APPLYING MACHINE LEARNING MODELS FOR FORECASTING HOUSE PRICES – A CASE OF THE METROPOLITAN CITY OF KARACHI, PAKISTAN

Hyder Ali Khan[1] and Junaid Rehman[2]*

ABSTRACT

The real estate market is a crucial component of the economy, and the accurate prediction of house prices is essential for buyers, sellers, investors, and policymakers in order to make informed decisions. Machine learning algorithms can provide a more accurate & efficient approach to predicting house prices by meaningfully utilizing the vast amount of available housing data. The research was aimed at addressing this gap in the extant literature by collecting firsthand data through web scraping from zameen.com and developing a user-friendly interface for accurate price estimation. To this end, seven machine learning algorithms that included: Ada Boost, Gradient Boosting, Random Forest, Ridge Regression, Lasso Regression, Elastic-Net, and Neural Network were evaluated for their performance based on the metrices like Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) value. The findings showed that the Random Forest model, showing the highest R-square value performed better than any other model. This suggested that in order to reliably forecast the house price using the given data of the Karachi market, the Random Forest model would be suitable for implementing the price prediction platform. In this regard, the additional GUI features further enhanced the usability, accessibility, and user-friendliness of the price prediction model platform that was proposed for sellers, buyers, and investors of property. Overall, the findings of this research have significant implications for the real estate sector, financial institutions, and government, offering valuable insights for more informed financial decision-making in the dynamic real estate market of Karachi, Pakistan.

Keywords: *House Price Forecasting; Machine Learning Models; Karachi Real Estate Market; Multivariate Regression; Random Forest.*

[1] Mohammad Ali Jinnah University, Karachi, Pakistan. Email: hyderkaimkhani@gmail.com
[2] Mohammad Ali Jinnah University, Karachi, Pakistan. Email: junaid.rehman@jinnah.edu
**Corresponding Author*

INTRODUCTION

The housing sector is considered one of the key growth drivers for an economy. There is a consensus among economists and policymakers that the role of this sector is phenomenal in the creation of employment opportunities, and at the same time, it complements the growth of other industries through indirect job creation. Determining the accurate value of a property is a complex task that affects various stakeholders of the real estate industry. While factors like size, rooms, and location are known to impact prices, other aspects such as market demand, population growth, living trends, socio-economic conditions, and government regulations also play a crucial role when it comes to house price determination. Governments, central banks, and policymakers in different countries are aware of the fact that the increase in housing costs not only fuels inflation but also affects the affordability of living for the common households. Nevertheless, the research in the areas of affordable housing and price prediction mechanisms has not been given much attention by prior researchers. For instance, the stock market crash of 2001 that resulted in a recession and a slowdown in the housing market in 2008 caused by the housing market uncertainties in the US could have been avoided (Miller et al., 2011).

Thus, this research focuses on the challenge of accurately predicting house prices in the Karachi market in order to cater to the needs of sellers, buyers, and investors. Accordingly, the primary aim is to develop a robust machine learning based platform that could effectively estimate house prices in the Karachi market in Pakistan, thereby enabling the stakeholders to analyze the complexity & variability and understand the factors that influence the local real estate market in order to make informed decisions. This consequently led us to frame the research question for this study: *'How do machine learning models help in predicting house prices and which machine learning algorithm performs better in terms of developing a more accurate house price prediction platform?'*

To this end, the development of an accurate predictive model for house price estimation required a comprehensive understanding of and the need to implement suitable machine learning algorithms. Accordingly, this research examined how well house prices could be predicted using commonly available techniques such as linear regression, random forest, decision trees, gradient boosting, and ada boost. As part of this research, we evaluated the effectiveness of these techniques from the viewpoint of their accurate estimation of the house prices. In this regard, regression analysis serves as a commonly employed approach for predicting property prices and utilizing target & predictor variables. Our proposed machine learning platform incorporated information about the housing features such as area, number of

bedrooms, number of bathrooms, location, and other key aspects with an aim to provide valuable insights to homeowners, buyers, agents, lenders, and investors.

For this purpose, the datasets were obtained from zameen.com and included data on apartments and houses. Seven machine learning algorithms were evaluated in terms of their suitability and effectiveness, and the best-performing technique was incorporated in developing a user-friendly model interface. We employed various data pre-processing steps to organize the data and enhance the accuracy of our proposed system. These steps involved data cleaning, normalization, and feature engineering. By following these steps, we ensured that the data were in a format that was suitable for utilizing the machine learning models. Subsequently, the performance of the models was evaluated using appropriate metrics. The limitations included data quality and external factors like market trends.

The remainder of this paper is ordered in a way that Section 2 provides an extensive review of the relevant literature, highlighting the existing knowledge and research gaps in the areas of machine learning models and their applications for the development of house price prediction platforms. Section 3 outlines the chosen research methodology and justifies its suitability for addressing the key research question. In Section 4, the results of the data analyses are presented along with a detailed discussion and interpretation of the findings. Section 5 focuses on the development of a user-friendly graphical interface for house price prediction based on the developed model. Section 6 discusses the unique contributions of this study, addressing the gaps identified in previous research. Section 7 explores the implications of the research findings, highlighting the practical applications and contributions of this study for various stakeholders. The limitations of the study along with the future recommendations are highlighted in Section 8, followed by Section 9 which presents the conclusion.

LITERATURE REVIEW

Housing Market

This section presents various related studies that were conducted to find appropriate models for the housing market. The income for many households in modern economies comes from the real estate sector. Since World War II, loans secured by the real estate sectors have accounted for a lion's share of the funding for the financial sector (Jordà et al., 2014). A sizable portion of the investments are residential investments which are vulnerable to significant market fluctuations.

House prices largely shrugged off the effects of the COVID pandemic in 2020 mainly because of the low interest rates, policy support, and increased demand for people to work from home.

This rise in house prices caused affordability issues and resulted in inflation. While house prices continue to soar around the world, an increase in interest rates, government intervention, and improved supply of building materials could lead to a normalization of housing prices. Economists and policymakers consider housing cost as one of the significant factors contributing towards the growth of or slowdown in an economy. According to Friedman's permanent income theory, if changes in property prices have an impact on their predicted lifetime wealth, people will alter their planned consumption. Research by economists suggests that rising property value supports easing the homeowners' borrowing restrictions and boosts their actual consumption (Ahtesham et al., 2020; Miller et al., 2011).

