data collected from secondary sources according to features and properties of data and after that will apply ML algorithms for prediction and analysis according to error rate. In the end, will compare the error rate and draw results which is a suitable algorithm for studying the research problem of residential asset prices data. In the last section, will draw a conclusion according to the result analysed.

Literature Review

Luo (2019) residential prices are analyzed and predicted according to the micro area by using SVM and random forest modal and prices are predicted according to the pool, area, etc. not by using traditional methods. Studied residential assets and understood how residential asset prices vary from region to region. Studied regression algorithm and how it can be implemented.

Kalehbasti et al. (2019) prices of rental property prediction using a model using deep learning, machine learning modal support-vector regression (SVR), neural networks (NNs) and others. The paper comparison between different algorithms of machine learning and is analysed for property prices of rental using different models.

Alfiyatin et al. (2017) predict housing values for the future according to the concept, physical conditions and location. And predict error for proved combination regression. Shehhi et al. (2020) predict hotel prices by four modal SARIMA, adaptive network fuzzy interference SVM machine model and Boltzmann machine for GCC cities.

Shamim (2022) Machine learning algorithm are used to predict stock prices with lower error rates paper aim to use various techniques for prediction and analyse which algorithm is best for stock price prediction.

Tziridis et al. (2017) predict airfare charges compared modals on eight states' air flight data and decide the prices according to factors that affect prices. In this research paper, we get a better understanding of the factors affecting the prices of plane tickets and how those factors can help in predicting the prices.

Erguven et al. (2012) Evaluation of the effectiveness of primary multiple linear regression in addition to constituent analysis for the Education and Science Ministry of Georgia.

He et al. (2005) support vector machines (SVMs) inverse problem is investigated and the inverse problem is divided into two clusters such that the margin between the two clusters for a given dataset.

Ogunleye et al. (2019) where modal XGBoost for the subject of high-performance chronic kidney disease the optimized and XGBoost by using its decision tree approach gives the best result using GPU.

Hu et al. (2017), Predict prices according to product review by using the frequency, regency and monetary (RFM) model and in the paper prices are predicted by using reviews given by users.

Mitchell (2017), Paper talks about the XGBoost algorithm working, it works using the graphics processing unit and forms a decision tree and its performance is very high on different datasets and its speed could be increased by using a higher version processor.

Li (2019), In the paper XGBoost gives the best prediction having the lowest error rate as compared to linear regression, D-Gex and KNN for Gene profiling and RNA-seq. It is a model of multiple trees hence it predicts gene expression perfectly for the given datasets of health problems.

Dong (2020), here in the paper talks about the prediction of electrical resistivity measurement using XGBoost algorithm and the algorithm gives a satisfactory result according to the fitting line as compared to values of RMSE.

Mo (2019), here in the paper XGboost algorithm acts best in predicting the window behaviour of a residential building which requires Heating Ventilation and Air Conditioning best performance.

Research objective

The chief objective of the paper for research is mentioned below:

1) To visualize also analyze residential price data with respect to various factors like neighbourhood, price, and location.
2) To study multiple regression approaches with respect to residential asset prices.
3) To study boosting technique (XG Boost) approach with respect to residential asset prices.
4) To study Support Vector Machine in order to predict better output and better prices on residential assets.
5) Compare the Multiple Regression, Support Vector Machine, XGBoost with respect to identification of better residential asset prices with better accuracy rate and minimum error rate.

Problem statement

As the tourism industry is growing day by day, the prices of hotels have been increasing. Due to which the middle-class people are facing renting issues when travelling. In order to understand the tourism prices and the factors affecting those prices we are analyzing the data of residential asset pricing. It can help us to understand major factors affecting the prices for example if the asset is located near a good Neighbourhood or bad neighbour or a beach etc. This research can benefit in promoting tourism as well as to set the best prices for a rented residential asset for a night.

