International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-11 Issue-3, January 2022

Housing Price Prediction with Machine Learning

Amena Begum, Nishad Jahan Kheya, Md. Zahidur Rahman

Abstract*: For socioeconomic development and the well-being of citizens, developing a precise model for predicting housing prices is always required. So that, a real estate broker or a house seller/buyer can get an intuition in making well-knowledgeable decisions from the model. In this work, a various set of machine learning algorithms such as Linear Regression, Decision Tree, Random Forest are being implemented to predict the housing prices using available datasets. The housing datasets of 506 samples and 13 feature variables from January 2015 to November 2019 were taken from the StatLib library which is maintained at Carnegie Mellon University. Since housing price is emphatically connected to different factors like location, area, the number of rooms; it requires all of this information to predict individual housing prices. This paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models to explore various impacts of features on prediction methods. This paper will also provide an optimistic result for housing price prediction by comprehensively validating multiple techniques in model execution on regression.

***Keywords*: Prediction, Machine Learning, Linear Regression, Random Forest, Decision Tree.**

I. INTRODUCTION

Data is the core of machine learning. Predictive models utilize data for training which gives some degree precise outcomes. Machine learning includes developing these models from data and using the data to predict new data. It has automation improvement in its ways to make our work easier s the world is pushing ahead to using variants technologies.

An accurate forecast on the house price is important to prospective homeowners, appraisers, developers, tax assessors, investors, and other real estates market participants, such as mortgage lenders and insurers. The price prediction of a traditional house is based on cost and sale price comparison lacking an accepted standard and a certification process. So, a model can help in this regard to forecast house price and to improve the efficiency of the real estate market. There has been a proliferation of empirical studies analyzing residential property values over the last two decades, with Ball (1973) [1] being the last major study. By emphasizing attributes of property value for example, such as housing site, quality, location, and environment, each succeeding research has generally improved the predictive power of the models. The growing interest rates in U.S. real estate market slowed down the market from the starting of 2005. This weakened the asset values, increased the slowing down of the economy combined with the subprime mortgage crisis, and resulted in a sharp decline in the housing prices; which ultimately caused a global crisis [2]. This continued until the end of 2011, the housing market fell especially in the large cities after the 2008 global crisis. At the starting of 2012, the market followed an upward trend with naturally increasing prices, increasing demand, and decreasing inventories. This made market analyzers and economists to be focused on more precise prediction models to shield the economy from predictable threats. The predictable threats could cause economic downturns [2]. With increasingly better results, nowadays machine learning methods have been used in prediction and changed the economic landscape. Practically every economic area presently profits from machine learning forecast models, and the current models are turning out to be more precise given the computational power accessible for handling enormous arrangements of data. In this work, the housing price problem is analyzed by utilizing several machine learning methods such as Linear regression, Decision tree, Random forest regression.

II. LITERATURE REVIEW

Trends in housing prices are a worry to the buyers and sellers and also indicate the current economic situation. There are many factors that have an impact on house prices, such as the number of bedrooms and bathrooms, per capita crime rate by town, proportion of residential land, property tax-rate, pupil-teacher ratio, etc. House price relies on its location too. A house with great accessibility to malls, employment opportunities, highways, schools, would have a notable price as compared to a house with no such accessibility. Predicting house prices are generally not very accurate and a difficult task, hence for housing price prediction, there are many systems developed.

Breiman (2001) [3][4] proposed Random Forest (RF) approach at first by blending bootstrap aggregation and classification and regression tree. It is an ensemble classifier to obtain better prediction performance that utilizes various models of decision trees. This method utilized a bootstrap technique and created many trees to train each of the tree of original sample set of training data.