Pakistani Housing Market

According to the Economic Survey of Pakistan (2022), the activities related to real estate showed an increase of 3.7%. About 13.4% of the industry revenue is generated by the construction sector, whereas industrial construction spending accounts for the majority of the value created in the sector. A rise in general government spending was the main driver of the construction industry's modest 3.1% gain. This mild growth rate is the result of an extraordinary rise in the applicable deflator, or Wholesale Price Index (WPI) building material of 30.1%. The second-highest percentage of Gross Fixed Capital Formation (GFCF) in the private sector at 18% is based on activities in the real estate sector which went up from 15% in FY 2016. Due to a larger growth in the deflator, private sector GFCF in the real estate sector has increased by 35%. The Pakistani government promoted these initiatives in October 2020 by introducing the Mera Pakistan Mera Ghar (MPMG) Scheme. This program enabled low-income individuals to borrow from the banks at cheaper interest rates for the construction or purchase of their houses.

Machine Learning

Data science has emerged as a rapidly growing field of study that builds on the foundations of statistics and machine learning. It incorporates diverse fields and techniques, driven by commercial motivations to improve business outcomes. The vision of data science emphasizes the importance of an interdisciplinary approach, an evidence-based approach to improving data-related activities and learning from data (Donoho, 2017). Modern machine learning research involves enabling computers to acquire new knowledge by processing various forms of data, such as numerical values, text, and images, aiming to enhance computer capabilities and allowing them to learn and make decisions autonomously similar to how humans do. As described by (Faggella, 2020), machine learning involves providing computers with data and real-world experiences, enabling them to improve their learning and decision-making abilities

over time (Garg et al., 2023). The underpinning goal is to create intelligent systems that can learn from and interact with their environment, opening up new possibilities in diverse fields.

Applications of Machine Learning Techniques

A subspecialty of artificial intelligence, machine learning, has gained traction in the field of computing. Deep learning has become an integral part of researchers' efforts to enhance the precision of machine learning systems (Garg et al., 2023). It has produced promising outcomes in several applications, and additional research into this area may yield new real-world uses. Numerous fields, including face recognition, speech, gene identification, diabetics, and tumor detection have already benefited from the application of advanced machine learning techniques. Sarker (2021) discusses various fields where machine learning techniques are applied, emphasizing the value of data and how machine learning techniques could be employed in creating intelligent applications.

Machine learning techniques are also becoming more and more common when it comes to estimating and forecasting property value. To this end, the development of an accurate predictive model for house price estimation required a comprehensive understanding of and the need to implement suitable machine learning algorithms (Ahtesham et al., 2020). Accordingly, this research examined how well house prices could be predicted using commonly available techniques such as Linear Regression, Decision Trees, Random Forest, Ada Boost, and Gradient Boosting.

Limitations of Machine Learning Techniques

In recent years, researchers working on artificial intelligence (AI) and Machine Learning (ML) have been able to make smart systems that think more like humans, handling challenging tasks and making decisions. While modern day robots are now able to perform a variety of jobs, nonetheless, machine learning algorithms still have several drawbacks. Stewart (2020) discusses the four limitations of machine learning. The first limitation is ethical, as machine learning algorithms may replace jobs and raise ethical questions about who is responsible if something goes wrong. The second limitation is that machine learning is stochastic, not deterministic, and may miss the physics of a system. The third limitation is the quality and quantity of data because machine learning algorithms need a lot of quality data to deliver useful outcomes. The fourth limitation is the misapplication of machine learning, where it is used to analyze deterministic or stochastic systems inappropriately, potentially leading to spurious correlations and false results.

Yet another issue with machine learning is overfitting, the capacity of a model to generalize new data suffers when it is intricate in nature and closely resembles training datasets. To reduce overfitting, techniques such as regularization, cross-validation, and early pausing can be applied. Data bias is another drawback of machine learning. Because machine learning algorithms heavily rely on the data they are trained on, if the data is biased or lacking, the model will pick up on these biases and continue to use them in its predictions. This can result in false or misleading findings. To overcome this limitation, it is critical to make sure that the training datasets are diverse and representative of the population.

Machine learning model interpretation can be highly complex and difficult, making it difficult to spot any biases or inaccuracies and comprehend how a model predicts. In this regard, techniques like feature importance analysis and model visualization can help increase interpretability, but they are not always successful. Besides, a lot of computing power may be needed to develop and implement machine learning algorithms which can restrict their adaptability and scalability in a practical context. Researchers are looking into more innovative, effective, and resource-saving algorithms and strategies to overcome this constraint.

While it is clear from the above discussions that machine learning is an effective tool for data analysis and making decisions, it is not without drawbacks. Hence, when applying machine learning techniques to real-world applications, the limitations of these techniques such as overfitting, data bias, interpretability issues, and computational resource needs, etc. as discussed above must be taken into account.

Machine Learning Models for House Price Prediction

Numerous research studies have investigated the use of machine learning models in anticipating housing value. In this regard, Komagome-Towne (2016) utilized house price prediction data in California that sought to forecast changes in real estate prices and identify the factors influencing those changes using machine learning techniques. The Random Forest model was determined to be the most appropriate model for this type of dataset as it produced the highest R-square and the lowest RMSE value. To estimate median house price, appropriate regression models were applied. The research concluded that machine learning was a viable substitute method for forecasting real estate prices, offering more precise forecasts, and assisting lenders in making more informed decisions.

The research also highlighted the limitations of machine learning models and recommended further research in order to create more reliable and accurate models (Isaac, 2022). Another study investigated how well Artificial Neural Networks (ANN) and machine learning

regression algorithms performed in forecasting house prices in Malmö, Sweden. In order to forecast house prices, the study used multiple linear Least Absolute Selection Operators (Lasso), Random Forest, and Ridge regression algorithms (Levantesi & Piscopo, 2020). It also examined the connection between variables to identify the most significant influences on home prices. The root square and RMSE scores were used to evaluate the accuracy of the prediction. When using the public dataset for training, the study discovered that Lasso provided the greatest scores among alternative algorithms. The results indicated that machine learning algorithms could be used to reasonably anticipate property values in specific places and Lasso could be the best method for doing so in Malmö, Sweden. However, the research also recommended conducting further research to achieve the robustness of the findings (Abdulal & Aghi, 2020).

In a different study, Kintzel (2019) investigated how machine learning and computer vision could be used to forecast a house price in the real estate market. The author developed a predictive model for forecasting property prices using descriptive data and highlighted the importance of deep learning in utilizing visual information to improve predictions. In order to enhance the predictive model, the study researched deep learning approaches for feature extraction from photos. It also looked at the effect of spatial information on house sale prices. The findings suggested that a combination of descriptive, spatial, and visual information could be used to create a more accurate predictive model. However, the research also reported some limitations to using machine learning and computer vision in real estate valuation such as the subjective factors like personal preferences and emotional attachments to a property which may not be accounted for. Additionally, the reliance on data and algorithms would lead to a lack of transparency and accountability in the real estate industry. This also served as an important consideration when evaluating the use of these technologies in predicting house prices.