Research Methodology

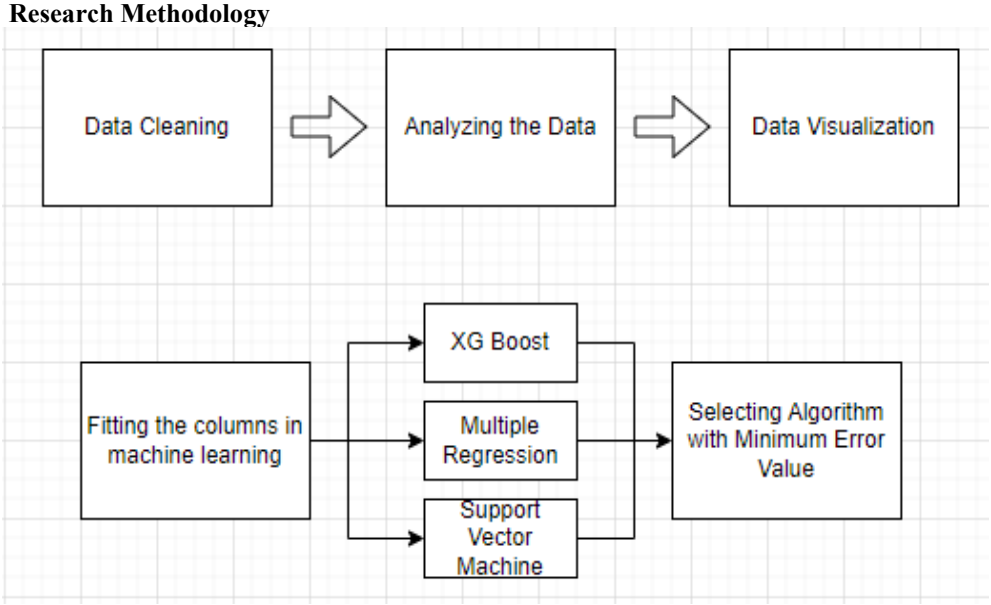

Figure 1: System Flow (source: https://ieeexplore.ieee.org/)

In Figure 1, Nguyen (2017), explains, the diagram represents a workflow of the research work. This diagram shows the separate steps of a process in sequential order. It helps others how a process is done. First, we remove unfitting, matching and inadequate data inside the dataset. After merging several data sources, there are likelihoods for data to be repeated or mislabelled and then thoroughly smearing logical and statistical techniques to define and estimate data. We can show information trends, graphically and weight patterns. It benefits the reader to attain rapid understanding. Then Multiple Regression, XG Boost and SVM models are used to train on the dataset.

The present study consists of four distinctive stages: (1) the assortment of the residential asset features that impact the prices, (2) to train and test the applied ML models on the group of adequate residential asset data (3) the choice of the regression ML models being compared and (4) investigational assessment of the ML models. Each phase of processing is discussed as follows:

Phase 1: - In this phase, the utmost enlightening structures of an asset that regulate the prices are fixed and it defines the problem which we are solving in the paper under. For all assets, the subsequent structures were measured:

Here is the list of Features we define as F1, F2 so on:
F1: Room type
F2: Ratings
F3: Nearby destinations to visit.
F4: Cost per night
F5: Facilities available
F6: Longitude
F7: Latitude
F8: Neighbourhood
F9: Neighbourhood Group
F10: Minimum nights
F11: Availability

F12: Number of reviews

Phase 2 (Collection of Data) – It is phase where we focused on the prediction of a single asset price. For trials, a set of asset data (From the Airbnp dataset from Kaggle) for every asset the features are (F1 to F12) were composed from the Web manually.

Phase 3 (ML Models Selection) – For current study ML models were selected and applied to the same data. i.e. Multiple Regression, Xgboost, and Support Vector Machine ML models

Phase 4 (Evaluation) – The residential asset data collected in phase 2, were used to train the mentioned ML models. The prediction accuracy indices are used here (error rate between the desired and predicted prices)

Data Analysis And Visualization

```
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x207902583d0>
```

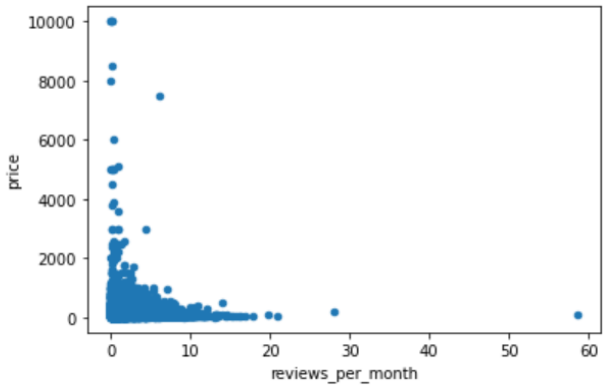

Figure 2: Price by Reviews

In Figure 2 We see that as the reviews are less the prices are more. Most tourists cannot afford costly homes. That's why they prefer affordable houses. Reviews are more where the price is less because Most of tourist thinks about why they spend a lot of money on a rental house where they only go to sleep at night.

```
<matplotlib.axes._subplots.AxesSubplot at 0x207908e5d30>
```