Manuscript received on January 20, 2022.
Revised Manuscript received on January 01, 2022.
Manuscript published on February 28, 2022.
* Correspondence Author
Amena Begum, Department of ICT, Comilla University, Cumilla , Bangladesh. Email: amenacou@gmail.com
Nishad Jahan Kheya, Department of ICT, Comilla University, Cumilla, Bangladesh. Email: kheya3331@gmail.com
Md. Zahidur Rahman*, Department of CSE, Britannia University, Cumilla, Bangladesh. Email: mzahidur.bd@gmail.com

This technique reduced the correlations between the generated regression trees, along with the randomly chosen features partitioned at each node. RF Regression has become a commonly used tool in multiple prediction scenarios due to their high accuracy and ability to handle large features with small samples [5].

In 2009, Li et al. [6] had utilized SVR, Support Vector Regression method first to forecast house values in China. In 2016, Rafiei and Adeli also used SVR [7] to conduct a study to determine whether a property developer should stop the construction or build another one at the beginning of a project based on the prediction of future housing values.

Using linear regression, Sifei Lu et al. (2017) [8] had proposed an advanced hybrid regression technique for house prices prediction. They combined Gradient boosting, Ridge and Lasso regression together to build a hybrid model; and achieved a significant result.

In 2018, Wang et al. [9] employed almost 30k housing assessment price data from Virginia USA, and recommend that Random Forest outperforms Linear Regression in terms of accuracy.

Based on a set of features, such as building age and floor area, number of beds, floor level, Mohd et al. (2019) [10] predicted housing prices in Petaling Jaya, Selangor, Malaysia by utilizing several machine learning algorithms such as Ridge Regression, Random Forest, Decision Tree and found that in terms of overall accuracy, Random Forest is the most preferred one, as evaluated by the root mean squared error (RMSE).

In 2020, Winky et al. [11] proposed Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM) to examine a data sample of about 40,000 housing transactions. Their study showed that the SVM method performed well since it can deliver sensibly exact forecasts within a tight time constraint.

A. Contribution of Researchers in this work

Currently, people are more comfortable with automated systems rather than manual systems. The aim of this work is to provide an efficient way to avoid the hassles of customers. Currently, to suggest suitable showplaces for customer's investments, the customer visits a real estate agent. But this approach is quite risky as the agent may be a fraud; also his forecasting may lead to loss of customer's investment. So, this manual approach currently used in the market has a high risk and is outdated. In this paper, researchers will show how three different machine learning algorithms namely Linear Regression, Decision Tree, and Random Forest Regression can help in housing prices prediction. The dataset used here was taken from the StatLib library which is maintained at Carnegie Mellon University; contained various essential parameters. Finally, exploration will be performed to find out the performances of these proposed methods and find the best one.

III. MATERIALS AND METHODS

The architecture of the system is depicted in Fig. 1.
The detailed description of system's modules will be provided in subsection A, B, C, D, and E.

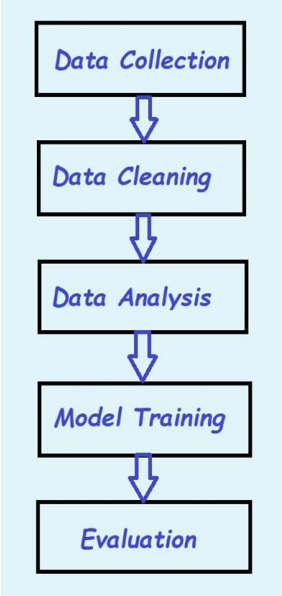

Fig.1.System Architecture

A. Data Collection

The dataset of housing contains information about different houses in Boston; originally a part of UCI Machine Learning Repository. This dataset for housing price detection was taken from the StatLib library [12] which is preserved at Carnegie Mellon University. The creators are Harrison, D. and Rubinfeld, D.L. The dataset consists of 506 samples and 13 feature variables. The goal is to anticipate the value of prices of the house utilizing the given features.

