House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan

Maida Ahtesham, Narmeen Zakaria Bawany, Kiran Fatima
Research Center for Computing,
Department of Computer Science and Software Engineering,
Jinnah University for Women, Karachi, Pakistan
Email: maidaahtesham@gmail.com, narmeen.bawany@juw.edu.pk, kiranahmedraza@gmail.com

***Abstract*— House prices are a significant impression of the economy, and its value ranges are of great concerns for the clients and property dealers. Housing price escalate every year that eventually reinforced the need of strategy or technique that could predict house prices in future. There are certain factors that influence house prices including physical conditions, locations, number of bedrooms and others. Traditionally predictions are made on the basis of these factors. However such prediction methods require an appropriate knowledge and experience regarding this domain. Machine Learning techniques have been a significant source of advanced opportunities to analyze, predict and visualize housing prices. In this paper, Gradient Boosting Model XGBoost is utilized to predict housing prices. Publicly available dataset containing 38,961 records of Karachi city is attained from an Open Real Estate Portal of Pakistan. Lot of work has been done in predicting house prices across many countries, however very limited amount of work has been done for predicting house prices in Pakistan. Our proposed house price prediction model is able to predict 98% accuracy.**

Keywords*— *Open Real Estate Portal, Gradient Boosting Model XGBoost, Housing Price Prediction, Machine Learning

I. INTRODUCTION

House price prediction refers to a concept of evaluating property prices by using various techniques. It serves as a first hand assistant for people in purchase or sale of properties [1]. Despite having a large number of increase property demands there is no appropriate mechanism that could help predict house prices in future.

Machine learning has been used for image recognition, spam reorganization, medical diagnosis for more than a decade. Machine Learning based predictions achieve better results when put in practice. Almost every economic domain now benefits from machine learning prediction models. In this research paper, House price prediction has been performed using machine learning technique XGBoost. The data set has been taken from an Open Real State Portal of Pakistan [9]. This is a huge dataset as it comprises housing records of many cities of Pakistan, including Karachi, Islamabad, Lahore, Rawalpindi and Faisalabad. This research paper focuses on predicting housing prices of Karachi using publicly available dataset. The housing dataset consist of 38,961 records with distinct set of features. Computational experiment has been performed to develop prediction model with high accuracy and low MAE.

The rest of the paper is structured as follow. Section 2 presents literature review of house price predictions and background study. Section 3 includes processing and analysis of data. Section 4 presents the implementation of machine learning technique that have been applied for house price prediction. Followed by empirical results and discussion in Section 5. Conclusion of the study is outlined in Section 6.

II. BACKGROUND AND RELATED WORK

Machine learning focuses on developing self-learning algorithms as to project future activity based on previous data. House price prediction works on similar phenomenon. This section presents various concepts and existing studies on this particular domain. Many researchers have worked on predicting a housing model, the process of developing an opinion of value is an important tool for evaluating property values when purchasing, selling, insuring, lending or taxing on residency property, said Zhao et al. [1] who applied deep learning in combination with extreme Gradient Boosting (XGBoost) for real estate price predictions, by analyzing historical property sale records. The dataset was extracted from Online Real Estate website. The data split into 80% as training set and 20% as testing test. Each record in dataset contains address, bedrooms, bathrooms, ensuites, garages, land size, and property image. XGBoost hybrid model achieved Mean Absolute percentage error of 8.70% whereas 13.01% for k-NN hybrid model. Experimental evaluation of this research propose that deep learning combined with XGBoost can help attained better results. According to Satish et al. [2] regression deals with specifying the relationship between dependent also called as response or outcome and independent variable or predictor. The study aimed to predict future house price with the help of machine learning algorithm. They compared and explored various prediction methods in order to select the method of prediction. Lasoo regression was selected as their model because of its adoptable and probabilistic methodology, other machine learning models were XGBoost and Neural System. The study found that Lasoo regression, in the view of accuracy, reliably outperforms in the execution of house price prediction.