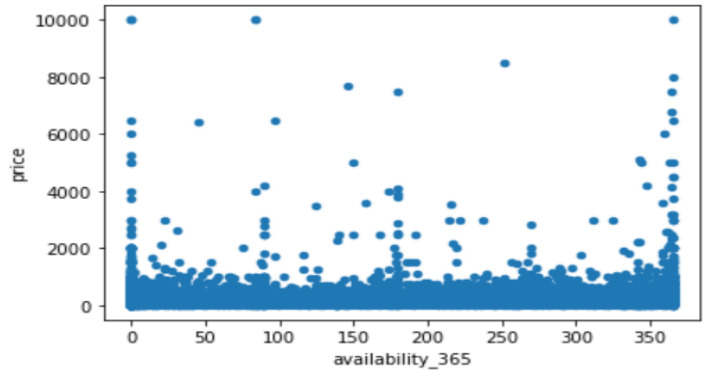

Figure 3: Price by Availability

In Figure 3, both the data columns are independent. There is no relation between price and availability but when one or more dimensions are added there might be a pattern occurring and this dimension can be added through ML models.

Dataset is a collection of data. This data helps to analyze the trends and hidden patterns and make decisions based on the dataset. These records in the dataset are organized in a way how we plan to access the information. Every column relates to a particular variable and every row relates to a given member of the data set.


```
In [68]: ml_data['price'].hist()
Out[68]: <matplotlib.axes._subplots.AxesSubplot at 0x2bb4ae26b80>
```

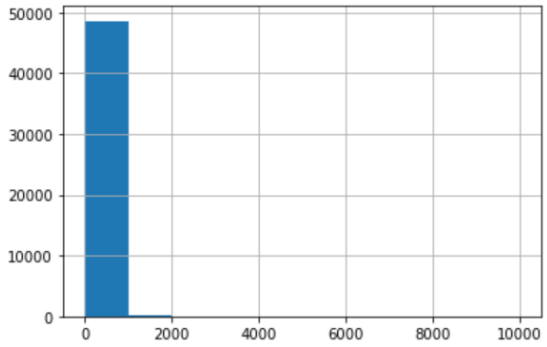

Figure 4: Price Range

Figure 4 represents the price range. This graph indicates which category has the maximum house rent.

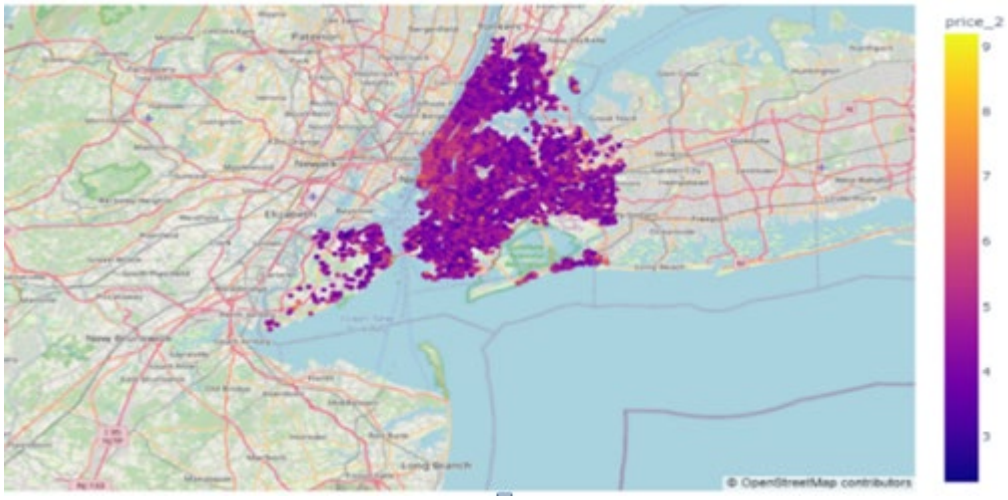

Figure: 5 Locations and Price

In Figure 5, the Visual Representation shows the location and price of homes. In this map, the dark blue colour shows the lower price of house availability and the yellow colour shows the higher price of house availability.