B. Data Cleaning

Raw data need to be processed as it has missing data, categorical data, etc. In this regard, we employ a data cleaning procedure. To take care of missing attributes we have three options:

- Get rid of the missing data points: If the missing data points are less, then remove the data points but if the missing data points are large then it should set a value for those.
- .Get rid of the whole attribute: Deleting the whole attribute containing the missing data points. But If the attribute contains a strong positive or negative correlation then the attribute should not
- Set the value to some value: Addition of zero, mean or median to the missing data points. The entire dataset has been cleaned up and after the completion of data cleaning, we separate the whole dataset into two parts - one part is for training the model called train set and another for testing the models, called test set. Train set has 376 data elements, while test set has 95 data elements.

C. Data Analysis

Before building a regression model, exploratory data analysis is an important step. Thusly, researchers can find the implicit patterns of the data, which in turn assists with picking suitable machine learning methods. Here, we mainly perform three operations - Correlation Analysis, Feature Scaling, and Pipelining.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

C.1. Correlation Analysis

Here correlation technique is used for investigating the relationship between two quantitative variables. Through the correlation analysis, one evaluates the correlation coefficient that tells how much a variable changes with respect to the other. It provides a linear relationship between two variables. The correlation of MEDV, RM, ZN, and LSTAT [13] with MEDV, RM, ZN, and LSTAT is shown in Fig. 2.

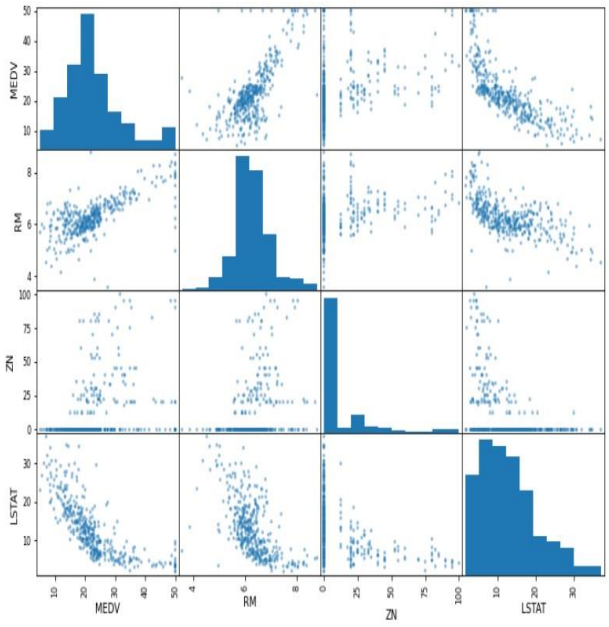

Fig. 2. Correlation of MEDV, RM, ZN, LSTAT with MEDV, RM, ZN, LSTAT [13]

The relationship between two variables is generally considered strong when their ***r*** value is larger than 0.7, and the relationship between two variables is generally considered strong when their ***r***-value is near 0. The correlation ***r*** estimates the strength of the linear relationship between two quantitative variables. A correlation of 0.10 is a weak positive correlation while a correlation of -0.97 could be a strong negative correlation.

C.2. Feature Scaling

Primarily, two types of feature scaling methods employed here.

- **Min-max scaling (Normalization):** In this way all values will be presented in a scale. That means all values will fall between 0 to1.
- **Standardization:** Here, we perform standard deviation.

C.3. Pipelining

In order to automate and codify the workflow, a machine learning pipeline is a way to produce a machine learning model. The pipelines of machine learning consist of several sequential steps that perform everything from preprocessing and data extraction to model deployment and training. We make pipeline in such way so that we can easily change it as per our requirements.

D. Model Training

Various methods can be used to predict the house price. In our work, we propose three models: Linear Regression, Decision Tree, and Random Forest Regression to see which one gives better performance.

D.1. Linear Regression

Linear Regression is a machine learning algorithm that performs a regression task. It is mostly used between variables and forecasting for finding out the relationship. Fig. 3 represents the workflow of the proposed Linear Regression method. Linear regression performs the task which is shown in Fig, that a given independent variable (x) is used to predict a dependent variable value (y). So, it finds out a linear relationship between x (input) and y (output) using Linear Regression.