In yet another study by Dimoski & Pettersen (2020), the machine learning models were utilized for forecasting quarterly and annual growth rates of houses in Norway. The performance of machine learning models viz., Random Forest, Elastic Net, and Neural Networks were evaluated in three different organizations. The findings showed that the Elastic Net was the most accurate model in predicting quarterly housing growth rates, while Random Forest was the most accurate in predicting yearly growth rates. Moreover, the study found that the most responsible factor that influenced Norwegian house prices was household debt. This study supported previous research claiming that debt was one of the critical factors governing the housing market price determination. Nevertheless, the study was subject to a small macroeconomic time series data, making it insufficient to fully explain and address the housing

market complexities. Moreover, the study only examined the effectiveness of machine learning models in Norwegian professional organizations and thus the results could not be generalized to other countries.

Komagome-Towne (2016) used visualizations to investigate the differences and similarities between various pricing models for single family homes in California. By using 5,142 data listings available at redfin.com and applying multiple regression models for forecasting single-family house prices, the findings provided insights into the application of different regression models in predicting housing prices. By comparing the results of multiple models, the study found that generalized additive models performed best and that a non-linear relationship between predictors and housing prices existed. The findings also pointed out the significance of considering numerous regression models and choosing the most suitable one for any given application. However, some counterarguments must also be considered. Primarily, the research only highlighted a specific region, limiting the applicability of the results in other regions. Secondly, the study only considered a limited number of features for prediction and therefore missed the other key factors that influence house prices.

Multiple Linear Regression for Housing Price Prediction

This literature provides an overview of the use of multiple linear regression as a methodology for predicting house prices based on various parameters. In predicting house prices, the main goal is to aid buyers and sellers in making informed decisions regarding the property value. The methodology involves collecting, cleaning, and pre-processing data and then training & evaluating the regression model. Recent studies in these areas suggest that multiple linear regression has shown promising results in accurately predicting house prices. However, the quantity & quality of data utilized for training can affect the accuracy of the model. Furthermore, other factors such as market trends and economic conditions may also impact the actual house prices, which might not be captured by the model. Regression analysis is the optimum method for model fitting and forecasting when two or more independent variables explain the pragmatic behavior of one or more dependent variables. Regression techniques come in a wide variety of forms in literature. The linear, ridge, stepwise, polynomial, elastic net, logistic, lasso, etc., are some popular types of regression (Levantesi & Piscopo, 2020; Sarojamma & AnilKumar, 2018).

Zhang (2021) presents an analysis of the application of multiple linear regression for predicting housing prices. He emphasized the importance of considering significant factors affecting general housing prices and determined a multiple linear regression model for estimation. He

examined the model on a real estate dataset in Boston and discovered that it could forecast and evaluate house prices to a certain extent.

These days, machine learning algorithms are increasingly being used to predict housing prices due to their encouraging results in several R&D areas. One such study used the Boston housing dataset from the UCI Machine Learning repository to examine the Random Forest algorithm's ability to predict the price variance of homes (Kaewchada et al., 2023; Adetunji et al., 2022). The study demonstrated that, in many real-world applications, forecasting a price variance rather than a particular amount is more appealing and practical. The outcomes also showed that the proposed model, having a 5% error margin, reliably predicted the value when compared to the actual value. On the other hand, there are some counterarguments to be taken into account that the price of a house is affected by a variety of factors, thus relying simply on machine learning algorithms may not always produce accurate predictions. Moreover, the over-reliance of the real estate sector on machine learning models that lack human experience and judgment could potentially result in inaccurate results (Garg et al., 2023; Adetunji et al., 2022).

From the above scholarly arguments and discussions, there is still a dearth of research governing house valuation using predictive models. Prior researchers have mostly predicted housing prices by employing traditional regression and autoregressive models. In particular, when it comes to the housing market of metropolitan cities like Karachi, there are very few studies that have utilized appropriate machine learning models for predicting house prices. It is therefore crucial to identify suitable machine learning techniques that could help accurately predict house prices by considering the socio-economic factors relating to the Karachi housing market. This research was thus aimed at determining the median house price of the Karachi property market using the most competitive machine learning models.

METHODS

Methodologically speaking, this research was conducted in five steps. These were data collection, preprocessing, feature processing, model training & model evaluation. Each of the steps is discussed below:

Data Collection

The housing related datasets from the earlier data science initiatives are widely available but they are mostly outdated. We determined that the best way would be to use web-based real estate data to develop a dataset for the purpose of this research. We mainly used data from Karachi, Pakistan. In this regard, a Python scrapper was used to extract data from zameen.com. The data were then transformed into a format that was processable by the machine learning

models. A machine learning platform can analyze the outcome of data parsing, but a human would find it difficult to understand. When collecting data from the internet, people are at significant risk of making mistakes, but computers transfer data between systems in the form of data structures that provide high data integrity. As the script is created for a specific data format, it is not always the case for any data to be in a compatible format. The inconsistency and incompatibility of the data could lead to a problem. As a result, data scraping only captures raw data requires laborious preparation, and even calls for occasional human intervention. Data scraping cannot be totally automated because it depends heavily on internet sources from which data are acquired. For example, when we scrape data from a website, the ideal way for a developer is to assign an attribute ID and an attribute of the class to each unique HTML element, and each item in the same group has these attributes assigned. This makes it easier to write scripts in virtually every programming language. In this research, we used BeautifulSoup to build the crawler for this study. As presented in Table 1, the dataset adapted from Ahtesham et al. (2020) included 10 characteristics or variables and 15,613 instances. It provides property listings for the city of Karachi, Pakistan.

Table 1. Data Set Description

S #	Name	Type	Description
i.	Webpage Link	Categorical	Ads link
ii.	Title	Categorical	Property Title
iii.	Property Type	Categorical	Flat, House
iv.	Location	Categorical	House Location
v.	Price	Numerical	House Price
vi.	Bedrooms	Numerical	No of Bedrooms
vii.	Baths	Numerical	No of Bathrooms
viii.	Area	Categorical	Property Area
ix.	Area Size	Numerical	House Area (Sq ft)
x.	Date	Numerical	Advertisement Date

Data Preprocessing

Data preprocessing is done to clean the datasets for achieving improved machine learning models. It is used on raw data that cannot be analyzed. As in the case of source data for this research, we drew the same from the website where the real estate agents and homeowners entered their data, so it was subject to missing values, formatting inconsistencies, and inaccurate information. Hence using data transformation techniques such as data wrangling and

data munging, the raw data records were transformed into a format that was appropriate for machine learning processing. Also, we deleted any incorrect or missing values from the dataset before doing an iterative analysis of the data. Furthermore, to create an integrated dataset, we performed data integration by combining data from various key sources (Zhan et al., 2023).