Another research by T. D. Phan [3] used machine learning techniques to analyze historical property transactions, the study aimed to get helpful information from recorded information of property markets in Melbourne city, Australia and to discover helpful models to anticipate the

978-1-7281-8855-3/20/$31.00 ©2020 IEEE

estimation of house given a set of attributes. Dataset consist of 34,857 observations and 21 attributes. The study showed high disparity between house costs in the most costly and most moderate rural areas in the city of Melbourne. In this paper different regression models were implemented in order to obtain better results. It was demonstrated that the blend of Stepwise and Support Vector Machine established on Mean Square Error measurement is an efficient approach. It is observed that regression tree is as good as linear regression but polynomial regression resulted with lower errors. Whereas, neural network didn't work efficiently with dataset. The study of A. Chouthai et al. [4] predicted house prices using different machine learning algorithms to build the prediction model for houses, such as logistic regression, support vector regression, Lasso Regression technique and Decision Tree employed. The study contains data of 100 homes along with their parameters. Dataset was ere divided with 50% to train the machine and 50% for testing purpose. This research resulted with accurate results.

A. Sinha [5] employed different machine learning techniques for predicting the house prices. Ordinary Least Squares algorithms used in this analysis. Various factors were taken to predict the price like lot size, bedrooms, bathrooms, location, drawing room, and material used in house, interiors, parking area and mainly on square feet per area, etc. As the scope of this paper to predict the house cost, several matrices are used for feature extraction and these variables are called feature data set. The study showed that the location of the house, along with the amenities were highly influenced. The study of Z. Peng et al. [6] aimed to predict the price of second hand houses more accurately. The dataset with 35,417 observations extracted from Chengdu HOME LINK network was taken. The dataset was ere preprocessed, cleaned by removing inconsistencies or incomplete data, corrected anomalies or outdated information and important characteristics were selected. In this way 27,961 records were obtained. Afterwards, multiple linear regression, decision tree and XGboost models were used for predicting housing price score curve, and the appropriate prediction model was selected with considerable preprocessing. The results showed that the accuracy obtained by using XGboost prediction model was highest, the score reached to 0.9251 and the XGboost model proved to be efficient among others having good classification and regression properties and possess positive aid in the processing of such unbalanced data.

C. S. Rolli [7] analyzed the real estate property prices of three counties in California with the help of machine learning algorithms. This study predicted selling and demand prices of house having features such as bedroom count, bathroom count, geographical locations, kitchen size, square feet etc. Multiple machine learning algorithms were used such as Linear Regression, Gradient Boosting, and Random forest Regression. 90% of the data was used as training dataset and 10% as testing dataset. It was concluded that among all regression techniques, XGboost achieved best results.

Linear Regression is commonly used as predictive analysis. This predictive analysis helps to determine effect or impact of change. XGBoost or Xtreme Gradient Boosting is machine learning algorithm used for regression problems and is known for its flexibility, performance and speed. The name XGBoost refers to drive the limits of computational resources in order to boost algorithms.

This algorithm have certain features that help in achieving greater efficiency and performance of model. A tree boosting algorithm by Tianqi Chen [8] is highly productive machine learning algorithm that supports parallel and distributed computing that speed up learning and predict high accuracy.[i]

House price prediction helps the developer in forecasting the prices in a genuine range which also helps clients to decide when and where to buy a house. Buying a suitable house is getting difficult due to rising prices. This paper aims to cover housing market problem of Karachi city which is third mega city of the world. Very limited work has been done for house price prediction in Pakistan. In order to explore the house pricing trends in Karachi- the biggest city of Pakistan, this work utilizes the dataset available at Open Real State Portal of Pakistan [9]. The study presented the results on the basis of various train/test ratios i.e. 60/40, 70/30 and 50/50.

III. DATA ANALYSIS

A. Data Exploration

The real estate property dataset was collected from property data for Pakistan website called Open Data Pakistan [1]. Original Dataset contains 168447 instances and 20 features or variables as given in Table I. It includes property listing of various cities of Pakistan i.e. Islamabad, Rawalpindi, Lahore, Faisalabad, and Karachi.