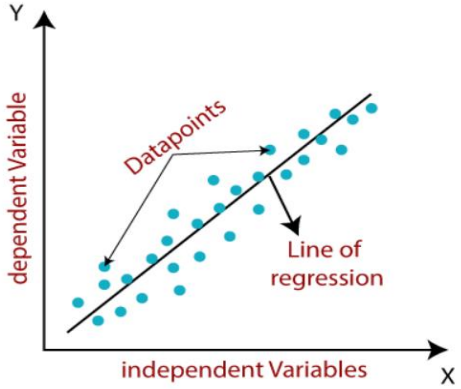

Fig. 3. Linear Regression

D.2. Decision Tree

Decision Tree is one of the most commonly used application, it can be used to solve both Regression and Classification tasks. It runs completely through the entire tree with a particular data point until it reaches the leaf node of the tree by answering True/False questions. The average value is the final prediction of the dependent variable in that particular leaf node. Through multiple iterations, for the data point, the Tree can predict a proper measure. Fig. 4 represents the workflow of the proposed Decision Tree method.

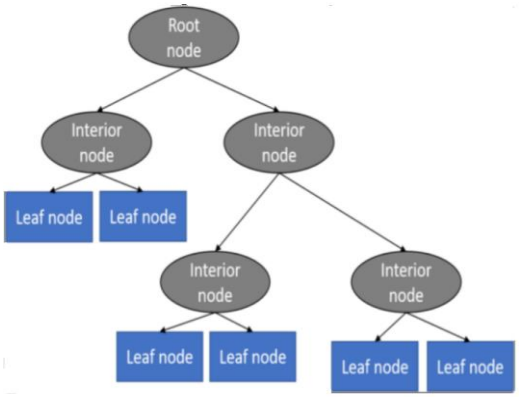

Fig. 4. Decision Tree

D.3. Random Forest

Random forest is a supervised learning algorithm. It can be used for both regression and classification. It is the most flexible algorithm. We have many decision trees in Random trees. To classify a new object, attributes of each tree are considered, as part of the voting process.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Based on the voting result, the forest chooses the classification. Shortly, with Random Forest for small amounts of data, we can train the model and can expect pretty good results. Fig. 5 represents the workflow of the proposed Random Forest method.

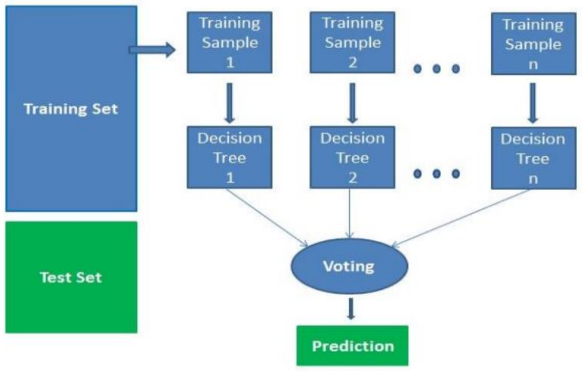

Fig. 5. Random Forest

E. Evaluation

To evaluate the performance of models we will apply two methods – MISE and Cross-Validation.

MSE: The Mean Squared Error (MSE) of an assessor measures the average of the squares of the error. MSE is used to check how close the actual values are to the estimates or forecasts. Higher the MSE, the oppose is forecast to actual and similarly the lower MSE, the closer is forecast to actual. This is utilized as a model assessment measure for regression models and the lower value indicates a better fit.

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

Cross-Validation: In machine learning, any model cannot be fitted on the training data,
and cannot be said that the model will work accurately for the real data. For this, it must be assured that the model is not getting up too much noise and got the correct patterns from the data. Cross-validation is a technique in which using the subset of the dataset, we train our model and then using the complementary subset of the data-set evaluate the model.

IV. EXPERIMENTAL RESULT

The experimental result is shown below.

Table- I: Models Comparison Table

Model Name	Loss
Linear Regression	4.150807182815247
Decision Tree	4.329810030701206
Random Forest	2.901010067504885

As shown in Table-I, Random Forest gives the best results for the training set. The accuracy of the Linear Regression and Decision Tree predictions were not on par with Random Forest on both training and test data.