In this research, we came across different measuring units and currencies for the asking price and property area of the houses. To ensure consistency and comparability, we undertook the following data pre-processing steps in particular:

- ***Asking Price Conversion:*** The asking prices were originally presented in various currencies and units, such as crore, lakh, and thousands. We standardized all prices to Pakistani Rupees (PKR) to eliminate any currency discrepancies. Additionally, we converted all prices to a consistent numerical format for ease of analysis and comparison.
- ***Property Area Conversion:*** The property area was measured in different units, including square yards, square meters, marla, kanal, and square feet. To ensure uniformity, we converted all area measurements to square feet. This conversion allowed us to have a consistent and standardized metric for property size across the dataset.

By performing these conversions and standardizations, we were able to remove any discrepancies arising from different currencies and units, making the data more coherent, and thereby facilitating accurate analysis and modeling for house price prediction (Ahtesham et al., 2020). We also employed different encoding techniques for transforming the 'Property Type' and 'Location' variables in the dataset. For the 'Property Type' variable, we utilized a one-hot encoding approach. One-hot encoding converts categorical variables into binary vectors and each classification is denoted by a binary number (0 or 1). In the case of 'Property Type', we created separate binary columns for each property type category. The dataset had property types such as 'House' and 'Flat' (apartment) we transformed them into two separate binary columns, with values of 1 indicating the presence of that particular property type and 0 otherwise. This encoding technique allowed the machine learning model to interpret the property type as distinct features.

Now for the 'Location' variable, we applied label encoding. With label encoding, each category of a categorical variable was given a distinct numerical value. In our case, we assigned a numeric label to each location in the dataset. For instance, if the original dataset had locations like 'City Center', 'CBD', and 'Downtown', we assigned them numerical labels as 1, 2 & 3 respectively. The application of label encoding helped preserve the ordinal relationship between different locations, enabling the models to capture any inherent hierarchy in the data

and effectively handle the categorical variables. By employing these encoding techniques, we effectively transformed the 'Property Type' variable into a set of binary features and the 'Location' variable into numerical labels, enabling our machine learning models to process and analyze these variables in a meaningful way during the prediction process.

Machine Learning Model Development

In the data modeling process, several steps were taken to prepare the dataset for machine learning algorithms. Firstly, houses with prices above PKR 100,000,000 (10 crore) were removed from the dataset to focus on a specific price range. Next, the location information was encoded using label encoding, which assigned a unique numerical value to each location category. To handle the property type column, one-hot encoding was applied, creating binary columns for each property type (flat or house). This allowed the model to understand and utilize the categorical property type information effectively. In order to capture non-linear relationships, polynomial features were generated for relevant columns such as bedrooms, bathrooms, area sq ft, location, and property type. This expanded the feature space by creating interactions between the variables.

To address the skewed distribution of the target variable (price), a logarithmic transformation (log(1+x)) was applied. This helped to normalize the data and reduced the impact of extreme values. During the prediction phase, a reverse transformation was performed using the equation e^{x-1} to obtain the actual predicted price. This ensured that the predicted prices were present in their original scale and format. Lastly, GridSearchCV with k-fold validation (k=5), was employed on Gradient Boosting, AdaBoost, and Random Forest models. This technique aided in finding the optimal hyper-parameters for these models, resulting in improved performance and reliability of house price predictions (Zhan et al., 2023; Ahtesham et al., 2020). Overall, these data modeling techniques enhanced the predictive accuracy and precision of the chosen machine learning models by preprocessing the data, transforming the target variable, and incorporating polynomial features.

RESULTS AND ANALYSIS

The results of this research have significant implications for various stakeholders of the real estate industry, including home buyers, sellers, and financial institutions. Improved accuracy in house price prediction can assist buyers in making informed decisions, sellers in setting competitive prices, and banking & financial institutions in risk assessment and making investment & lending decisions. The sample of the dataset loaded into Python is shown in

Figure 1. While figure 2 shows the type and non-null count of each variable. Following that, the statistics of the dataset were generated which are presented in Figure 3.

	Price	Location	Bedrooms	Baths	Area Sq ft	Date	Property Type
0	35500000.0	Bahria Town - Precinct 1, Bahria Town Karachi	5	6	2448	2023-05-06	House
1	67000000.0	Askari 5 - Sector J, Askari 5	5	5	3375	2023-05-06	House
2	36000000.0	Bahria Town - Precinct 1, Bahria Town Karachi	5	5	2250	2023-05-06	House
3	69500000.0	Askari 5 - Sector J, Askari 5	5	7	3375	2023-05-06	House
4	21000000.0	Bahria Town - Precinct 35, Bahria Sports City	4	4	3150	2023-05-06	House

Figure 1. *Sample Dataset*

Variable	Type	Non-null Count
Price	float64	15613
Location	object	15613
Bedrooms	int64	15613
Baths	int64	15613
Area Sq ft	int64	15613
Date	datetime64[ns]	15613
Property Type	object	15613

Figure 2. *Field Types of Datasets*

	Price	Bedrooms	Baths	Area Sq ft
count	1.561300e+04	15613.000000	15613.000000	15613.000000
mean	4.664278e+07	3.761097	3.886825	2325.304362
std	5.422817e+07	1.601804	1.577266	1778.234482
min	7.000000e+05	1.000000	1.000000	117.000000
25%	1.400000e+07	3.000000	3.000000	1080.000000
50%	3.000000e+07	3.000000	4.000000	1800.000000
75%	6.700000e+07	5.000000	5.000000	3150.000000
max	8.500000e+08	11.000000	10.000000	32400.000000

Figure 3. *Summary of Statistics*

Moreover, figure 4 below shows the bar graph of property type. We have two types of properties in our data (house & flat/apartment). In terms of price, the area is seen to be highly correlated with price in Figure 5 with a value of 0.86 once the correlation analysis has been completed.