TABLE I. FEATURE DESCRIPTION OF ACTUAL DATASET

	Name	***Type***	***Description***
1	Property_id	Numerical	Different types of properties i.e. House and flat
2	Location_id	Numerical	Locations or areas where the property situated
3	Page_url	Categorical	Property advertisement link
4	Property_type	Categorical	House type, Flat, Portion etc
5	Price	Numerical	House price (Prediction outcome)
6	Location	Categorical	House location
7	City	Categorical	City located
8	Province	Categorical	Where the city (Karachi) is located
9	Latitude	Categorical	House latitude
10	Longitude	Categorical	House longitude
11	Baths	Numerical	Number of bathrooms
12	Area	Categorical	House Area
13	Purpose	Categorical	For sale or rent
14	Bedrooms	Numerical	Number of Bedrooms
15	Date_added	Numerical	Date of advertising
16	Agency	Categorical	Advertising Agency
17	Agent	Categorical	Advertising agent

	Name	***Type***	***Description***
18	Area_type	Categorical	Square Foot (1 Square = 0-0036 Marla)
19	Area Size	Numerical	House area
20	Area Category	Categorical	House area category

B. Data Preprocessing

Data preprocessing is a technique that is applied on the dataset to make sure the data is effective to use. Before applying models on house price prediction data preprocessing is applied. Among dataset of various cities of Pakistan, dataset of Karachi city is specifically selected for house price prediction. After selecting Karachi dataset investigation of missing data is performed. Rows with missing values were removed from Karachi dataset. This preprocessing also involved removing features that are less effective. Preprocessed dataset consist of 38961 records and 14 features.

Final dataset that was obtained after preprocessing is given in Table II.

TABLE II. FEATURE DESCRIPTION OF DATASET AFTER PREPROCESSING

	Name	***Type***	***Description***
1	Property_id	Numerical	Different types of properties i.e. House and flat
2	Location_id	Numerical	Locations or areas where the property situated
3	Property_type	Categorical	House type, Flat, Portion etc
4	Price	Numerical	House price (Prediction outcome)
5	Location	Categorical	House location
6	City	Categorical	City located
7	Province	Categorical	Where the city (Karachi) is located98
8	Latitude	Categorical	House latitude
9	Longitude	Categorical	House longitude
10	Baths	Numerical	Number of bathrooms
11	Area	Categorical	House Area
12	Bedrooms	Numerical	Number of Bedrooms
13	Area Size	Numerical	House area
14	Area Category	Categorical	House area category

IV. IMPLEMENTATION

A. Reading Data

This study is performed using python machine learning libraries. Dataset files are loaded using pandas library. Other Libraries used numpy, xgboost, scikit-learn, Matplotlib, Seaborn.

B. Removing Missing Data

In order to fit data into model. Pre-assessment of missing data is performed. Dataset contained feature with missing or NaN values. Number of entries in dataset were initially high, hence missing values entries are removed.

Fig. 1. SalePrice Correlation Matrix

C. Separating Categorical and Numerical Data

In order to prepare data for training purpose, categorical data is separated from numerical data. After which categorical data is transformed into numerical data. Features having low correlation with the target variable were removed. The sales price correlation matrix of dataset generated by using Pearson correlation on 38960 records shown in Fig 1.

D. Feature Selection

Identifying important features is essential step. Generalizing model with less data is difficult as less features fails to represents data well. Identifying the key features that are less or more important in house price prediction. Less Important features are selected away or removed using "crcols.remove" and important or partially important features are kept for further processing

E. Split DataSet into Training and Testing

Process of Splitting data set into training and testing divides data into smaller set for building and validating model. Training set has known output on which model learns. Whereas Testing set is to test our models prediction. Prediction Model is check for accuracy and MAE with multiple train and test ratio of 60/40, 50/50 and 70/30

F. Using XGBRegressor

XGBBoost algorithm is used for House price prediction. This algorithm provides flexibility, speed and performance. This study prefers XGBoost in order to get better accuracy results and lower MAE. Model is trained using XGBRegressor and is validated using validation dataset.

V. RESULT AND DISCUSSION

This experiment of predicting house price has been deployed using XGBoost algorithm on python notebook. XGBoost is an application of gradient boosting decision tree algorithm. It was designed to push the computational limits of boosted tree algorithm. Idea of selecting optimized distributed gradient boosting library is being its fast and flexible nature and best used for tabular dataset and classification and regression model. XGBoost allows parallel processing that makes it 10 times faster than other models. Data set employed in this experiment is taken from open data real state property dataset of Pakistan where specifically Karachi based dataset were selected.