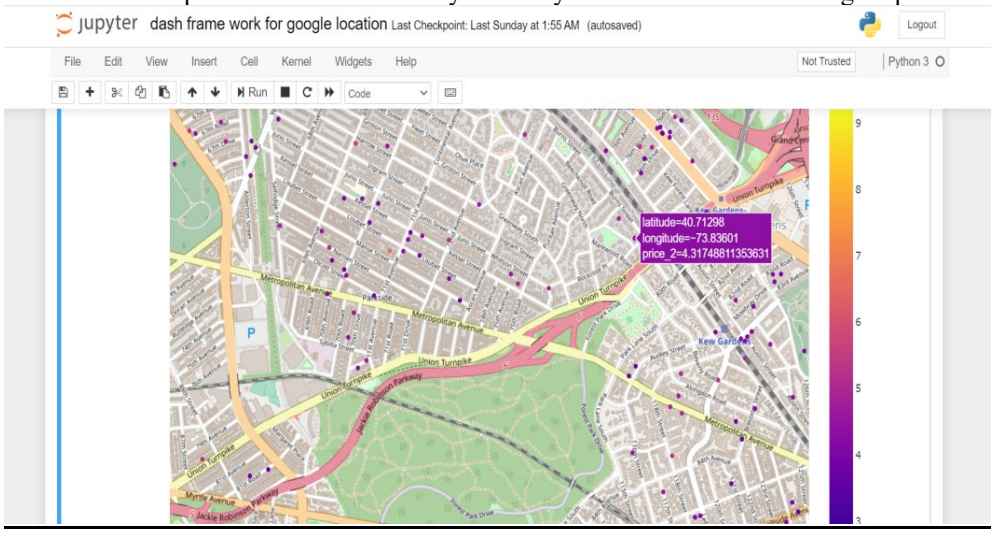

Figure 6: Location

The above Figure 6 is to get a better idea about the location which is preferable.

Result And Findings

Multiple Regressions: The relation between a single dependent variable and several independent variables can be analysed by using the technique of multiple regressions. Multiple regression analysis is a technique by using the independent variables whose values are known to predict the value of the single dependent value. The value of every predictor is weighed, where the weights denote their relative contribution to the overall prediction. Multiple regressions help to compare various columns such as price, neighbourhood, availability, reviews, etc. at the same time which draws a line in multiple dimensions which can only be done through machine learning.

XG Boost: It is a machine learning algorithm which helps to convert a weak tree into a strong tree. It makes use of boosting techniques. XG Boost is a framework that can run on multiple languages. So you can very well run XG Boost in R, Python, etc. XG Boost is very much a platform free means it is portable. So you can run on Windows, Linux, iOS, etc. XGBoost gain so much popularity because of two main things: It's the speed of processing and the kind of result or output it gives.

SVM: Supervised learning algorithm Support Vector Machine is majorly used for Classification and its aim is to create the finest decision boundary or line which can segregate space into classes of n-dimensional. The hyper plane is the top decision boundary and SVM uses a hyper plane for choosing the extreme points/vectors. The following tables give the error value of each machine learning algorithm used on the residential asset data:

Comparative analysis of Multiple regression, Support Vector Machine, XG Boost algorithms error rate with respect to identification of better residential asset prices	
Algorithm	**Error Rate**
Multiple Regression	60
XG Boost	37
SVM	26

Table 1 Comparative analysis of multiple regressions, SVM, XG Boost algorithms error rate with respect to identification of better residential asset prices