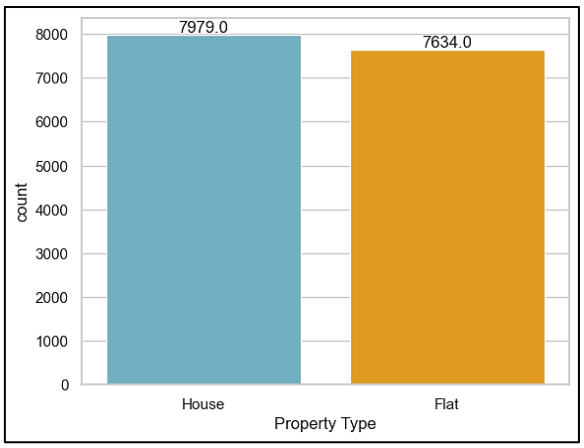

Figure 1. *Property Type Bar Graph*

Figure 2. *Correlation Analysis*

The experimental outcomes of the machine learning models used in our work to predict house prices are shown here. An attempt was made to identify any price field outliers. It can be seen in Figure 6 that the dataset had some outliers, and the interquartile range is from 0 to 10 crore. The scatter plot and histogram of different variables are shown in Figure 7. The histogram of prices shows that most houses have 3 bedrooms and prices are going up as the area increases.

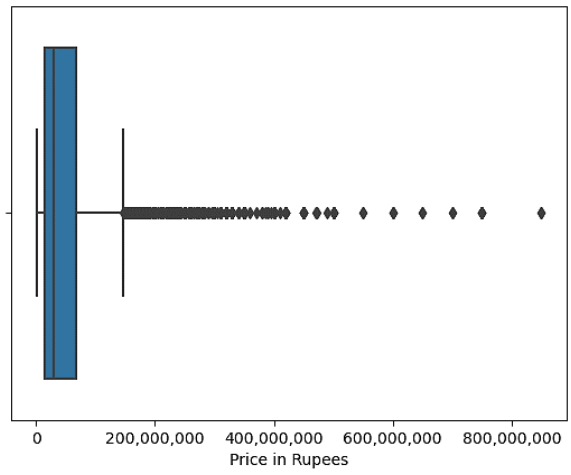

Figure 6. *Box Plot of the Price*

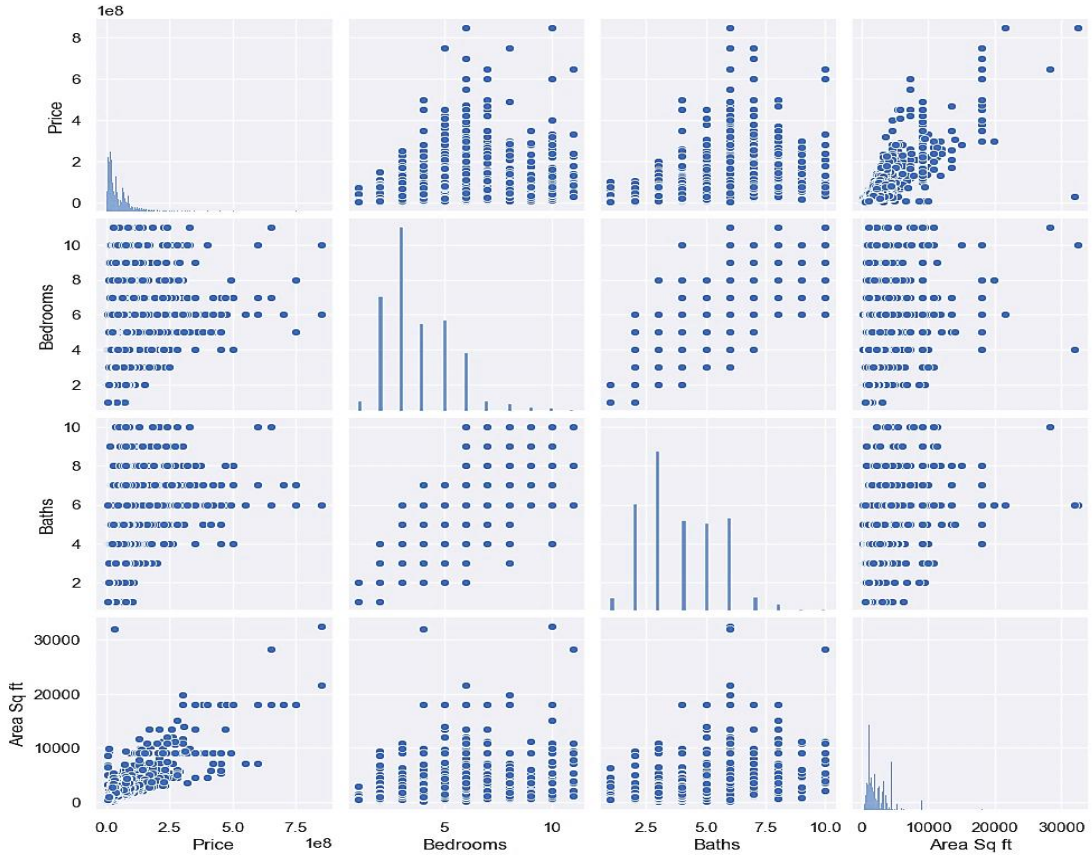

Figure 7. *Pair Plot*

We observed outliers in the pricing data throughout the training phase, which had a negative impact on our assessment metrics including R-squared value, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). To address this issue, we performed outlier removal by eliminating data points that fell outside the interquartile range. This preprocessing step helped improve the performance of our models and resulted in more accurate predictions. The results obtained from applying the AdaBoost algorithm to predict house prices are discussed below:

The Mean Absolute Error value is 0.298, indicating that the predicted house prices on average deviate from the actual prices by approximately 0.298 units. The Mean Squared Error value is 0.149, reflecting the average squared difference between the predicted and actual house prices. The Root Mean Squared Error value is 0.387. Lastly, the R-squared (R2) value is 0.827, indicating that approximately 82.7% of the variance in house prices can be explained by the independent variables. These results showed that the AdaBoost algorithm performed well in predicting house prices, with relatively low MAE and MSE values, a reasonable RMSE value, and a high R-squared value. Using the Ridge Regression method to forecast housing prices, the anticipated housing prices often differ by about 0.264 units from the actual values, according

to the Mean Absolute Error. The average squared difference between the projected and actual home values is 0.126, which is represented by the Mean Squared Error (MSE) number. The average difference between projected and actual prices, considering both direction and size, is measured by the Root Mean Squared Error (RMSE), which has a value of 0.355. Finally, the R-squared (R2) value is 0.8546, indicating that the independent variables employed in the Ridge Regression can account for about 85.46% of the accuracy in house prices.