In order to predict house prices several metrics are used such as feature selection. **Error! Reference source not found.**II Shows list and details of features that are being used for house price prediction. This dataset comprise of 14 features and 38961 records. Feature selection is a procedure used in this process that required to manually or automatically selecting attributes that contributes to prediction variable. Initially this feature selection technique is used in preprocessing stage where number of features are omitted on the basis of their less association with predicting attribute. Later on while implementing XGBoost model more feature were dropped to assure the efficient results. Data set is being tested using various testing and training ratios to obtained multiple Model accuracy values and Mean Absolute errors.

Accuracy and mean absolute error obtained are presented in Table III. Results have been presented on the basis of various training and testing ratio.

TABLE III. TRAIN TEST RATIO

	Train/Test Ratio	***Learning Rate***	***Model Accuracy***	***Mean Absolute Error***
1	60/40	0.01	98%	22502.0824694
2	50/50	0.01	98%	22502.0824694
3	70/30	0.01	98%	22502.0824694

VI. CONCLUSION

In order to purchase real state property accurate estimation of house prices is necessary. A real state property contains various factors. In order to predict house prices, machine learning algorithms are considered to be efficient techniques. This paper provides insight of XGBoost algorithm as one of the useful algorithm for house price prediction and to provide flexible and efficient results. Dataset used for the experiment were obtained from Open Data Pakistan website. A real state property data set of Karachi city that contain 38,961 records and features including location, city, property type (Flat, House), number of bedrooms, baths, longitude, latitude, area, price. Preliminary assessment of dataset is done including removing NaN values and finding the characteristics that best match with predicting value. Model accuracy and Mean absolute error are being calculated using XGBoost algorithm. Observations made from the obtained results shows that compared to all models used in predictions, XGBoost model outperformed and provided with high rate of accuracy

REFERENCES

[1] Y. Zhao, G. Chetty and D. Tran , "Deep Learning with XGBoost for Real Estate Appraisal," in IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 2019.

[2] G. N. Satish, C. V Raghavendran, M. D. S. Rao, and C. Srinivasulu, "House Price Prediction Using Machine Learning," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 9, pp. 717–722, 2019.

[3] T. D. Phan, "Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia," Proceedings - International Conference on Machine Learning and Data Engineering, iCMLDE 2018, pp. 8–13, 2019.

[4] A. Chouthai, M. A. Rangila, S. Amate, P. Adhikari, and V. Kukre, "HOUSE PRICE PREDICITION USING MACHINE LEARNING," pp. 4403–4406, 2019.

[5] A. Sinha, "Utilization Of Machine Learning Models In Real Estate House Price Prediction," vol. 4, no. 1, pp. 18–23.

[6] Z. Peng, Q. Huang, and Y. Han, "Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm," 2019 IEEE 11th International Conference on Advanced Infocomm Technology, ICAIT 2019, pp. 168–172, 2019.

[7] C. S. Rolli, "ZILLOW HOME VALUE PREDICTION USING XGBOOST," California State University San Marcos, 2019.

[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[9] "Open Data Pakistan," https://opendata.com.pk/dataset/property-data-for-pakistan

[10] S. Xiong, Q. Sun and A. Zhou, "Improve the House Price Prediction Accuracy with a Stacked Generalization Ensemble Model," in International Conference on Internet of Vehicles, 2019.

[11] Q. Truong, M. Nguyen, H. Dang and B. Mei, "Housing Price Prediction via Improved Machine Learning Techniques," in 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019), 2019.

[12] L. Mrsic, H. Jerkovic and M. Balkovic, "Real Estate Market Price Prediction Framework Based on Public Data Sources with Case Study from Croatia," in Asian Conference on Intelligent Information and Database Systems, 2020.

[13] U. K. Cinar, "Combining Domain Knowledge & Machine Learning: Making Predictions using Boosting Techniques," in ICAAI 2019: Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence, 2019.

[14] T. Mohd, N. S. Jamil, N. Johari, L. Abdullah and S. Masrom, "An Overview of Real Estate Modelling Techniques for House Price Prediction," in Charting a Sustainable Future of ASEAN in Business and Social Sciences, 2020.

[15] "Machine Learning Mastery," https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/