V. CONCLUSION

For housing price prediction, this paper investigates three different types of Machine Learning methods including Linear Regression, Decision Tree, Random Forest Regression. Their performances are compared and dissected for the best solutions. Although these methods accomplished desired outcomes, different models have their pros and cons. Experimental results show that the proposed CNN Random Forest method gives better result than the other two methods; it has the lowest error and performed very well for both training and testing data. We hope this study will help to provide some methodological and practical contributions to property appraisal and presenting an alternative way to deal with the valuation of housing costs. The future direction of research may consider a bigger geological area with more features, incorporating additional property transaction data beyond housing development.

REFERENCES

1. Michael J. Ball, "Recent Empirical Work on the Determinants of Relative House Prices," in *Urban Studies*, 10, pp. 213-233, 1973.
2. Park, B., & Kwon Bae, J., "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," in Expert Systems with Applications, vol. 42, issue 6, pp. 2928–2934, 2015.
3. Breiman, L., "Random forests," in Machine Learning, vol. 45, issue 1, pp. 5–32, 2001.
4. Breiman, L et al., "Classification and regression trees," New York: Chapman & Hall/CRC Press, 1984.
5. Ranadip Pal, "Overview of predictive modeling based on genomic characterizations," in Predictive Modeling of Drug Sensitivity, 2017.
6. Li et al., "A SVR based forecasting approach for real estate price prediction" in International Conference on Machine Learning and Cybernetics, Hebei, 2009.
7. Rafiei and Adeli, "A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units," in Journal of Construction Engineering and Management, vol.142, issue 2, 2016.
8. Sifei Lu et al., "A hybrid regression technique for house prices prediction," in IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2017.
9. Wang, C. C., & Wu, H., "A new machine learning approach to house price estimation," in New Trends in Mathematical Sciences, vol. 6, issue 4, pp. 165–171, 2018.
10. Mohd et al., "Machine learning housing price prediction in Petaling Jaya, Selangor, Malaysia," in International Journal of Recent Technology and Engineering, vol. 8, pp. 542–546, 2019.
11. Winky et al., "Predicting property prices with machine learning algorithms," in Journal of Property Research, vol. 38, issue 1, 2020.
12. Machine Learning Databases – Housing (StatLib library) [Online]. Available:https://archive.ics.uci.edu/ml/machine-learning-databases/housing/
13. Linear Model-Boston Housing [Online]. Available: https://rpubs.com/TaylorCastro/719068

AUTHORS PROFILE

Amena Begum, is currently working as an Assistant Professor at the Department of Information and Communication Technology, Faculty of Engineering, Comilla University, Bangladesh. She earned a Master of Science (M.Sc) in Engineering with a thesis and a Bachelor of Science (B.Sc) in Engineering from Comilla University's Department of Information and Communication Technology. Deep Learning, Network Forensics, block chain technology and Internet of Things (IoT) are some of her current research interests. Her research publications were published in a number of international journals. She participates in educational and co-curricular activities on a regular basis.

Retrieval Number: 100.1/ijitee.C97410111322
DOI: 10.35940/ijitee.C9741.0111322
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Nishad Jahan Kheya, is studying as an M.Sc (Engg.) student in the Department of Information and Communication Technology, Comilla University, Bangladesh. Her research interest includes Data Science, Network Security and Cyber Law, Bioinformatics, Computer-assisted education.

Md. Zahidur Rahman, passed BSc (Engineering) and MSc (Engineering) from Department of Computer Science and Engineering, Comilla University; and ranked first in both. He is very much interested and passionate to do his research in the area of Machine Learning, Natural Language Processing, Image Processing, Pattern Recognition and Computer Vision. He did some research works in these areas. He has two publications in IEEE, one in Springer, two in Scopus-indexed journals and others in reputed peer reviewed journals. He has been working as a Lecturer in Department of CSE, Britannia University, Cumilla, Bangladesh since October 2019.