The Lasso Regression model was applied to forecast home prices, the MAE is 0.283, the MSE is 0.146 and the RMSE stands at 0.383. The independent variables utilized in the model can account for about 83.07% of the variance in house prices, according to the R-squared value of 0.8307. The Lasso Regression model exhibits somewhat higher errors and a slightly lower R-squared when compared to the earlier Ridge Regression model. The Elastic Net model resulted in an MAE of 0.279, MSE of 0.142, RMSE of 0.377, and an R-squared value of 0.835. These metrics measure the performance of the model using average deviation, squared differences, and overall variance explained in house prices.

Now comparing these results with the Lasso Regression model, the Elastic Net model shows similar accuracy and predictive power. These findings highlight the effectiveness of the Elastic Net algorithm in forecasting housing prices and its potential application in real estate (Zhan et al., 2023). The Gradient Boosting model was used to predict house prices, and evaluation metrics like MAE of 0.133, MSE of 0.049, RMSE of 0.223, and a strong R-squared value of 0.942 were obtained. These measurements show how well the Gradient Boosting algorithm predicts house price changes and how accurate it is. The high R-squared value shows that a sizable percentage of the variation in housing prices is explained by the Gradient Boosting model. These results highlight the Gradient Boosting algorithm's capability to estimate property prices with accuracy and consistency.

The performance parameters for the Random Forest model used to forecast house prices were MAE of 0.125, MSE of 0.046, RMSE of 0.2149, and an outstanding R-squared value of 0.9467. These measurements show how well the Random Forest algorithm predicts house prices and how accurate it is. In comparison to earlier models, the model shows lower MAE, squared differences, and RMSE, demonstrating its greater performance. A significant percentage of the variation in housing prices is also explained by the Random Forest model, as indicated by the high R-squared value.

The MAE, MSE, RMSE, and R-squared values produced by the Neural Network model for predicting house prices are 0.563, 0.436, 0.660, and 0.496, respectively. In comparison to the

earlier models, the current model displays higher MAE, squared differences, and RMSE, indicating a comparatively higher level of prediction mistakes. Although the Neural Network method has some promise, these findings point to the need for additional adjustments and fine-tuning to increase its precision and predictive capacity for tasks involving house price prediction. A comparative summary of various models as discussed above is presented below:

Table 2. Comparison of Models

Models	MAE	MSE	RMSE	R-Squared
Ada Boost	0.298	0.149	0.387	0.827
Random Forest	0.125	0.046	0.215	0.947
Gradient Boosting	0.133	0.049	0.223	0.942
Ridge Regression	0.264	0.126	0.355	0.854
Lasso Regression	0.283	0.146	0.383	0.83
Elastic Net	0.279	0.142	0.377	0.835
Neural Network	0.563	0.436	0.66	0.496

Among the evaluated machine learning models, Random Forest demonstrated superior performance in predicting house prices. These metrics indicate that the Random Forest algorithm achieved the lowest prediction errors and significantly explained the variance in house prices. The robustness and accuracy of the Random Forest model make it a promising approach for house price prediction tasks.

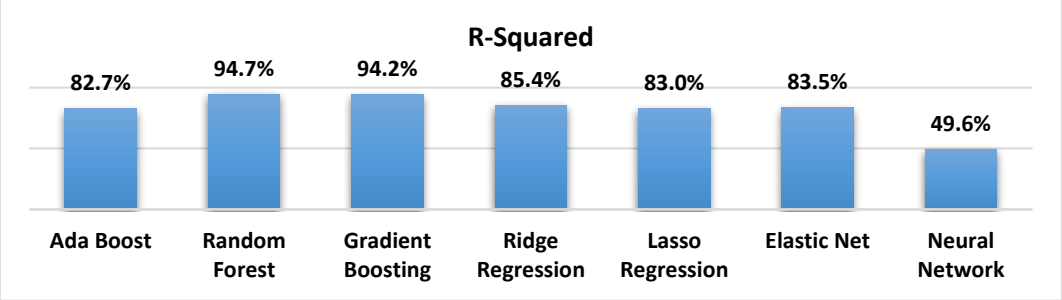

Figure 8. *R-squared Values for Various Models*

Graphical User Interface for Proposed House Price Prediction Platform

The development of a Graphical User Interface (GUI) was an essential part of this research, providing a user-friendly platform for users to interact with the house price prediction model. The Random Forest model, which appeared to be the most suited and accurate model for projecting property prices in the Karachi market, formed the basis for the design and development of GUI. Because it can manage complicated datasets, and high-dimensional feature spaces, and produce reliable predictions, the Random Forest model was chosen for implementing the model platform. The model was trained and evaluated using the collected

dataset and several assessment matrices were used to assess the model performance (Jessica et al., 2023).

The finally developed GUI based platform enabled the users to enter pertinent details about a residential property, such as its size, number of bedrooms, bathrooms, and location. The Random Forest model was used to process and pass these inputs, and it produced a prediction for the house price based on the available data. The development of GUI required making a user-friendly and aesthetically attractive interface that was both intuitive and easy to use. The goal of the design was to create a seamless user experience, enabling the users to easily enter the relevant data and get precise price results. To make sure that the inputs were accurate and within the desired range, the GUI additionally included the necessary error handling and validation procedures. Following is the GUI layout of the model price prediction platform.