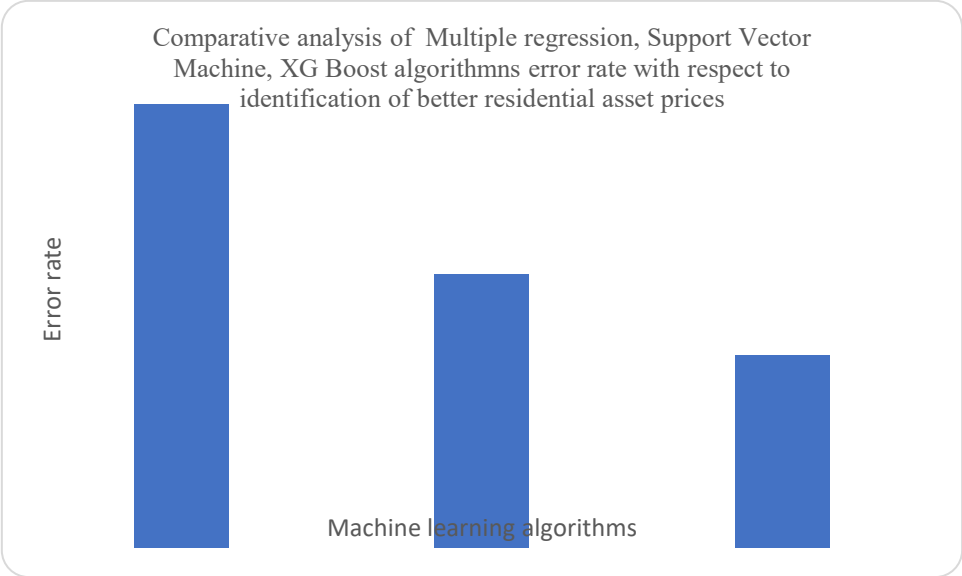

Figure 7: Error Rate Graph for Machine Learning Algorithms for Given Dataset

In Figure 7 the graph shows the error rate of machine learning algorithms regression, XG Boost and SVM using the Residential price dataset.

Conclusion:

The paper predicts charges of Residential assets using regression, XG Boost and SVM machine learning approaches. We conclude that Residential asset prices vary based on various factors like the type of rooms as well as the location. The investigational outcomes demonstrate that for predicting asset rents and prices ML models are suitable tools. The locality and the neighbourhood affect the most in-house pricing. The Comparative Analysis of Machine learning modal multiple regression, Support Vector Machine, and XG Boost algorithms error rate concerning the identification of better residential asset prices shows that SVM is the best method as the error rate is 26. This research can help Residential owners to set the house rent for tourism. From this paper, they get an idea about the range of asset rent according to its location. We conclude that this research paper can help an organization to improve tourism. This research can help organizations to predict prices and add competition between organizations. It can help the tourist to buy places at the best prices and can afford to stay near tourist spots. We hope this can be helpful in the future.

References

Alfiyatin N., Febrita R., Taufiq H., & Mahmudy W. (2017), "Modelling house price prediction using regression analysis and particle swarm optimization", Malang, East Java, Indonesia. International Journal of Advanced Computer Science and Applications, 8(10).

AlShehhi, M., & Karathanasopoulos A. (2020), "Forecasting hotel room prices in selected GCC cities using deep learning", Journal of Hospitality and Tourism Management.

DongW., Huang Y., LehaneB., & Ma, G. (2020) , "XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring", Automation in Construction, 114, 103155.

Erguven M. (2012), "Comparison of the efficiency of principal component analysis and multiple linear regressions to determine students' academic achievement", In 2012 6th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-5). IEEE.

He Q., & Chen F (2005), "The inverse problem of support vector machines and its solution", In 2005 International Conference on Machine Learning and Cybernetics (Vol. 7, pp. 4322-4327). IEEE.7)

Hu H., Chen K, & Lee J. (2017), "The effect of user-controllable filters on the prediction of online hotel reviews" Information & Management, 54(6), 728-7

Luo Y. (2019, December), "Residential Asset Pricing Prediction using Machine Learning", In 2019 International Conference on Economic Management and Model Engineering (ICEMME) (pp. 193-198). IEEE.

Li, W., Yin Y., QuanX.& Zhang, H. (2019),"Gene expression value prediction based on XGBoost algorithm", Frontiers in genetics, 10, 1077.

Mitchell, R. & Frank, E. (2017), "Accelerating the XGBoost algorithm using GPU computing", PeerJ Computer Science, 3, e127.

Mo, H., Sun, H., Liu, J., & Wei, S. (2019), "Developing window behavior models for residential buildings using XGBoost algorithm", Energy and Buildings, 205, 109564.

Nguyen& Dong(2017),"Joint network coding and machine learning for error-prone wireless broadcast", 10.1109/CCWC.2017.7868415

Kalehbasti P., Nikolenko L., & Rezaei, H. (2019), "Airbnb price prediction using machine learning and sentiment analysis", arXiv preprint arXiv:1907.12665.

Shamim R. (2022), "Machine learning's algorithm profoundly impacts predicting the share market stock's price", IJFMR-International Journal For Multidisciplinary Research, 4(5)..

Tziridis K., Kalampokas T., Papakostas, G., & Diamantaras, K. (2017, August), " Airfare prices prediction using machine learning techniques.",In 2017 25th European Signal Processing Conference (EUSIPCO) (pp. 1036-1039). IEEE.

Ogunleye A., & Wang Q. (2019)., "XGBoost model for chronic kidney disease diagnosis", IEEE/ACM transactions on computational biology and bioinformatics, 17(6), 2131-2140.44.