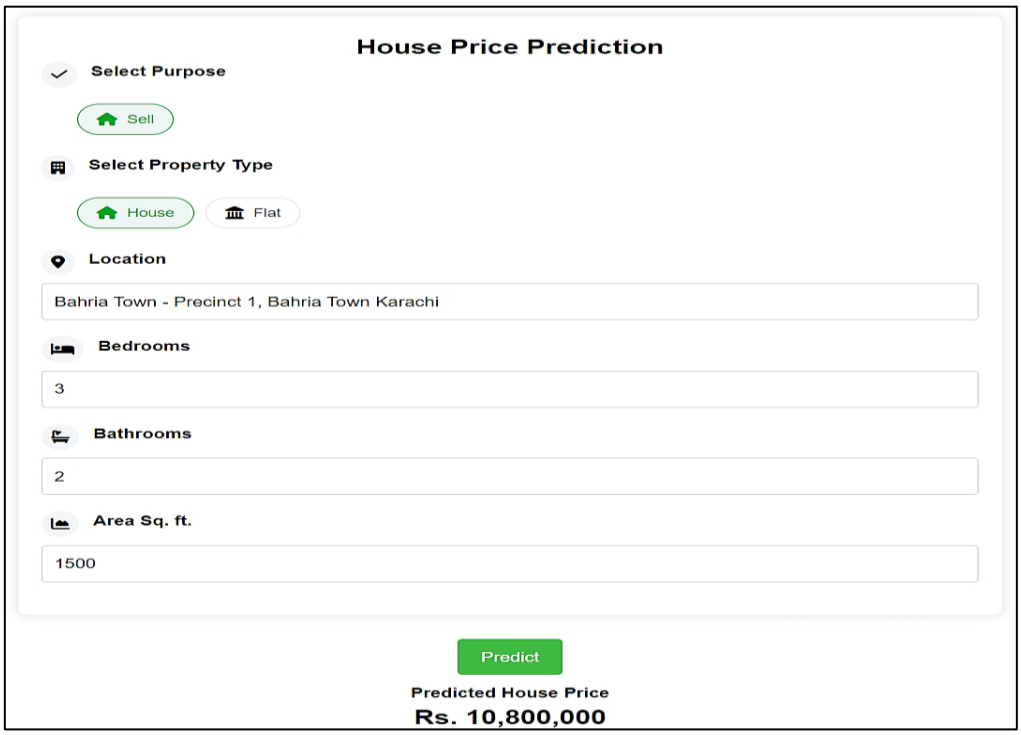

Figure 9. *GUI for House Price Prediction Model Platform*

The obtained price value as shown in figure 9 is based on the various parameters input by the user. To accurately predict the median house value, we employed the random forest model. By leveraging this model, the application can provide reliable estimates of house prices based on the specific input provided by the user. This ensures that users receive key insights on the potential value of houses or units they are interested in.

DISCUSSION

The housing sector is vital for the economic growth of a nation. Accurate predictions of house prices are crucial for making informed decisions. Traditional methods of estimating house

prices often rely on limited information and subjective assessments, leading to potential risks. However, the application of machine learning techniques offers promising opportunities for more accurate and reliable predictions. By leveraging advanced machine learning models and algorithms, the analyses done in this research were aimed at effectively examining vast historical data and recognizing the key factors that influence house prices.

A review of the literature revealed a sizable gap in the research on machine learning-based house price prediction models, especially in the context of the property market of Karachi. Most of the previous studies have focused on other geographical locations and utilized pre-existing datasets.

As part of this research, we addressed the gap by developing a user-friendly interface and data that was acquired through web scraping from zameen.com. This unique approach allowed us to capture the specific dynamics and details of the Karachi real estate market. By incorporating seven different machine learning algorithms and using the available comprehensive datasets, this research offers a robust platform for accurately estimating house prices in the Karachi property market. The inclusion of a user interface enhanced the usability and accessibility of the prediction model, making it practical and beneficial for sellers, buyers, and investors operating in the Karachi real estate market.

Nevertheless, the researchers can explore more sophisticated machine learning algorithms and techniques, especially the ones designed for house price prediction. This may include the inclusion of new characteristics, perfecting present models, or innovating existing methods to enhance forecasting accuracy. They can evaluate the impact of various types of data on house price prediction such as by combining demographic, geographic, or environmental factors to augment the predictive efficiency of the models. The effectiveness of machine learning based house price forecasting models can be enhanced by taking into account the missing, incomplete, or potentially inaccurate datasets coupled with improving the interpretability and openness of the models for better results.

In addition, researchers can focus on integrating the machine learning models with other technologies viz., natural language processing, image recognition, etc. to extract supplementary data from the property images or descriptions, thereby further augmenting the accuracy of the predicted price. Moreover, machine learning models can examine distinct geographic locations or housing markets, identifying the need for localized models rather than generalized models. Also, the machine learning researchers by joining hands with the real estate sector and relevant government departments can contribute to the development of ethical

guidelines and frameworks governing the fair, transparent, and responsible use of machine learning techniques in the housing and real estate sector.

CONTRIBUTIONS AND STUDY IMPLICATIONS

This research greatly contributes to the theory of and research on machine learning models utilized for determining house prices. On a practical front, the developed model platform offers numerous benefits in terms of increased accuracy, accurate pricing analysis, well-informed decision-making, effective risk assessment, and data-driven market insights. In particular, it has the following implications for various stakeholders.

Implications for Theory & Research

From a theory and research perspective, this study makes two significant contributions.

- First, a comprehensive housing dataset specifically collected and tailored to the Karachi housing market serves as a valuable resource for future research and analysis of the Karachi housing market and understanding its dynamics.
- Second, this research offers a sophisticated model platform that leverages machine-learning techniques to predict house prices based on user input variables. This platform allows users to input specific parameters related to the type and class of accommodation they are interested in. Using the collected dataset and advanced machine learning models, the platform provides accurate predictions of house prices, enabling users to make informed decisions when it comes to buying and selling properties.

Implications for Practice

Banking and Financial Institutions

- Banking institutions greatly contribute to the growth of the housing industry because they offer mortgage loans, risk assessment services, and opportunities for investments. Accurate prediction of house prices can aid them in better decision-making and review of their existing business plans. In addition, as part of loan application processing, inspections and appraisals are often conducted by lending institutions in order to ensure that the assessed property value would cover the risk of the loaned amount and help avoid financial loss in case of loan default.
- For financial institutions, the proposed machine learning-based house price forecasting platforms would assist the financial institutions in terms of efficient investment planning, better risk assessment & management, and improved services for the customer as a whole. Financial institutions would be able to use these insights while offering customized

financial services & solutions to their clients, thereby maximizing their return on investment.

Implications for the Real Estate Market of Karachi

The proposed house price prediction tool has transformative implications for the real estate sector. By leveraging advanced algorithms and data analysis techniques, machine learning models can provide valuable insights into property valuation and market dynamics. In particular, it offers profound benefits to buyers, sellers, and realtors via timely property market insights, accurate pricing recommendations, better risk assessment, and fraud detection, empowering them to predict the right time to sell/purchase a property or identify lucrative investment opportunities.

Implications for the State Bank

The results would also assist the State Bank of Pakistan in more meaningfully devising their monetary policies and taking appropriate measures that promote healthy growth of the housing market, thereby ensuring affordable living and genuine investments in the housing sector. The more informed and data-driven policy decisions by state bank would also boost multiple other industries at the same time, leading to new employment opportunities and overall economic growth of the country.

Implications for Federal, Provincial & Local Governments

Governments have always been very much interested in making accurate predictions of house prices since these have a direct impact on many socio-economic areas including housing policies, urban planning, resource allocation, taxation on the property market, new employment creation, economic growth, etc. Given the availability of price forecasting platforms, federal, provincial & local governments in Pakistan now have the opportunity to make data-driven decisions and review the current taxation mechanisms, housing market policies, and regulations. In general, by utilizing any machine learning platform, these governments can take into account the available housing market data & information to identify the key factors and emerging trends that affect the housing market prices. Besides, the additional improvements in these areas would further help these governments ensure affordable housing, thereby improving governance and achieving sustainable economic growth as a whole.

LIMITATIONS AND FUTURE RECOMMENDATIONS

On the one hand, this research attempted to provide valuable insights into house prices in the Karachi market, but there are several limitations that should be considered.

- First the accuracy of the prediction model heavily relies on the quantity & quality of the datasets trained for the purpose. Despite the efforts made in gathering and preprocessing a large dataset, there could still be issues governing the completeness and potential biases of the data. As the dataset was obtained from a specific source (zameen.com), which may not fully represent the entire housing market of Karachi.
- Second, the house prices in the real estate market are affected by several external factors such as market trends, economic conditions, government policies, socio-cultural dynamics, etc. These factors may not be fully captured by the model, as they are subject to continuous changes and uncertainties.
- Third, the developed prediction model and user interface were specifically tailored for the Karachi housing market and may not be directly applicable to other cities or regions.
- Last, the Random Forest model was selected based on the evaluation metrics of specific datasets, different models may perform differently depending on the dataset and problem domain.

Given the unique environment and potential uncertainties of the housing market, it is critical to be aware of these constraints and interpret the study results accordingly. Future researchers should address these limitations through further research and model upgrades with an aim to enhance the precision and applicability of the suggested model. Moreover, the regular merging of new property transaction data from a greater geographic area with other features may also be included in future studies. This will make the prediction system more reliable and accurate. Last but not least, another improvement that can be made is to add more factors like garden, floors, build type, nearby hospital, etc. in order to incorporate improved price prediction capabilities in the existing model.

CONCLUSION

This research was primarily aimed at determining the most effective and useful machine learning method for predicting house prices and creating a model platform based on factors such as location, area, property type (flat or house), number of bedrooms, bathrooms, etc. The source of the data was zameen.com. In the training process, the datasets were split into 2 groups, 80% for the training phase, and 20% for the testing phase. In this regard, we investigated seven machine learning algorithms Ridge Regression, Lasso Regression, Elastic Net, Ada Boost, Random Forests, Gradient Boosting Machines, and Neural Networks (Levantesi & Piscopo, 2020). The Random Forest model, having the highest R-square values and the lowest MAE, MSE, and RMSE appeared to be the most suitable model for

implementing the house price prediction platform. Accordingly, the machine learning model was integrated into a user-friendly interface. The model accuracy and performance were evaluated using appropriate metrics. Limitations include data quality and external factors like market trends. Overall, this study was aimed at providing valuable insights and property-related recommendations to buyers, sellers, investors, and realtors in addition to government and banking & financial institutions.

REFERENCES

Abdulal, A., & Aghi, N. (2020, August 13). *House Price Prediction*. Kristianstad University Research Portal. https://researchportal.hkr.se/en/studentTheses/house-price-prediction-5

Ahtesham, M., Bawany, N.Z., & Fatima, K. (2020). House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan, *21st International Arab Conference on Information Technology (ACIT), 2020*.

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science, 199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

Dimoski, M., & Pettersen, M. (2020). *Predicting housing prices with machine learning: A macroeconomic analysis of the Norwegian housing market* [Master thesis]. https://openaccess.nhh.no/nhh-xmlui/handle/11250/2734788

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Faggella, D. (2020, February 26). *What is Machine Learning? - An Informed Definition*. Emerj Artificial Intelligence Research. https://emerj.com/ai-glossary-terms/what-is-machine-learning/

Garg, K., Das, N.N. & Aggrawal, G. (2023). *A Review On: Autism Spectrum Disorder Detection by Machine Learning Using Small Video*, 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT), 2023.

Housing Prices Continue to Soar in Many Countries Around the World. (2021, October 18). IMF. https://www.imf.org/en/Blogs/Articles/2021/10/18/housing-prices-continue-to-soar-in-many-countries-around-the-world

Isaac, A. (2022). *Combining Machine Learning models to predict House Prices*.

Jessica, A., Raj, F.C. & Sankaran, J. (2023). *Credit Card Fraud Detection Using Machine Learning Techniques*, 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), 2023.

Jordà, Ò., Schularick, M., & Taylor, A. M. (2014). *The Great Mortgaging: Housing Finance, Crises, and Business Cycles* (Working Paper No. 20501). National Bureau of Economic Research. https://doi.org/10.3386/w20501

Kaewchada, S., Ruang-On, S., Kuhapong, U. & Songsri-in, K. (2023). Random forest model for forecasting vegetable prices: a case study in Nakhon Si Thammarat Province, Thailand, *International Journal of Electrical and Computer Engineering (IJECE)*.

Kintzel, J. D. (2019). *Price Prediction and Deep Learning in Real Estate Valuation Models*.

Komagome-Towne, A. (2016). *Models and Visualizations for Housing Price Prediction*. California State Polytechnic University, Pomona.

Leamer, E. E. (2015). Housing Really Is the Business Cycle: What Survives the Lessons of 2008–09? *Journal of Money, Credit and Banking*, *47*(S1), 43–50. https://doi.org/10.1111/jmcb.12189

Levantesi, S. & Piscopo, G. (2020). The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach, *Risks*, 2020.

LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27*.

Miller, N., Peng, L., & Sklarz, M. (2011). House Prices and Economic Growth. *Journal of Real Estate Finance and Economics*, *42*. https://doi.org/10.1007/s11146-009-9197-8

Pakistan Economic Survey. (2022). *Pakistan Economic Survey 2021-22 Real Estate Highlights*.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sarojamma, B., & AnilKumar, K. (2018). *A Study On Comparison Among Ridge, Lasso, And Elastic Net Regressions*. *6*(1).

Stewart, M. (2020). *The Limitations of Machine Learning*. Medium. https://towardsdatascience.com/the-limitations-of-machine-learning-a00e0c3040c6

Tzanis, G., Katakis, I., Partalas, I., & Vlahavas, I. (2006). *Modern Applications of Machine Learning*.

Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. *Scientific Programming*, *2021*, e7678931. https://doi.org/10.1155/2021/7678931

Zhan, C., Liu, Y., Wu, Z., Zhao, M., Chow, T.W.S. (2023). A hybrid machine learning framework for forecasting house prices, *Expert Systems with Applications*.